What's Hot in The Theme: Query Dependent Emerging Topic Extraction from Social Streams

Yuki Endo, Hiroyuki Toda and Yoshimasa Koike
NTT service evolution laboratories, NTT Corporation
Yokosuka, Kanagawa, Japan
{endo.yuki, toda.hiroyuki, koike.y}@lab.ntt.co.jp

ABSTRACT

Analyzing emerging topics from social media enables users to overview social movement and several web services to adopt current trends. Although existing studies mainly focus on extracting global emerging topics, efficient extraction of local ones related to a specific theme is still a challenging and unavoidable problem in social media analysis. We focus on extracting emerging social topics related to user-specified query words, and propose an extraction framework that uses a non-negative matrix factorization (NMF) modified for detecting temporal concentration and reducing noises. We conduct preliminary experiments for verifying our method using a Twitter dataset.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

Keywords

Social media; local clustering; NMF; topic models

1. INTRODUCTION

Topic modeling for emerging topics in social streams (e.g. Twitter and Weibo) plays a crucial role in the fields of recommendation [5] and social analysis. SocialTransfer [5] was recently proposed as an effective application for trend-aware video recommendation, which learns topic modeling from social text streams and then uses the topics as a bridge between the social and the video domains. For such applications, people desire to promptly analyze the emerging topics related to a specific theme. For example, video service providers desire to offer users contents related to social hot topics that match individual viewing histories such as sports and animation to promote utilization of many contents.

Although several studies focus on extracting global emerging topics [4][6], these are not designed for local ones related to a specific theme. Additionally, applying these approaches to a whole dataset takes a lot of computational time for optimization. Chen et al. [3] proposed a system that can extract specific emerging topics for organizations. They use an SVM classifier to collect relevant data in a supervised manner and cluster the collected data incrementally to extract emerging topics. The system requires training sets including documents posted by known accounts of organizations for building a classifier. However, searching for relevant accounts takes a certain time and such accounts may not always exist. Moreover, performance of the data collection will decrease with time because supervised training is done with past data while social streams are time varying data.

To solve the above problems, this paper proposes an unsupervised framework for efficiently extracting emerging topics related to user-specified query words from social streams. Especially, we propose a NMF technique modified for detecting temporal concentration and reducing noises.

2. PROPOSED FRAMEWORK

Figure 1 shows an overview of our framework that consists of the following two steps:

Step 1: Given query words, the system broadly and efficiently collects documents related to the query words from social streams in a predefined time window.

Step 2: The system then learns topic modeling based on NMF from the collected data.

In the step 1, we collect documents following the idea of pseudo-relevance feedback for query expansion [2], which assumes that the top-ranked documents in the initial search results contain good topic-related words that help discriminate relevant documents from irrelevant ones. In practice, we adopt EvoCut [1] that is a local graph clustering algorithm that can efficiently find subgraphs including given seed nodes without looking at the whole graph. We construct a graph that represents relationships between documents on the basis of word co-occurances and use documents including query words as seed nodes. The graph is updated online every time a document is posted.

2.1 Emerging topic extraction

In the step 2, we consider that there are two requirements for extracting emerging topics: (1) temporal concentration detection and (2) noise reduction. For (1), we adopt an existing NMF based method with a temporal regularizer [6]. For (2), we extend the existing method to more precisely extract relevant emerging topics. This is because the broadly collected data may contain irrelevant noisy documents since there are various themes in the social media. Specifically, we propose a query dependent regularizer and modify the

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742727.

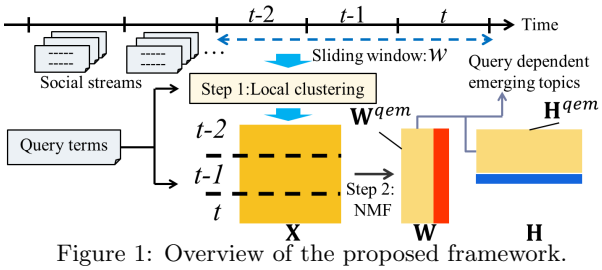

Figure 1: Overview of the proposed framework.

existing NMF based method. As shown in Figure 1, our key idea is to decompose a matrix into two matrices that consist of emerging topics related to queries and the other irrelevant noisy topics using the two regularizers.

Given query words and a document-word matrix $\mathbf{X}(t-w+1:t) \in \mathbb{R}^{m \times n}$ at any given time-point t over the sliding window w, our matrix factorization is formulated as the problem of minimizing the following objective functions:

$$\begin{aligned} &||\mathbf{X}(t-w+1:t) - \mathbf{WH}||_F^2 \\ &+\lambda_q \sum_{i}^{k_{qem}} \sum_{j}^{n} \max(0, \mathbf{Q}_{i,j} - \mathbf{H}_{i,j}^{qem})^2 \\ &+\lambda_t \sum_{\mathbf{w}_j \in \mathbf{W}^{qem}} \sum_{i=1}^{T-1} c_i \max(0, \nu - (\mathbf{DFSw}_j)_i)^2. \end{aligned} \quad (1)$$

The first term is a typical NMF formulation based on Frobenius norm, in which $\mathbf{W} \in \mathbb{R}_{\geq 0}^{m \times k}$ and $\mathbf{H} \in \mathbb{R}_{\geq 0}^{k \times n}$ are the factorized document-topic matrix and topic-word matrix respectively, and m, n and k are the the number of documents, words and topics respectively. The second term is a query constraint that facilitates the occurrence of query words and relevant words in specific topics by penalizing the k_{qem} topics in which query words do not occur. $\mathbf{H}^{qem} \in \mathbb{R}_{\geq 0}^{k_{qem} \times n}$ is the partial matrix of $\mathbf{H}$ and $\mathbf{Q} \in \mathbb{R}_{\geq 0}^{k_{qem} \times n}$ is the query dependent topic-word matrix. The columns of $\mathbf{Q}$ that correspond to the query words are assigned constant η and the other columns are assigned 0. If the intensities of the query words in the k_{qem} topics are larger than η, the loss becomes 0. In addition, it is expected that the intensities of the relevant words with which the query words co-occur increase in the k_{qem} topics because topics are inferred by observing word co-occurrence based on the first term. The third term is a temporal constraint that penalizes the static or decaying topics. This constraint is similar to the existing method [6] and refer to their paper for more details, but the difference is that our method applies the temporal regularizer to only the partial matrix of $\mathbf{W}$, that is, $\mathbf{W}^{qem} \in \mathbb{R}_{\geq 0}^{m \times k_{qem}}$. This enables extracting precisely relevant emerging topics in the partial matrix and gathering irrelevant noisy topics in the other partial matrix. λ_q and λ_t are the hyper parameters and empirically determined. We solve $\mathbf{W}$ and $\mathbf{H}$ in the equation (1) alternately using a L-BFGS-B method.

3. EXPERIMENTS AND RESULTS

We use a Japanese tweet dataset collected by Twitter API. To evaluate our approach, We uses Olympic as the query word and extract emerging topics related to Olympic. We prepares a dataset that consists of tweets related to Olympic and irrelevant ones from 2012/7/30 to 2012/8/5. For this, we annotated the tweets with labels related to the topics of Olympic (e.g. judo and swimming) using manually selected hashtags, and annotated the other tweets with the irrelevant label in the same way. Afterwards, the hashtags were removed from the tweets themselves. As a result, 57,306 Olympic tweets and 329,238 irrelevant ones are obtained.

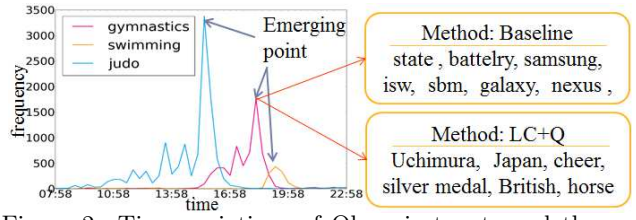

Figure 2: Time variations of Olympic tweets and the extracted topics at the most emerging point.

Table 1: The JSDs of emerging topic extraction at the most emerging points of each theme.

method	judo	swimming	gymnastics	average
Baseline	0.212	0.212	0.228	0.217
LC	0.178	0.195	0.235	0.203
LC+Q	0.106	0.100	0.113	0.106

As a quantitative evaluation metric, we use Jensen-Shannon divergence (JSD) [6] to compute errors between word distributions of true topics and extracted topics. Each True topics is estimated using the standard NMF from the tweets with each corresponding label by assuming that tweets with the same label have the same topics. When the number of tweets of each theme (e.g. judo) increases, a small JSD means that the emerging topics are precisely detected. Refer to the paper [6] for more information of the metric.

In Figure 2, the left side shows the time variations of frequency of tweets related to the Olympic themes. Table 1 shows the JSDs at the most emerging points of each theme by the several methods, where the baseline denotes NMF with temporal regularization [6] using a whole data, LC denotes the same algorithm as the baseline but input of NMF is data collected using the local clustering (step 1), and LC+Q denotes our total framework (step 1 and step 2). The JSDs of LC are sometimes equal to or lower than those of the baseline due to the noises in the collected data. On the other hand, LC+Q achieves better performance than the other methods. The right side in Figure 2 shows the examples of the extracted gymnastic topics obtained by each method at the most emerging point. Here, the most related topics are selected and the most related words are selected from the top 15 words with high probability in the topics. LC+Q can detect a local topic related to Olympic at the emerging point while the baseline extracts global topics that are dominant themes in the time point. For computational time, the baseline took 1388 sec, whereas LC+Q took 0.8 sec for data collection and 65.4 sec for NMF on a Xeon2.66GHz CPU.

4. REFERENCES

[1] Andersen, R. and Peres, Y.: Finding sparse cuts locally using evolving sets, In STOC'09, pp.235-244, 2009.
[2] Buckley, C., Salton, G. and Allan, J.: Automatic retrieval with locality information using SMART, In the 1st Text Retrieval Conference (TREC-1), pp.59-72, 1992.
[3] Chen, Y., Amiri, H., Li, Z. and Chua, T.-S.: Emerging topic detection for organizations from microblogs, In SIGIR'13, pp.43-52, 2013.
[4] Diao, Q., Jiang, J., Zhu, F., Lim, E.-P.: Finding bursty topics from microblogs, In ACL'12, pp.536-444, 2012.
[5] Roy, S. D., Mei, T., Zeng, W. and Li, S.: SocialTransfer: cross-domain transfer learning from social streams for media applications. In ACM MM'12, pp.649-658, 2012.
[6] Saha, A. and Sindhwani, V.: Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In WSDM'12, pp.693-702, 2012.