ABSTRACT
With the overwhelming information from social media networks and news portals, it is crucial to provide users a complete package of visual and textual information with popular interests automatically. To this concern, we present a news detection and pushing system, called Me-Digger (Multimedia News Digger), which not only effectively detects emerging topics from social streams but also provides the corresponding information in multiple modalities. Me-digger is the first systematic effort to leverage three sources of data, that is, Twitter, Flickr and Google news, to output with vivid visual and textual contents on emerging topics. Enabled by a novel general-structured high-order co-clustering approach, it has a more accurate detection of emerging topics compared to the existing methods on micro-blog social streams.

Categories and Subject Descriptors
H.4.0 [INFORMATION SYSTEMS APPLICATIONS]: General

Keywords
Co-clustering, Cross media, Emerging topic detection

1. INTRODUCTION
Nowadays, people are more and more used to receiving news from social media networks and news portals. However, due to the overwhelming information, it costs much time to seek the information related on the emerging topics from those websites. Digg receives a huge success as it dynamically shows the most popular news voted by readers at the front page. However, it heavily relies on the users’ contribution. Some efforts on emerging topic detection in social media streams are constrained by the short message length limitation in micro-blog type media like Twitter [2]. Furthermore, existing industrial and academic works are developed upon processing data from single media source, while users prefer to receive comprehensive information from multiple media sources on one stop, such as published news, images, and the public opinions. To these concerns, we design a novel news detection and pushing system, called Me-Digger, with following properties: (1) Auto detection of emerging topics on Twitter without requirement of user intervention; (2) Accurate keyword detection without limitation on the message length; (3) Auto compilation of multi-modality information from Google news, Flickr and Twitter on emerging topics.

Figure 1: The framework of Me-digger.

Figure 1 shows the framework of Me-digger with four parts: Data Collection, Emerging Keywords Extraction, Emerging Topic Detection, and Cross Media Retrieval. The most challenging part is Emerging Topic Detection, in which we enrich the correlation between detected Twitter keywords by introducing Google news and Flickr images. The proposed General-Structured High-Order Co-Clustering approach (GS-HOCC), is to detect emerging topics on Twitter by simultaneously co-clustering Twitter emerging keywords as well as Google news and Flickr images.

2. GENERAL-STRUCTURED HIGH-ORDER CO-CLUSTERING APPROACH
As aforementioned, similarities between the detected Twitter keywords, which is the key to detect the emerging topics, are hardly revealed due to message length limitation. Thus, we introduce Google news and Flickr as complementary to Twitter, and simultaneously co-cluster them respectively.

2.1 Data from Twitter, Flickr and Google News
In k-th time interval, let detected emerging keywords be \(w \in D_W\), Flickr data be \(f = (f^t, f') \in D_F\), and Google news data be \(n = (n^t, n') \in D_N\), where \(f^t\) and \(n^t\) indicate the text features, while \(f'\) and \(n'\) indicate the corresponding visual features. Let \(S_{WF}, S_{WN}, S_{WF}\) be textual similarities, and \(S_{WF}, S_{WN}, S_{WF}\) be the visual ones, the correlations among three sources are,

\[
C_{WF}(w, f) = S_{WF}(w, f), \quad C_{WN}(w, n) = S_{WN}(w, n),
\]

\[
C_{NF}(n, f) = \frac{\tau^t \delta(n^t, f^t) S_{NF}(n, f) + \tau' \delta(n', f') S_{NF}(n, f)}{\tau^t \delta(n^t, f^t) + \tau' \delta(n', f')}
\]

where \(\tau^t\) and \(\tau'\) are to balance two terms, and \(\delta(x, y) = 0\) when \(x = 0\) or \(y = 0\), otherwise \(\delta(x, y) = 1\). The joint probabilities \(P_{WN}, P_{NF}\) and \(P_{WF}\) can be calculated by,

\[
P_{WN} = \frac{C_{WN}}{\sum_{w,n} C_{WN}}, \quad P_{NF} = \frac{C_{NF}}{\sum_{n,f} C_{NF}}, \quad P_{WF} = \frac{C_{WF}}{\sum_{w,f} C_{WF}}.
\]
2.2 GS-HOCC

Different from star structure in traditional high-order co-clustering problem [4], three sources in our case are pairwise dependent. [3] proved that an optimal W-N co-clustering should be minimize the loss of mutual information as,

$$\arg \min_{W, N} \{I(W, N) - I(\hat{W}, \hat{N})\},$$

where $W, \hat{N}$ are the discrete random variables on the objective clusters $D_{\text{w}}, D_{\hat{N}}$, and $I(W, N)$ measures the amount of information shared between $W$ and $N$, that is

$$I(W, N) = \sum_w \sum_n p_{WN}(w, n) \log \frac{p_{WN}(w, n)}{p_W(w)p_N(n)}.$$

Since $W, N$ and $F$ are independent and identically distributed, the “global” optimization can be divided into three sub-problems: $W-N$ co-clustering, $N-F$ co-clustering and $W-F$ co-clustering. Then the “global” clustering can be evaluated as a linear combination of all the biclusters on correlated domains, shown in Eqn. (3).

$$\triangle I = \min_{\hat{W}, \hat{N}, \hat{F}} \alpha I(W, N) - I(\hat{W}, \hat{N}) + \beta I(N, F) - I(\hat{N}, \hat{F}) + \gamma I(W, F) - I(\hat{W}, \hat{F})$$

$$= \min_{\hat{W}, \hat{N}, \hat{F}} \alpha D(P_{WN}(W, N)||Q_{WN}(W, N))$$

$$+ \beta D(P_{NF}(N, F)||Q_{NF}(N, F)) + \gamma D(P_{WF}(W, F)||Q_{WF}(W, F)),$$

where $\alpha + \beta + \gamma = 1$, $D(\cdot; \cdot)$ is the KL divergence, and $Q_{WN}(W, N)$ is the distributions of the following forms [3]:

$$Q_{WN}(w, n) = P_{\hat{W}N}(\hat{w}, \hat{n})p(w|\hat{w})p(n|\hat{n}), \ w \in \hat{w}, n \in \hat{n},$$

so as $Q_{NF}(N, F)$ and $Q_{WF}(W, F)$.

3. MULTIMEDIA NEWS DIGGER

We introduce the other three modules in this section. (1) Data Collection Module is to collect data from three sources. The text modalities are preprocessed by following Porter’s stemming algorithm [5], and stop-words filter is conducted to remove stop words. TF-IDF is used to represent them. For image modality, each image is represented as a vector of Bag-of-Word features. (2) In Emerging Keywords Extraction Module, we utilize age theory to extract emerging keywords by modeling a life span of frequency for each keyword. By following [2], we assign high weight to the emerging keyword from Twitter. (3) Cross Media Retrieval Module works on retrieving the top relevant multi-modality samples from Twitter, Flickr and Google new to the detected emerging topics. The core method is manifold ranking.

4. EXPERIMENTS

The time period is chosen from 03/04/2011 to 03/31/2011, and we equally separate it into 4 time intervals. The baselines are Pair-wise Structure Co-clustering with Flickr data (PS-CC-F), Pair-wise Structure Co-clustering with Google news data (PS-CC-N), Start-structure High-order Co-clustering algorithm (SS-HOCC) and General-structure High-order Co-clustering only with text modality (GS-HOCC-T). The criteria is normalized discount cumulative gain (NDCG). “Top twitter trends each week” [1] provides the ground truth for the emerging topics in each week. We also ask for 5 participants to assign the top 10 detected emerging topics with the ranks of corresponding topics in the ground truth. The results are shown in Figure 2. Fig. 3 is the outputs of Me-digger in the second time interval. It shows that Me-digger provides attractive information on the top emerging topics.

5. CONCLUSIONS

In this paper, we have designed Me-digger, which is the first systematic attempt to push textual and visual information from Google news, Flickr and Twitter on automatically detected emerging topics. We believe this system will bring the customer experiences to a whole new level and achieve great commercial success.

6. ACKNOWLEDGEMENT

This work is supported in part by the National Natural Science Foundation of China (Grant No. 61003161), the China Post-Doctoral Science Foundation (Grant No. 2011M500430). The author also gratefully acknowledges the support of K.C. Wong Education Foundation.

7. REFERENCES


