WEB TOPIC DETECTION USING A RANKED CLUSTERING-LIKE PATTERN ACROSS SIMILARITY CASCADES

Fei Jia[†], *Junbiao Pang*[‡], *Weigang Zhang*[♭], *Guorong Li*[†], *Chunjie Zhang*[†], *Qingming Huang*[†‡♯], *Yugui Liu*[†]

[†]School of Computer and Control Engineering, University of Chinese Academy of Sciences, China
[‡]Beijing Key Laboratory of Multimedia and Intelligent Software Technology,
College of Metropolitan Transportation, Beijing University of Technology, China
[♯]Key Lab. of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences, China
[♭]School of Computer Science and Technology, Harbin Institute of Technology, China
{jiafei, jbpang, wgzhang, grli, cjzhang, qmhuang }@jdl.ac.cn liuyg@ucas.ac.cn

ABSTRACT

In multi-media and social media communities, web topic detection poses two main difficulties that conventional approaches can barely handle: 1) there are large inter-topic variations among web topics; 2) supervised information is rare to identify the real topics. In this paper, we address these problems from the similarity diffusion perspective among objects on web, and present a clustering-like pattern across similarity cascades (SCs). SCs are a series of subgraphs generated by truncating a weighted graph with a set of thresholds, and then maximal cliques are used to describe the topic candidates. Poisson deconvolution is adopted to efficiently identify the real topics from these topic candidates. Experiments demonstrate that our approach outperforms the state-of-the-arts on two datasets. In addition, we report accuracy v.s. false positives per topic (FPPT) curves for performance evaluation. To our knowledge, this is the first complete evaluation of web topic detection at the topic-wise level, and it establishes a new benchmark for this problem.

Index Terms— Web Topic detection, maximal cliques, unsupervised ranking, Poisson process, similarity cascade

1. INTRODUCTION

Web topic detection is a practical requirement of many today's applications such as information retrieval and monitoring. It provides a core technology to organize, understand and analyze the new and interesting trends happening on web.

Web topic detection is totally different from the traditional topic detection and tracking (TDT) that aims at finding trends on news [1]. The reason is various: the emergence of new way to transmit information (e.g., twitter, facebook, blog), information cascade behaviors on web (e.g., "abandon their own information in favor of inferences based on earlier people's action" [2]), etc. Moreover, supervised information is rare and expensive to identify the real topics.

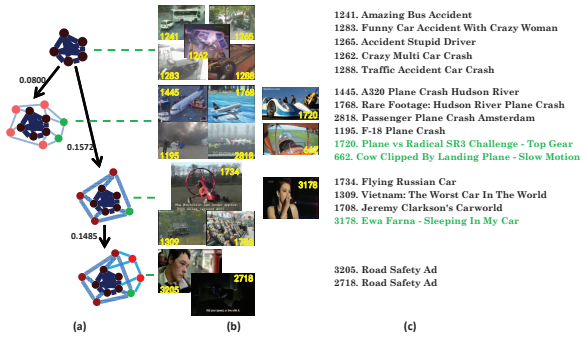

Fig. 1. An illustration of 4 detected topics on MCG-WEBV (False positives are illustrated in green color). In (a), the nodes and the lines represent the objects and their similarity separately. The darker the lines are, the higher similarities they are. (b) shows the new objects identified at a new layer. (c) gives the corresponding video titles.

Although some existing works have been aware of these differences, the solution has been quite unanimous: a web topic is considered as a clustering. For example, many approaches are based on partition-based clustering [3] [4], such as Non-negative Matrix Factorization (NMF) [5]. However, the partition-based methods assign each object (in this paper, objects mean the elements on web, e.g. the texts, images, videos, etc.) to a topic, but according to our experience, each object may belong to several topics. For instance, "Oil production" can belong to both "Energy sources" and "Economy". Moreover, these ideas ignore the fact that topic is rare – enormous data have been updated on web, but only a few objects trigger the interests among people and further evolve into topics.

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, 61332016, 61202322, 61202234, 61303153, 61303154 and 61175115, by Municipal Natural Science Foundation of Beijing: 4132010 and KZ201310005006, and by China Postdoctoral Science Foundation: 2012M520436.

Motivations: This paper tries to discuss the nature of web topic detection in terms of similarity diffusion. To reduce the influence of other factors as small as possible, we get rid of the auxiliary characteristics on web, e.g., cross media, hyperlink, time stamp, etc, and try to answer two primitive questions with text information as a case study.

1)What is the universal scheme to generate topics on web?

Most of the related work is based on the "close" clustering assumption. Chen *et al.* [6] deal with web video topic detection via the closely interlinked tags. Zhang *et al.* [7] introduce graph shift (GS) [8] to find highly similar subgraphs as topics. However, as observed in our study, the "close" clustering assumption barely reflects the content shift phenomenon among web topics, where the random and latent factors tend to shift the content of topics. Fig. 1, for instance, illustrates 4 examples about how human beings organize web videos into topics. As is unexpected, "Car accident" gradually evolves into "Plane crash", "Car news" and "Road safety ads".

In this paper, we study the scheme to generate topics from a view of the evolution of information, i.e., the diffusion of similarity among objects. Based on this, the evolution of information is first formulated as similarity cascades, and then topic candidates are identified across the different cascade layers. A layer (LY) is a subset of objects where the similarity among each other is higher than a certain threshold. The multi-granularity topics across cascade layers naturally simulate the evolution. In this way, the union of candidates identified in different layers forms the topic candidate set.

2) How to identify which topic candidate is a real one?

It is a practical requirement to identify which one is real from a large number of candidates, due to the following two main reasons: 1) we do not know how many topics really form on web; 2) web topic is rare. However, previous works ignore this problem, and just output the topics with a defined number, e.g., 179 clusterings are considered as topics [7].

In this paper, we assume that the relative importance among candidates makes them as real topics. As a reasonable assumption, Poisson process is used to describe the similarity diffusion process. An unsupervised expectation maximization (EM) algorithm is introduced to rank topic candidates by calculating their weights. Candidates with higher weights can be considered as real topics than these with lower ones. The advantage of the unsupervised approach is that the ranking framework not only handles the rareness problem, but also deals with the unknown number of topics.

Contributions: To the best of our knowledge, this paper is the first to investigate the similarity diffusion for web topic detection, presenting a series of comprehensive experiments to illustrate the benefits of this novel viewpoint for web topic detection. The proposed method is simple, yet remarkably powerful. Simply by finding maximal cliques to represent topics, with an unsupervised ranking, we develop a web topic detection method that outperforms the state-of-the-arts.

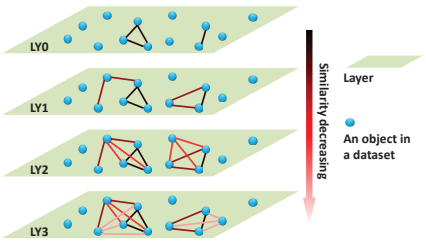

Fig. 2. An illustration of similarity cascades. A node corresponds to a text, an image, a video or any object on web.

2. GENERATING TOPIC CANDIDATES

2.1. Representing Correlations as Graph

Given a dataset, we establish a graph to represent the correlations among the objects. The nodes v_i represent objects in a dataset, and edges e_{ij} between two nodes v_i and v_j denote the correlations. Any similarity between two objects can be used to represent the correlation into a graph (V, E), where $V = \{v_i\}$ and $E = \{e_{ij}\}$. Given any two nodes represented as feature vectors h_i and h_j, this paper uses the normalized histogram intersection (NHI) to measure the similarity as edge $G(i, j) = e_{ij}$:

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

where $h_i(k)$ and $h_j(k)$ are the k-th bins of the histogram h_i and h_j, respectively.

Although NHI similarity is relatively simple, the primary motivation is to discover an universal scheme to grasp topics with cues as few as possible. Moreover, graph is a general tool to represent the correlation among objects. For instance, the canonical correlation analysis (CCA) between heterogenous media, the learned metrics, or hyperlinks can also be simultaneously incorporated into graph or hypergraph [9].

2.2. Similarity Cascade and Multi-granularity

Although the diffusion of information can be observed at every time stamp, the time stamp can not directly indicate the topics. In contrast, similarity is a general cue to find topics. Fig. 2 illustrates the similarity diffusion process: a topic absorbs more objects at a lower similarity level, or a topic absorbs different objects at different similarity levels to form different topics, and then this process recursively happens in or across different similarity layers. We call this process as similarity cascade (SC).

SC is totally different from information cascade [10] where the decision of people follows the other's options, almost being independent of their own private information. In contrast, SC assumes that the evolution of topics has been driven by weak or even random correlation among objects.

Multiple topics evolve into multi-granularity over a series of layers in SC. When a topic C propagates over SC, it leaves a trace, a tree-like cascade, in the form of a set of tuple $(C, V_+)_l$ which means that the topic C absorbs a set of nodes V_+ above the similarity l. If we denote the fact that the cascade initially starts from some active topic C at the level l_0 as $(C^{l_0}, V_+)_{l_0}$, two-granularity topics across SCs are represented as:

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

where $C^{l_{t+1}} = (C^{l_t}, V_+^{l_t})_{l_t}$ is the topic at the layer above the similarity l_{t+1}. Therefore, multi-granularity topics can be represented as $\{C^{l_t}, t = 0, 1, \cdots, T\}$ across a set of layers in SC. For instance, Fig. 2 illustrates that topics propagate from LY0 to LY3. Now, we only observe that topics evolve across SCs but do not know how they propagate, and we only know that V_+ is absorbed by topics but do not know which V_+ is involved. Therefore, given an initial topic, we aim to recover the diffused topics and understand this process.

2.2.1. Similarity Diffusion Process

The topic is represented by the indicator vector b_k, $b_k \in \Delta^N$ where $\Delta = \{0, 1\}$, and $k = 1, 2, \ldots, K$ is the indicator for the k-th topic. The i-th bin of the b_k is denoted as b_{ki}, where $b_{ki} = 1$ or 0 means that the topic k whether contains the i-th node or not. The k-th topic can be represented as subgraph:

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

Therefore, a similarity-preserving topic can be approximately represented as μC_k, where the μ is an relative weight to indicate the similarity among the nodes.

Poisson process is used to describe the similarity diffusion process among topics:

$$SG_i = \text{Poisson}\,(\mu C_k). \tag{4}$$

(4) is a reasonable assumption, topics are independently and randomly focused by people in a time interval.

2.2.2. Generating Topics Across SCs

SCs can be simulated at different similarity levels from the minimal value of similarity to the maximal value. Given a similarity graph G built by any method, all the unique values in graph are theoretically considered as possible thresholds, i.e., $\mathcal{L} = \{l_0, l_1, \ldots, l_T\} = \text{unique}(\{G(i,j)\})$, $i, j = 1, 2, \ldots, N$, where T is the number of thresholds, N is the number of nodes in a graph. At each LY, topic candidates are generated by any algorithm, and all topic candidates identified from all LYs are merged into a candidate set (see Alg. 1).

2.3. Maximal Cliques as Clustering-like Pattern

In this paper, rather than using "close" clustering patterns [7], we adopt the "loose" pattern, i.e., maximal clique (MC), to represent topics. The reasons to use loose pattern have twofolds: 1) it is difficult to design a perfect clustering pattern for all types of topics; 2) as observed, the objects in topics, especially at lower layers in SCs, tend to be weakly correlated.

Algorithm 1: Generate topic candidates across SCs

Input: A graph G, thresholds $\mathcal{L} = \{l_1, l_2, \cdots, l_T\}$
Output: Topic candidates $\mathcal{C}$

Initialize $\mathcal{C} = \varnothing$;
for *each* $l \in \{l_0, l_1, \cdots, l_T\}$ **do**
  $$SG^l(i,j) = \begin{cases} G(i,j), & if\ \ G(i,j) \geqslant l \\ 0, & if\ \ G(i,j) < l \end{cases}$$
  $\mathcal{C}^t$ = all candidates output by any algorithm on the SG^l;
  $\mathcal{C} = \mathcal{C} \cup \mathcal{C}^t$;
end

Bron-Kerbosch algorithm [11] is a classical one to find MCs, but it is time consuming. Although some fast algorithms [12] [13] are proposed, it is still too slow in our application scenario, as we try to find all MCs from all layers in SC. So how can we overcome the problem when the fact that multi-granularity topics generated across SCs is available? Next, we propose an accelerated algorithm by identifying meaningful topic candidates from subgraph.

2.3.1. An Accelerated Algorithm (AA)

The multi-granularity topics are formed at different layers with a same "seed" topic. Based on this observation, many candidates that do not include the "seed" topics have been safely filtered out. First, the similarity between an object and a topic is defined in Defination 1.

Definition 1. *Given a node v and a topic $C (x \in C)$, the similarity between v and the topic C, sim$(v, C) = \min_{x \in C}$ (sim(v, x)), where sim(A, B) is any similarity function.*

The accelerated algorithm is proposed to generate meaningful candidates with three steps. Firstly, the AA algorithm divides the thresholds $\mathcal{L}$ into two sets $\mathcal{L}^L$ and $\mathcal{L}^S$ by a parameter τ. That is, $\mathcal{L}^L = \{l_t \in \mathcal{L} | l_t \in \tau\%$ highest values of $\mathcal{L}\}$, $\mathcal{L}^S = \mathcal{L} \setminus \mathcal{L}^L$. Secondly, Alg. 1 is used at $\mathcal{L}^L$ to identify the "seed" topic candidates that are denoted by $\mathcal{C}^L$. Given a "seed" topic $\mathcal{C}_0^L$, for each low level $\mathcal{L}_t^S$, we find a candidate node set $I = \arg \underset{v \in V(G)}{\text{sim}} \left(v, \mathcal{C}_0^L\right) \geqslant \mathcal{L}_t^S$ and extract the subgraph SG induced by I. Thirdly, Alg. 1 is run on SG at $\mathcal{L}_t^S$ to identify new topic candidates at this layer. Fig. 3 explains this procedure on a toy example. A "seed" topic $\{2,3,4\}$ is first identified in Fig. 3(a), and then Fig. 3(b) finds $I = \{1,2,3,4,5,6\}$, and the induced subgraph (SG). Fig. 3(c) shows the expansion of this topic, i.e., $\{1,2,3,4\}$ and $\{2,3,4,5,6\}$, which are identified from SG, include the "seed" topic $\{2,3,4\}$, and are also maximal cliques of G. Moreover,

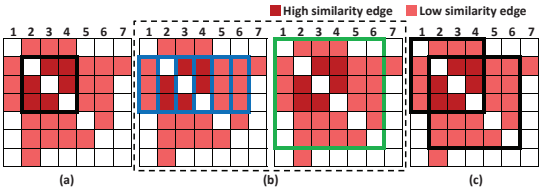

Fig. 3. An illustration of the AA algorithm on a 7-node graph with two similarity levels. (a) MC (shown in black box) is identified as the "seed" topic at the high threshold. (b) Set I is shown in blue box, and the extracted subgraph SG is shown in green box. (c)Identifies all MCs from SG at low threshold (shown in black box).

$\{2,7\}$ is a maximal clique at low threshold but has been filtered out.

2.3.2. Theoretical Justification

In this subsection, we theoretically justify that all candidates output by the AA algorithm are meaningful.

Theorem 1. [1] *$\forall C$ identified from SG, whose corresponding "seed" topic is C^L , we have: 1) $C^L \subseteq C$, 2) C is a maximal clique of the graph G.*

The Theorem 1 theoretically justifies that although the MCs are identified from SG, they are also maximal cliques of the graph G. More specially, the AA algorithm can find the weakly correlated objects, and many topic candidates which don't include "seed" topics will be filtered out. It can speed up greatly to find candidates, and also decrease the number of topic candidates sharply.

3. UNSUPERVISED RANKING TOPIC CANDIDATES

This section introduces an unsupervised method to rank topic candidates to handle the rareness problem, as a large number of candidates can be generated across SCs.

3.1. Ranking As Deconvolution

A reasonable assumption for topic detection is that the larger weight a topic candidate has, the higher probability that it is a real topic. Therefore, the unsupervised ranking can be formulated as the graph reconstruction task [14]:

$$G = \text{Poisson}\left(\sum_{k=1}^{K} \mu_k C_k\right). \tag{5}$$

Given the topic candidates C_k, an EM algorithm is adopted to estimate the weights μ_k. This algorithm consistently estimates the weights by filtering out the weights associated with the unnecessary candidates.

$$\begin{aligned} \text{E-step:} \quad & P_{k,ij}^{t+1} = \mu_k^t \frac{G_{ij} C_{k,ij}}{Q_k \sum_m \mu_m^t C_{m,ij}}, \\ \text{M-step:} \quad & \mu^{t+1} = \textstyle\sum_{ij} P_{k,ij}^{t+1}. \end{aligned} \tag{6}$$

[1] We omit the proof because of the limited space.

where $Q_k = \sum_{ij} C_{k,ij}$, $\mu_k^0 = 1$, $k = 1, 2, \cdots, K$ and the iteration is terminated when $|\mu^{t+1} - \mu^t| < \varepsilon$. This is the Richardson-Lucy algorithm [15] for Poisson deconvolution. It originally aims at recovering images that were blurred by a point spread function.

4. EXPERIMENT AND DISCUSSION

In the experiments, we evaluate our method on two diverse datasets, MCG-WEBV [16] and YKS [7]. MCG-WEBV is downloaded from the "Most viewed" videos of "This month" on *YouTube*. While YKS is a cross media dataset crawled from *YouKu* and *Sina*, respectively, but we only use news articles on YKS in the following experiments. Table 1 lists the differences between two datasets.

For the MCG-WEBV dataset, we consider the surrounding text of each video as a set of words, and filter out the stopwords. While YKS is tokenized by *NLTK* package in Python as pre-processing, and then the TF-IDF is used to measure the importance of each word. Finally, the bag-of-words (BoW) is used to represent these text cues into features.

Following the common baselines, we evaluate the performance with Top-10 F_1 score. That is, every detected topic D_t is matched with its most similar groundtruth topic, and then the top-10 F_1 scores are averaged to measure the performance [7]: $F_1 = \frac{2 \times Pr \times Re}{Pr + Re}$, where $Pr = \frac{|D_t \cap G_t|}{|D_t|}$, $Re = \frac{|D_t \cap G_t|}{|G_t|}$, D_t is a detected topic, G_t is a groundtruth topic, and $|\cdot|$ denotes the number of objects in a topic.

4.1. Analysis of Our Approach

We first use MCG-WEBV to demonstrate the effectiveness of different components in our method, i.e., maximal cliques to represent topics, generating topic candidates across SCs, the AA algorithm and the unsupervised ranking.

4.1.1. The Analysis of Maximal Cliques

In our experiments, NMF and GS are used as partition-based method and "close" clustering method to compare with MCs. Since the input of NMF [5] is a document-term matrix, it's meaningless to threshold it at different LYs. Therefore, the number of topics in NMF is assigned with a series of numbers from 50 to 1000 with the step size 20, and then all generated topics are merged into a candidate set. While GS and MC are provided a series of similarity thresholds $\mathcal{L} = [0.05 : 0.03 : 0.3, 0.3 : 0.1 : 1]$[2]. After the candidates are generated by these algorithms, the unsupervised ranking (6) is Applied. 73 highest weighted topics are evaluated with the groundtruth, as 73 is the number of topics in the groundtruth.

As shown in Table 2, the performance of MC is much better than GS, and is slightly better than NMF. As is expected,

[2] It's a Matlab expression.

Table 1. A comparison between MCG-WEBC and YKS.

Dataset	#Topics	#Objects	#Objects in topics	Comments (the cues used in our experiments are indicated in bold.)
MCG-WEBV	73	3282	832	Videos and their surrounding **titles, tags and descriptions** on *Youtube* from Dec 2008 to Feb 2009
YKS	298	7325	990	**News articles on *Sina*** and Web videos on *YouKu* from May 2012 to June 2012

Table 2. The results of different representation of topics.

Alg.	NMF	GS	MC
Pr	0.957	0.953	0.967
Re	0.943	0.781	0.948
Top-10 F_1	0.945	0.852	**0.953**

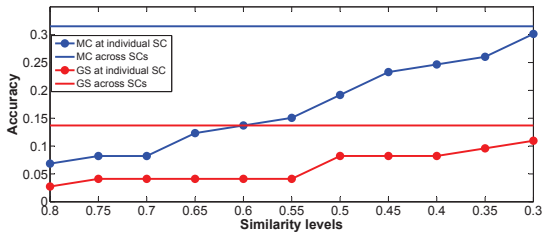

Fig. 4. The effectiveness of the SC. Each node shows the accuracy at the corresponding similarity layer, and horizontal line means the accuracy across all layers in the SC.

NMF is a partition-based method that is difficult to handle the rareness of topics. On the other hand, GS fails to handle weakly correlated topics.

4.1.2. The Effectiveness of SCs

In order to demonstrate the effectiveness of SCs, we use 11 layers with the thresholds $\mathcal{L}$ ranging from 0.3 to 0.8, and identify topic candidates in or across layers. To evaluate the effectiveness of SCs, the detection accuracy is defined as:

$$\text{Accuracy} = \frac{\#\text{Successful}}{\#\text{Groundtruth}}, \tag{7}$$

where a topic candidate D_t is recognized as successful detection, if the matching ratio satisfies that, $r = \frac{|D_t \cap G_t|}{|D_t \cup G_t|} \geqslant 0.5$.

Although both GS and MC curves increase when the similarity level is decreased, the performance of MC increases faster than GS at the low similarity level (see Fig. 4). It means that only a few topics have "close" correlation pattern at high level similarity, and the low similarity level can handle the "loose" correlation in topics. On the other hand, the accuracy across SCs is consistently higher than that at each individual layer, i.e., MC increases 4.55% accuracy rate while GS increases 25.00%. Obviously, detected topics across SCs can handle the large inter-topic variations.

4.1.3. The Performance of the Accelerated Algorithm

In this subsection, we compare the performances of the AA algorithm with the Alg. 1 on three aspects, i.e., running time, the number of topic candidates ($\#\mathcal{C}$) and Top-10 F_1. In Alg. 1, all topic candidates are identified across SCs with the threshold $\mathcal{L}$. While in AA, we first divide $\mathcal{L}$ into two subsets by setting $\tau = 0.03$.

Table 3. A comparison between two algorithms.

Alg.	Alg. 1	AA
time(day)	11	**0.8**
$\#\mathcal{C}$	623488	**13646**
Top-10 F_1	**0.969**	0.953

Table 4. Effectiveness of unsupervised ranking.

Alg.	No rank	Rank
Pr	0.790±0.018	0.967
Re	0.580±0.021	0.948
Top-10 F_1	0.654±0.013	**0.953**

Table 5. A comparison between MMG and our method.

Datasets	MCG-WEBV		YKS	
Alg.	MMG	Our	MMG	Our
Pr	0.937	0.967	0.975	1.00
Re	0.942	0.948	0.936	0.960
Top-10 F_1	0.937	**0.953**	0.952	**0.979**

Table 3 shows clearly that the AA algorithm has greatly reduced the time of identifying meaningful candidates, because we only find the "seed"-topic-related candidates from subgraph. On the other hand, comparing with the Alg. 1, the performance of the accelerated algorithm only slightly decreases. These results indicate that the AA algorithm may filter out some meaningful candidates, but the multi-granularity topics are very universal in topic detection.

4.1.4. The Effectiveness of Unsupervised Ranking

We select 73 highest weighted topic candidates as ranked results. For unranked algorithm, we randomly select 73 topics from all 13646 topic candidates as detected topics. The random selection process is repeated 5 times and their mean and standard deviation is calculated. The comparison between ranked and unranked algorithm is shown in Table 4.

As is expected, the Top-10 F_1 score has been improved about 45% after unsupervised ranking. Ranking successfully returns more meaningful topics, which justifies the correctness of our hypothesis: the similarity diffusion satisfies Poisson process.

4.2. Comparisons With Other Algorithms

In this subsection, we compare our method with other algorithms. Table 5 shows that our method performs better than MMG [7], which uses multiple cues (including texts, videos and time stamp), and is the state-of-the-art on both datasets.

Top-10 F_1 ignores the influence of the number of detected topics. Therefore, Fig. 5 shows the performance of our

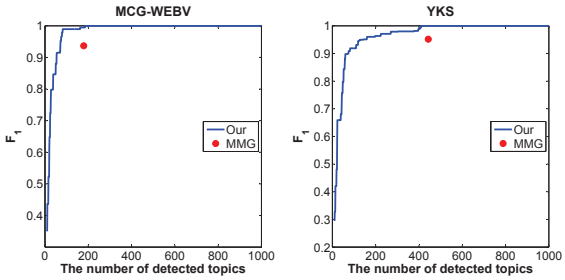

Fig. 5. Comparisons between MMG and our method on two datasets with average Top-10 F_1.

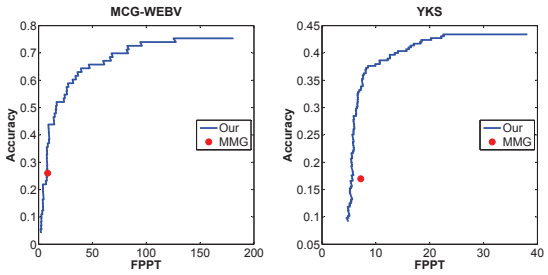

Fig. 6. Comparisons between MMG and our method on two datasets with accuracy v.s. FPPT.

method at different number of detected topics. Fig. 5 illustrates that the curves increase quickly – our method can accurately retrieval the top-10 topics at cost of the slightly increased number of topic candidates. Moreover, at the same number of detected topics (179 for MCG-WEBV and 443 for YKS), our method outperforms MMG obviously.

4.2.1. Evaluations With New Measurement

Top-10 F_1 ignores the rareness problem in web topic detection – top-10 makes it impossible to measure the number of false positives. We therefore propose a more reasonable evaluation method, accuracy v.s. false positives per topic (FPPT): if a topic is successfully detected, how many number of false positives are introduced by detection systems. Accuracy is defined as (7) and FPPT $= \frac{\#\text{Detected}-\#\text{Successful}}{\#\text{Successful}}$, where the detected topic with matching ratio $r_t \geqslant 0.5$ is considered as successful detected topic. The closer to the upper-left corner the curves are, the better the performances are (see Fig. 6).

Fig. 6 shows the accuracy v.s. FPPT curves of MMG and our method. The curve shows that our algorithm is consistently better than MMG [3] on both datasets.

5. CONCLUSIONS

In this paper, we have described a method based on representing web topics as a set of "loose" correlated objects, similarity diffusion among topic candidates, and the results outperform the state-of-the-arts. Moreover, a new benchmark on the topic-wise level evaluation is proposed to describe the rareness problem in web topic detection.

The promising results of this paper motivate a further examination. First, more efficient "loose" patterns, like random methods, may scale up well to large-scale problems over MCs used here. Furthermore, the heterogenous cues should be embedded into graph, as is currently being investigated in multimedia community.

[3]Note that the results of MMG in the previous experiments are directly copied from their paper. While in this experiment we run the source code supplied by the authors [7] with the same setting, as the reported results can not be compared with the new measurement.

6. 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] D. Easley and J. Kleinberg, *Networks, Crowds, and markets - Reasoning about a highly conected world*, Cambridge University Press, 2010.

[3] A. Saha and V. Sindhwani, "Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization," in *WSDM*, 2012, pp. 693–702.

[4] D. Zhang, C. Lin, S. Chang, and J. Smith, "Semantic video clustering across sources using bipartite spectral clustering," in *ICME*, 2004, pp. 117–120.

[5] D. Lee and H. Seung, "Learning the parts of objects using non-negative matrix factorizations," *Nature*, 1999.

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

[7] Y. Zhang, G. Li, L. Chu, S. Wang, W. Zhang, and Q. Huang, "Cross-media topic detection: a multi-modality fusion framework," in *ICME*, 2013, pp. 1–6.

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

[9] A. O'Sullivan, N. Adams, and I. Rezek, "Canonical correlation analysis for detecting changes in network structure," in *ICDM Workshops*, 2012, pp. 250–257.

[10] S. Bikhchandani, D. Hirshleifer, and I. Welch, "A theory of fads, fashion, custom, and cultural change in informational cascades," *Journal of Political Economy*, vol. 100, no. 5, pp. 992–1026, October 1992.

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

[12] E. Tomita, A. Tanaka, and H. Takahashi, "The worst-case time complexity for generating all maximal cliques and computational experiments," *Theoretical Computer Science*, vol. 363, no. 1, pp. 28–42, 2006.

[13] K. Makino and T. Uno, "New algorithms for enumerating all maximal cliques," in *SWAT*, 2004, pp. 260–272.

[14] H. Soufiani and E. Airoldi, "Graphlet decomposition of a weighted network," *Journal of Machine Learning Research - Proceedings Track*, vol. 22, pp. 54–63, 2012.

[15] W. H. Richardson, "Bayesian-based iterative method of image restoration," *Journal of the Optical Society of America (1917-1983)*, vol. 62, pp. 55–59, Jan. 1972.

[16] 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.