This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

Sequential Summarization: a Full View of Twitter Trending Topics

Dehong Gao, Wenjie Li, Xiaoyan Cai, Renxian Zhang and You Ouyang

***Abstract*—As an information delivering platform, Twitter collects millions of tweets every day. However, some users, especially new users, often find it difficult to understand trending topics in Twitter when confronting the overwhelming and unorganized tweets. Existing work has attempted to provide a short snippet to explain a topic, but this only provides limited benefits and cannot satisfy the users' expectations. In this paper, we propose a new summarization task, namely sequential summarization, which aims to provide a serial of chronologically ordered short sub-summaries for a trending topic in order to provide a complete story about the development of the topic while retaining the order of information presentation. Different from the traditional summarization task, the numbers of sub-summaries for different topics are not fixed. Two approaches, i.e., stream-based and semantic-based approaches, are developed to detect the important subtopics within a trending topic. Then a short sub-summary is generated for each subtopic. In addition, we propose three new measures to evaluate the position-aware coverage, sequential novelty and sequence correlation of the system-generated summaries. The experimental results based on the proposed evaluation criteria have demonstrated the effectiveness of the proposed approaches.**

***Index Terms*—Twitter Trending Topics, Sequential Summarization, Subtopic Detection**

I. Introduction

Twitter, as a popular micro-blogging service, collects millions of real-time short text messages (up to 140 characters) every second. It acts as not only a public platform for posting trifles about users' daily lives, but also a public reporter for real-time

Manuscript submitted on January, 2013 and accepted on September, 2013.
The work described in this paper was supported by the Hong Kong RGC projects (No. PolyU 5230/08E, and PolyU 5202/12E) and a National Nature Science Foundation of China (NSFC No. 61272291).
Dehong Gao, Wenjie Li, and Renxian Zhang are with the Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (e-mail: {csdgao, cswjli}@ comp.polyu.edu.hk).
Xiaoyan Cai is with Northwest A&F University, Yangling Shaanxi, China.
You Ouyang was with the Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. He is now with Miaozhen Systems, Beijing, China.

news. It has shown the powerful ability in many public events, like the wildfires in San Diego and the earthquake in Japan [1]. Nevertheless, the side effect is that the individual users are usually sunk in the flood of millions of tweets.

To alleviate this problem, many applications have evolved from Twitter, like *echofon* (www.echofon.com), *whatthetrend* (whatthetrend.com), which provide services to explain why the term becomes a trending topic (topic for short hereafter in this paper) or to give a short description of the topic [2] [3] [4]. These systems generally track the topics in Twitter and use existing tweets or encourage users to edit a new tweet to explain the topics. For example, *whatthetrend* encourages users to edit explanatory tweets about topics. It ranks the submitted explanatory tweets by readers' agreements. These explanatory tweets can be regarded as tiny summaries about the topic, providing a good way to help users understand the topic. However, a short summary can only sketch the topic in a simple way. Some researchers [5] [6] attempt to aggregate several explanations into one long summary using traditional summarization approaches, but it still loses much useful information, such as the change of twitters' focus and the temporal information. A well generated traditional summary can reflect the overall picture of topic, but performs poorly in summarizing these temporal changes of the crowds' focus in Twitter. Note that the focus of tweets changes much more frequently than that of the traditional mainstream media. Considering this, we propose the following work. ***Sequential Summarization*** aims to generate a serial of chronologically related *sub-summaries* for a given topic while retaining the order of information presentation and each sub-summary attempts to concentrate on one theme or subtopic. The ***sequential summary***, made up by the chronologically ordered sub-summaries, shall provide the entire development of the topic, or the full view of the topic.

We expect that the whole sequential summary covers the entire development of the topic and in particular the sub-summaries can depict the important subtopics sequentially. We approach the sequential summarization with two sub-tasks: subtopic detection and sequential summary generation. To automatically detect the subtopics, we propose two approaches, the stream-based approach and the semantic-based approach. Typically, when a subtopic is popular enough to catch users' attentions, it will lead to a surge of a certain level in the tweet stream. According to this idea, we use the stream-based approach, the Offline Peak Area Detection (OPAD) algorithm [7], to locate such surges by tracing the volume changes of tweet streams. This approach regards the collections of tweets at such surge time interval as subtopics. The algorithm is easy to implement and intuitively works, but it may fail to handle the cases where some subtopics have their bursts at different time points in different countries due to the geographical or time-zone differences. This often makes some subtopics appear in more than one surge and/or mingle with other subtopics, especially when facing short-time topics. In order to locate each subtopic well, the semantic-based approach leverages Dynamic Topic Modeling (DTM) [8] to mine the dynamic subtopic shift in tweet streams. We use different strategies to score the significance of tweets corresponding to the stream-based and semantic-based subtopic detection. Sub-summaries are generated with the highest scored tweets for each detected subtopic. The experiment confirms the superiority of semantic-based approaches in handling the sequential tweet streams. During the experiment we observe that a proper subtopic number plays an important role in the semantic-based approach. Motivated by these observations, we further combine these two approaches. Recall that DTM in nature is a clustering approach which requires a pre-specified cluster number. When the cluster number is determined by the number of the peaks identified by the stream-based approach and only tweets within surges are used in DTM, the best performance is achieved.

To sum up, this paper makes the following contributions. (1) We set out a new summarization application, sequential summarization, which will be of great benefit for Twitter users to better understand trending topics. The techniques developed for this application are also valuable for other researches, such as information visualization, topic detection and tracking, etc. (2)In evaluation, we emphasize the temporal nature of the generated summaries and propose three measures including position-aware coverage, sequential novelty, and sequence correlation. These measures can also be applied to other time-related evaluations. (3)In experiment, we conclude that a reasonable subtopic number plays the pivotal role in sequential summarization and the crowding endorsement is an effective indicator for the importance of the tweet.

II. Related Work

A. Twitter summarization

Document summarization aims to produce a short synopsis for one or more documents which conveys the most important information stated in these documents. The early summarization researches mainly focus on formal text-based summarization, e.g., news articles in DUC (duc.nist.gov). Recognized as the chief driving force in Internet development, Web 2.0 plays a crucial role in spurring social network applications and generates new types of texts, like online discussion forums [9], blogosphere [10], and micro-blogs. This technique innovation further brings new research resources and new challenges for cyber-based text summarization.

The popularity of micro-blogging services, such as Twitter and *Plur*(www.plurk.com), has caught increasing attention from worldwide researchers. There exist some pioneering researches working on Twitter summarization. [2] explains Twitter topics by providing a list of significant terms. Users can utilize these terms to drill down to the tweets which are related to the topics. [3] and [4] attempt to provide a one-line summary for each topic using phrase reinforcement ranking. The relevance model employed by [11] generates summaries in a larger size, i.e., 250-word per summary, by synthesizing multiple high rank tweets. [5] further explores multiple text sources including tweets and web contents linked by tweets to discover real essential information for topics. The main objective of the above works is to explain the topics with a few sentences or just a few words. [6] generates summaries specially for sports topics. To use the proposed summHMM algorithm, one needs to estimate the parameters with manually annotated data first.

In practice, these annotated data are hard to obtain and the learned parameters may overfit the annotated data from a specific domain. [12] detects the main entities involved in the events and attempts to select significant tweets to describe the entities. However the temporal information is neglected when generating the descriptions. Another latest work has been conducted by [13]. They use spike detection algorithm to detect the important moments within sporting events like World Cup Soccer matches, which take place over a short period of time. But chances will be high that sub-events are within one spike in these short-period topics. It will be difficult to discriminate every sub-event happened in one spike. Our work resembles update summarization promoted by TAC (www.nist.gov/tac) which requires creating summaries with new information assuming that the reader has already read some previous documents about the topics. The update summarization is defined as: given two chronologically ordered documents sets about the same topic, the system is asked to generate two summaries, and the second one should inform the user of new information only. In order to achieve this goal, existing approaches mainly emphasize the novelty of the subsequent summary [14]. Different from update summarization, we focus more on the temporal change of topics. In particular, we need to automatically detect the time points at which we need to generate sub-summaries.

B. Event detection and tracking in social networks

Another line of the related work is the event detection and tracking. Several research efforts have focused on identifying events in social media, like Twitter [1] and Flickr [15].

To detect the event, [16] presents an algorithm based on locality-sensitive hashing to make the event detection feasible on the web-scale Twitter corpus. [17] implements a news processing system, called TwitterStand, to capture the latest breaking news in Twitter stream. [1] investigates the real-time interaction of events such as earthquakes in Japan and proposes an algorithm to detect the events in Twitter. [18] uses topically similar message clusters to distinguish between messages about real-world events and non-event messages. [19] applies the Newton's method to detect a point of change in the slope of a given function. The basic idea of this method is to monitor volume change in Twitter stream and detect the peaks (or spikes). [7] proposes a peak-finding algorithm based on Transmission Control Protocol (TCP) congestion detection which uses the weighted moving mean and variance to find the peaks in Twitter streams. The research on event detection and tracking usually relies on the corpus that contains a variety of topics and it aims to dig them out. In sequential summarization, we focus on how to fully reveal the development of topic.

III. Subtopic Detection

In sequential summarization, we assume that a topic is composed of a serial of chronological-ordered subtopics. We approach the sequential summarization by means of two models, i.e., subtopic detection and sequential summary generation. The goal of subtopic detection is to identify a serial of time-ordered tweet sets such that each tweet set represents a subtopic of the topic. Two approaches,

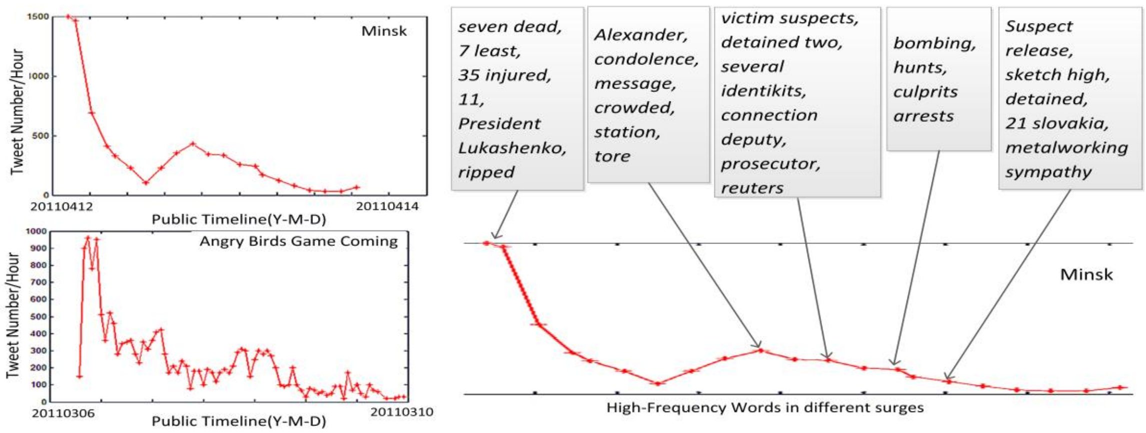

Fig. 1. (a) Surges in tweet stream and (b) high-frequency words in surges

i.e., the stream-based and semantic- based approaches, are proposed to achieve this goal.

A. *Stream-based subtopic detection*

The assumption behind the stream-based approach is that if a subtopic is popular enough to attract attentions from Twitter users, it often creates a certain level of surge in the tweet stream. To provide evidences of our assumption, we investigate two problems: 1) if the volume change appears in most tweet streams; 2) if there exit topic development or change of Twitter user focus in the surge areas.

Fig. 1 shows the tweet streams of two topics, "Minsk" and "Angry Birds Game Coming", and the high-frequency words in different areas of the tweet stream related to the topic "Minsk". For both topics, an explosive growth is clearly observed at the beginning of the tweet stream. The volume of the tweets then fluctuates up or down for a while, either dramatically or slightly, and gradually declines with a long tail. Fig. 1.(b) further provides potential explanations that trigger the volume changes and creates the serial of surges. For example, in the topic "*Minsk*", we can trace the development of "*Minsk Metro Blast*" from the updating number of casualties (from "*seven death at least*" to "*eleven reported by the present*" in the first set of keywords), to the investigation of explosion ("*detained two suspects*" in the second set), to hunting the culprits ("*culprits arrests*" in the third set), and to condolence and praying for the victims ("*sympathy*" in the following sets). From these observations, we draw the conclusion that in most cases tweets in surges can reflect subtopics under the topic to some degree.

Based on this idea, we propose the stream-based approach to detect the surges based on the volume change of the tweet stream and we use the tweets in these surges to represent the subtopics. We define the concept, **Peak Area** (PA) which represents the tweets within the time spans of each surge. The stream-based subtopic detection then is to identify these peak areas from the tweet stream. Let $S = \{s_1, s_2, .. s_n\}$ denote the tweet stream about a given topic and each tweet s_i is associated with a timestamp t_i which denotes the time when the tweet is created. For any two tweets s_i and s_j , if $i<j$, then $t_i<t_j$. The subtopic detection can be formalized as: g*iven the*

tweet stream S, detect a serial of peak areas $P = \{p_j^t \mid s_{t \in <t_j^s, t_j^e>} \in p_j^t, j = 1,..k\}$*, where* p_j^t *denotes a set of tweets* $s_{t \in <t_j^s, t_j^e>}$. t_j^s *and* t_j^e *denote the start and end time point of the peak area* p_j^t*, respectively. The number of tweets in* p_j^t *is significantly larger than the historical mean.*

Algorithm 1 in the next page describes the Offline Peak Area Detection (OPAD) algorithm. It is based on the idea of TCP congestion control, which uses a weighted moving mean and variance to determine if a new peak area appears [7]. We first divide the tweet stream *S* into a serial of time slices with the time window Δt . Initializing the mean and variance with the first time interval, we loop through the whole tweet stream. If the tweet number in the current interval is more than τ mean deviations and the tweet number is increasing (line 5), we regards there appears a new peak area (line 6). When the tweet number still increases (line 7), it performs up-hilling and updates the mean and variance (line 8). The peak area stops when the tweet number is back to the number of the starting time interval (line 10) or when a new peak area appears (line 11). In **Function** *Update* (line 28), π is set to 0.125 as in [7].

B. Semantic-based subtopic detection

Stream-based subtopic detection utilizes the volume change of the tweet stream to identify the subtopics. However, as mentioned before, due to the differences of geographical and time zones, Twitter may collect tweets about a same subtopic at different times. As a result, a subtopic may be separated into different peak areas or a peak area may contain a mix of more than one subtopic. To identify subtopics from the semantic perspective, semantic-based subtopic detection is proposed to employ Dynamic Topic Model [8] to capture subtopics in the tweet stream.

DTM, an extension of Latent Dirichlet Allocation (LDA) [20], is a generative model widely used in dynamic topic modeling researches. Given the tweet stream S and its vocabulary $W = \{w_1, w_2, ..w_m\}$, DTM assumes that a serial of latent topics $Z = \{z_1, z_2, ..z_k\}$ is hidden in the stream. Each tweet s_i is viewed as a mixture of topics in *Z*, and each topic z_j is a distribution over the word vocabulary *W*. Unlike LDA, DTM regards the topics evolve over time and supposes that the data is divided by a special time interval. The tweets in each time interval are modeled by *K*-component topic model, and the subtopic associated with the time interval *t* evolves from the subtopic associated with the time interval *t*-1. The graphical model for this generative process is shown in Fig. 2, in which α_t denotes the parameters of the per tweet subtopic distribution at time *t*, and θ_t and β_t denote subtopic and word distributions respectively. *M* and *N* denote the word and tweet numbers. With time moving, the k^{th} subtopic $\beta_{k,t}$ at the time interval *t* is smoothly evolved from the k^{th} subtopic $\beta_{k,t-1}$ at the time interval (*t*-1). The variational distribution of tweet-level latent variables can be defined as

$q(\theta_{1:T}, Z_{1:T} \mid \gamma_{1:T}, \phi_{1:T}) = \prod_{t=1}^{T} (\prod_{s=1}^{S} q(\theta_{t,s} \mid \gamma_{t,s}) \prod_{n=1}^{N} q(z_{t,s,n} \mid \phi_{t,s,n}))$ The subtopic $\theta_{t,s}$ for tweet s is endowed with a free Dirichlet parameter $\gamma_{t,s}$ and each topic indicator $z_{t,s,n}$ is endowed with a free multinomial parameter $\phi_{t,s,n}$. Because the Dirichlet is not amenable to sequential modeling, the natural parameters of each subtopic $\beta_{k,t}$ are chained with Gaussian distribution in a state space model, defined as

$$\beta_{k,t} \mid \beta_{k,t-1} \sim \mathrm{N}(\beta_{k,t-1}, \sigma^2 I) \text{ and } w_{t,n} \mid \beta_t \sim Mult(\pi(\beta_t)), \pi(\beta_t)_w = \frac{\exp(\beta_{t,w})}{\sum_w \exp(\beta_{t,w})} \quad \text{Eq.1}$$

and the sequential structure between models is captured with

$$\alpha_t \mid \alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I) \quad \text{Eq.2}$$

Due to the non-conjugacy of the Gaussian and Multinomial distributions, posterior inference is intractable. Thus instead of Gibbs sampling, [8] approximates the posterior inference with variational methods and introduces the variational Kalman filtering to model β_t evolving over time. The variational state space model can be formed as

$$\hat{\beta}_t \mid \beta_t \sim \mathrm{N}(\beta_t, \hat{v}_t^2 I) \quad \text{Eq.3}$$

where $\hat{v}_t$ is the variational parameters. The forward-backward algorithm is employed to compute the expectations for updating the variational parameters. For a certain term w, the variational forward distribution $p(\beta_t \mid \hat{\beta}_{i,i\le t})$ is a Gaussian distribution [21], and the mean m_t and the variance V_t are given by

$$\begin{aligned} m_t &= (\frac{\hat{v}_t^2}{V_{t-1} + \sigma^2 + \hat{v}_t^2}) m_{t-1} + (\frac{V_{t-1} + \sigma^2}{V_{t-1} + \sigma^2 + \hat{v}_t^2}) \hat{\beta}_t \\ V_t &= (\frac{\hat{v}_t^2}{V_{t-1} + \sigma^2 + \hat{v}_t^2})(V_{t-1} + \sigma^2) \end{aligned} \quad \text{Eq.4}$$

with initial conditions specified by the fixed m_0 and v_0. Similarly the mean $\tilde{m}_{t-1}$ and the variance $\tilde{V}_{t-1}$ of the variational backward distribution $p(\beta_t \mid \hat{\beta}_{i,i\le T})$ are given by

$$\begin{aligned} \tilde{m}_{t-1} &= (\frac{\sigma^2}{V_{t-1} + \sigma^2}) m_{t-1} + (\frac{V_{t-1}}{V_{t-1} + \sigma^2}) \tilde{m}_t \\ \tilde{V}_{t-1} &= V_{t-1} + (\frac{V_{t-1}}{V_{t-1} + \sigma^2})^2 (\tilde{V}_t - V_{t-1} - \sigma^2) \end{aligned} \quad \text{Eq.5}$$

with initial conditions $\tilde{m}_T = m_T$ and $\tilde{V}_T = V_T$. Finally from DTM, we can obtain two distributions, the topic distribution of the tweets and the word distribution of topics.

Algorithm 1. OPAD Algorithm

```
Input: tweets stream S, interval window \Delta t
Output: Peak Areas P = \phi
Initial: Mean and Variance (E,V) = Fresh( t_0 )
WHILE (t_i = t_{i-1} + \Delta t) < t_{n-1}
    IF \frac{Mean(t_i) - E}{V} > \tau AND Mean(t_i) > Mean(t_i - \Delta t)
        Peak area starts time: t_j^s = t_{i-1}
        WHILE (t_i = t_{i-1} + \Delta t) < t_{n-1} AND Mean(t_i) > Mean(t_i - \Delta t)
            (E, V)=Update (E, V, Mean(t_i) );   // perform hill-climbing
        END WHILE
    WHILE (t_i = t_{i-1} + \Delta t) < t_{n-1} AND Mean(t_i) > Mean(t_j^s)
        IF \frac{Mean(t_i) - E}{V} > \tau AND Mean(t_i) > Mean(t_i - \Delta t)
            Peak area stops: t_j^e = t_i - \Delta t
            BREAK
        ELSE
            (E, V) = Update (E, V, Mean(t_i) ); //perform down-hill
            update peak area stop time: t_j^e = t_i + \Delta t
        END IF
    END WHILE
    Output one peak area p_j^t =< t_j^s, t_j^e > and add it to peak area P
 ELSE
    (E, V) = Update (E, V, Mean(t_i) );
 END IF
END WHILE
Function Mean(t_i)
 tweets number in time interval t_i + \Delta t
Function Variance( t_i, ..., t_j )
 variance of tweet number in time interval ( t_i + \Delta t ,..., t_j + \Delta t )
Function Update(old_E, old_V, New_value)
 Diff=|old_E-New_value|;
 New_V= \pi * Diff + (1 - \pi) * old_V ;   (0 \le \pi \le 1);
 New_E= \pi * New_value + (1 - \pi) * old_E ;
```

IV. Sequential Summary Generation

Once the subtopics are identified by the stream-based or the semantic-based approaches, the tweets in each subtopic are ranked to generate the sub-summaries. The most significant tweets are extracted for each subtopic. Two ranking strategies are adopted to conform to the two different subtopic detection mechanisms.

A. Stream-based sequential summary generation

In stream-based subtopic detection, we obtain a serial of peak areas $P = \{ p_j^t \mid s_{t \in <t_j^s, t_j^e>} \in p_j^t, j = 1,..k \}$. The target of sequential summary generation is to generate a serial of sub-summaries $D = \{d_1, d_2, ..., d_k\}$ with each d_j summarizing the content of the peak area p_j . As expected, an effective sub-summary not only needs to be related to the topic (defined by Global Relevance), but also can generally summarize the content of all the tweets in the corresponding peak area (defined by Local Relevance). Besides, we define the Crowding Endorsement which suggests that the tweet gaining more endorsements from the crowds will be regarded as more important. In Twitter, a tweet can be re-tweeted by many others if they think it is important and we thus use re-tweeting to calculate the importance of crowding endorsement.

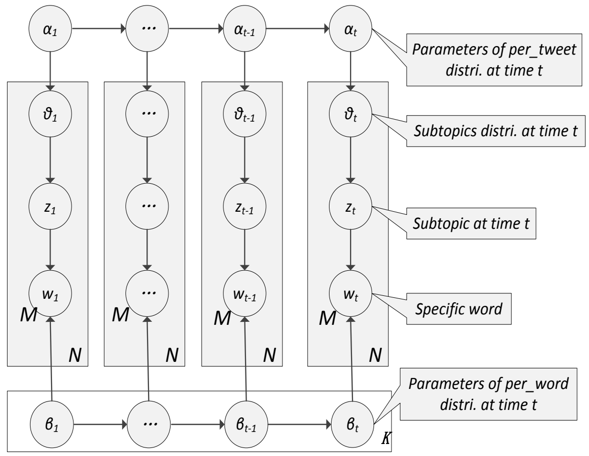

Fig. 2. Graphical representation of DTM

Global Relevance. The global relevance of the tweet s_i is defined as the cosine similarity between the tweet s_i and the entire stream *S*, i.e.

$$\mathrm{GRel}(s_i) = \mathrm{cosine}(s_i, S) = \frac{V_{s_i} \cdot V_S}{\|V_{s_i}\| \|V_S\|} \qquad \text{Eq.6}$$

Local Relevance. Assume that the tweets in a peak area represent a subtopic in the topic. The local relevance of the tweet s_i is defined as the cosine similarity between the tweet s_i and the tweets in the peak area that s_i belongs to, i.e. p_j .

$$\mathrm{LRel}(s_i) = \mathrm{cosine}(s_i, p_j) = \frac{V_{s_i} \cdot V_{p_j}}{\|V_{s_i}\| \|V_{p_j}\|} \qquad \text{Eq.7}$$

Crowding Endorsement. The endorsement of the tweet s_i from the crowds is measured by the normalized re-tweeting count.

$$\mathrm{Eds}(s_i) = \frac{\mathrm{RetweetCount}(s_i)}{\|\mathrm{TotalRetweetCount}\|} \qquad \text{Eq.8}$$

The weighed linear combination of the above three measures gives the final significance score of a tweet

$$Score(s_i) = \alpha_1 \cdot \mathrm{GRel}(s_i) + \alpha_2 \cdot \mathrm{LRel}(s_i) + \alpha_3 \cdot \mathrm{Eds}(s_i) \qquad \text{Eq.9}$$

where $\sum_i \alpha_i = 1$. The tweets in different peak areas are scored and ranked independently.

B. Semantic-based sequential summary generation

Before generating the sequential summary for the semantic-based approach, we first need to sort the K subtopics detected from DTM according to their temporal information. Particularly, we use the output probabilistic relationships between tweets and subtopics to assign each tweet to the subtopic that it most likely belongs to. Then the subtopics are ordered by the mean timestamp of the tweets in the corresponding subtopics. For the k^{th} subtopic, we evaluate the significance of the tweets directly by word distribution β_k, which can be also written as $p(w \mid z_k)$. The scores of tweets can be computed by

$$\text{Score}(s_i) = \sum_{w \in s_i} p(w \mid z_k), s_i \in z_k \qquad \text{Eq.10}$$

Then the tweets with the highest scores are selected for each subtopic. MMR [22] is used to remove redundancy in the generation of each sub-summary in both cases. For each tweet to be selected into sub-summary, it is compared against the tweets that are already selected in the previous sub-summaries. The tweet is selected only when it is considered not significantly overlapping any previously selected tweets.

V. DATASET AND EVALUATION MEASUREMENTS

A. Dataset

Using Twitter APIs (dev.twitter.com), we track Twitter official trending topics and download the real-time tweets about these topics. Totally 13,211,258 tweets of 2,113 topics are downloaded from March, 2$^{\text{nd}}$ to June, 4$^{\text{th}}$ in 2011. For research purpose, we choose 24 topics in which the topic development or the user focus change can be observed. These 24 topics cover most of the categories, like news ("*Minsk*", and "*Libya Release*"), technology ("*Apple iPhones*" and "*Google Fiber*"), sports ("*#bbcf1*" and "*World Champions*") and so on. The tweets for each topic are pre-processed by removing stop-words (Google English stop-words) and stemming, the details of which are provided in Table I.

Due to lack of benchmarks for sequential summarization evaluation, we invite two human summarizers to read all the tweets (chronologically ordered) for the 24 topics, and then ask them to write a serial of sub-summaries for each topic independently. Since the sub-summary can be regarded as a tweet aiming to summarize the tweets in its peak area, we require the summarizers to constrain the length of the sub-summary to 140 characters to comply with the limitation of tweet. But we do not limit the number of sub-summaries. For the two summarizers, the means of the sub-summary numbers are 4.84 and 4.4, and the variances are 6.33 and 4.06, respectively. The Kappa coefficient is leveraged to evaluate the agreement of word usages in their sequential summaries. The Kappa coefficient of 0.612 indicates a fair agreement between the two summarizers. These two sets of sequential summaries are then used as the gold standard for evaluation. Each system-generated sequential summary is evaluated against the two human-written sequential summaries and the average score is regarded as the evaluation result.

TABLE I

DETAILED INFORMATION OF DATASET

Category	#Topic	Topic Examples	Tweets Number
News	6	*Minsk, Libya Release*	25145
Sports	6	*#bbcf1, Lakers/Heat*	17204
Technology	5	*Google Fiber, Apple iPhones*	13281
Science	2	*AH1N1, Richter*	10935
Entertainment	2	*Midnight Club,*	6573
Meme	2	*#ilovemyfans, Night Angels*	14595
Lifestyle	1	*Goose Island*	6230
Total	24	----------------------	93963

B. Sequential summary evaluation measurements

The system-generated sequential summary is evaluated from three aspects, i.e., coverage, novelty and correlation. The coverage evaluates the overlap between system-generated and human-written sequential summaries. The novelty measures the average information increment in two successive sub-summaries. The correlation examines the sequential/order consistency of system-generated and human-written sub-summary serials. The definitions of these measures are described below.

1) Sequence coverage measurement

ROUGE [23] is a widely recognized metric in traditional summarization evaluation. The main idea of it is to count the matches between a candidate summary and a reference summary. The matches in two compared summaries are order-free. However, in our task the temporal information is an important factor. We expect the sub-summary-by-sub-summary matches and the matches in misaligned positions need to be penalized. Another reason that we do not adopt the ROUGE directly is that it requires the two summaries to be evaluated are equal in length. In our task though the length of each sub-summary is constrained in 140 characters, the number of sub-summaries for a topic can be different in human-written and system-generated sequential summaries. Thus it is not fair to directly employ ROUGE into our coverage evaluation. Out of these two considerations, we propose a position-aware coverage by accommodating the position information in matching. Let $D=\{d_1,d_2,\ldots,d_k\}$ denote a sequential summary and d_i denote the i^{th} sub-summary. The *N*-gram coverage is defined as:

$$\text{Coverage}(N-\text{gram}) = \frac{1}{|D^H|}\sum_{d_i \in D^H}\frac{1}{\omega_{ij}}\frac{\sum_{d_j \in D^S}\sum_{N-\text{gram}\in d_i^H, d_j^S} Count_{match}(N-\text{gram})}{\sum_{d_j \in D^S}\sum_{N-\text{gram}\in d_h^S}\text{Count}(N-\text{gram})} \text{ and } \omega_{ij} = \begin{cases} |j-i|+1, & j \neq i \\ 1, & j = i \end{cases} \quad \text{Eq.11}$$

where *i*, *j* denote the serial indexes of the sub-summaries in the human-written summary D^H and the system-generated summary D^S, respectively. $|D^H|$ is the number of sub-summaries in the sequential summary D^H. ω serves as a penalty coefficient to discount mismatches in sub-summaries. Each system-generated sub-summary is compared with all the human written sub-summaries and the match is weighted by $1/\omega$. We normalize the summing score to make sure the value is between 0 and 1. We evaluate unigram, bigram, and skipped bigram matches. Like in ROUGE [23], the skip distance is up to four words.

2) Sequence novelty measurement

To evaluate the novelty of sequential summaries, the information contained in a sub-summary is measured first. Here, we introduce the information content (*I*), which is used to measure the novelty of an update summary in [24]. In our evaluation, the novelty of a sequential summary is defined as the average of *I* increments of two adjacent sub-summaries:

$$\text{Novelty} = \frac{1}{|D|-1}\sum_{i>1}(I_{d_i} - I_{d_i,d_{i-1}}) \quad \text{Eq.12}$$

where $|D|$ is the number of sub-summaries. $I_{d_i} = \sum_{w\in d_i} I_w$ denotes the information containing in the sub-summary d_i and $I_{d_i,d_{i-1}} = \sum_{w\in d_i \cap d_{i-1}} I_w$ denotes the overlapped information of the two adjacent sub-summaries. I_w denotes the information content of a word *w* in sub-summaries and is computed as $I_w = \text{ITF}(w)\times\text{Relevance}(w, W_s)$ where $\text{ITF}(w)$ is the inverse tweet frequency of *w*, and W_s refers to all the tweets in the topic. The relevance function is introduced to ensure that the information brought by sub-summaries is not only novel but also related to the topic.

3) Sequence correlation measurement

For sequence correlation evaluation, we employ Kendall's tau coefficient Γ, which is often used to measure the association between two sequences [25]. The basic idea of Γ is to count the concordant and discordant pairs in two sequences. Borrowing this idea, for each sub-summary in a human-generated summary, we find its most matched sub-summary (judged by the cosine similarity measure) in the system-generated summary and then define the correlation according to the concordance between the two matched sub-summary sequences. The correlation Γ is defined as

$$\Gamma = \frac{2(|\#\text{ConcordantPairs}| - |\#\text{DiscordantPairs}|)}{n(n-1)} \quad \text{Eq.13}$$

where n is the number of human-written sub-summaries. In particular, we find out all the sub-summary two-tuples in the two sequential summaries. Any pair of two-tuples from each sequential summary is said to be concordant if the sub-summaries in the two-tuples are in the same order, otherwise, it is said to be discordant. Γ is in the range $-1 < \Gamma < 1$. If the agreement between two sequences is perfect, $\Gamma = 1$.

VI. EXPERIMENTS AND DISCUSSION

We compare the stream-based and semantic-based sequential summarization approaches with the following two baselines. ***Heuristic* (*stream-based*) *approach***: A simple approach to generate a sequential summary is to equally segment tweet streams into several time intervals and then select tweets from each time interval as sub-summaries. Considering the average subtopic number of the human-written summaries is 4.6, we equally segment each tweet stream into five intervals. To compare with our stream-based model with OPAD, we score tweets with the same measurements mentioned in Section IV.*A*. ***LDA-based* (*semantic-based*) *approach***: To compare with the semantic-based sequential summarization, we employ original LDA, which is a method widely used in topic modeling. LDA breaks the time order of the tweets and regards each tweet as an individual short document. Once the subtopics are identified by LDA, we use the same method to generate the sequential summary as seen in Section IV.B.

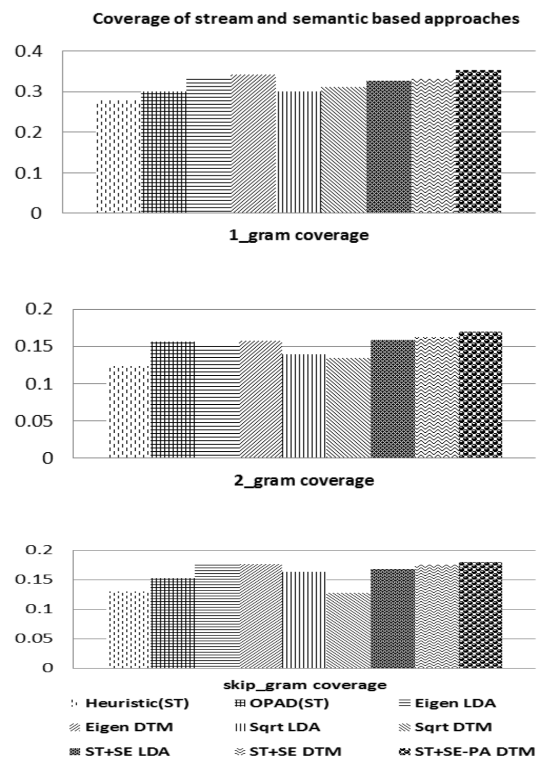

Fig. 3. Coverage evaluation of stream-based and semantic-based approaches

Both LDA-based and DTM-based topic modeling require a pre-determined subtopic number K. To avoid heuristically determining this topic number, we compare three approaches. **Eigen**: Li et al. [26] propose an approach to automatically identify the topic number for spectral clustering. Based on the tweets similarity using the 1-norm for its eigenvalues $\lambda_i^{norm}(i=1,2,..n)$, the ratio $\gamma_i = \lambda_i^{norm}/\lambda_2$ is calculated. The topic number is defined as $k=i+1$ if $\gamma_i - \gamma_{i+1} > \delta$. **Sqrt**: In Wan [9], the topic number is set to the square root of the total document number, that is $K=\sqrt{n}$. n denotes the tweet number in our study. **ST+SE**: We combine the stream-based and semantic-based subtopic modeling (ST+SE) by defining the topic number K as the number of peak areas detected by the OPAD algorithm.

The DTM-based topic modeling assumes that the tweet streams are divided by a time window. Most methods using DTM separate the streams with a fixed-size time window, like by year [8]. In our experiment, we first set the time window as one hour. But in actually, the subtopics hardly evenly distribute in the tweet stream. The evenly segmented time window may separate a subtopic or include partial content of two subtopics. Taking this into consideration, we attempt to use unfixed-size time window to separate the tweet streams. In the ST+SE approach, the OPAD algorithm detects a serial of peak areas $P=\{p_j^t \mid s_{t\in<t_j^s,t_j^e>} \in p_j^t, j=1,..k\}$. The time intervals ($<t_j^s, t_j^e> \ j=1,...,k$) already segment the tweet streams into discrete sets of tweets. Thus in the experiment of ST+SE-PA, we not only set the topic number with the peak area number, but also use the tweets in peak areas as the input to DTM.

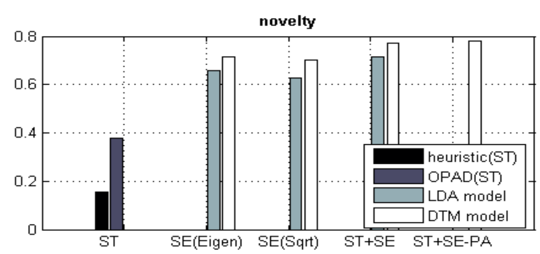

Fig. 4. Novelty evaluation of stream/ semantic-based approaches

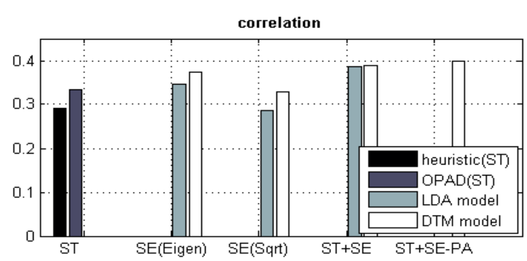

Fig. 5. Correlation evaluation of stream/semantic-based approaches

A. Evaluation of stream /semantic-based sequential summarization

In the heuristic-baseline and the OPAD-based approaches, two kinds of coefficients affect the performance of sequential summarization, i.e., the window size Δt in peak area detection and the weights $\alpha_1, \alpha_2, \alpha_3$, representing the importance of global relevance (GRel), local relevance (LRel) and endorsement (Eds), in sub-summary generation. In the steam-based approaches, Δt is set to be 3 hours and we equally treat the weights of the three measurements, that is $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$. In the Eigen (ssemantic-based) approach, δ is set to be 0.5.

Fig. 3 illustrates the coverage evaluation of the stream-based and semantic-based approaches. We find that the OPAD approach is superior to the heuristic-baseline approach in coverage, which suggests that the sequential summaries generated by the OPAD approach tends to include more information than those generated by the heuristic-baseline approach. We manually check some sub-summaries generated by the heuristic-baseline approach and find that the sub-summaries generated by the heuristic-baseline approach tends to repeat the main information and the subtopics cannot be described in time. On the contrary, the OPAD-based sequential summary can cover more diversity information. Besides, the performances of the semantic-based approaches (except the Sqrt approach) remarkably exceed those of the stream-based approaches. The reason is that in some situations LDA and DTM can detect the frequently-change subtopics by the semantic change and then select tweets for all these subtopics. For example, when two or more subtopics nearly happen in the same time, the stream-based approaches can only detect one volume change in the tweet stream. Though the peak area contains the information about these subtopics, it may not have enough space to select several tweets for each subtopic when generating sub-summaries. In this situation we find that the stream-based approach tends to select a tweet, which only contains the main information about the topic, to balance both the subtopics (or not to prefer any subtopic). Thus the subtopics detected by semantic-based approach are more fine-grained and the generated sub-summaries tend to contain more diversity information, which leads to the superior of the semantic-based approaches in sequential summarization as seen in Fig. 4 and Fig. 5.

In comparison with LDA, a remarkable superior can be seen in all three measures in DTM. The main difference is that DTM models subtopics not only by the word semantic change, but also accounts for the word temporal-related information as seen in Eq. (2). Consequentially, the detected subtopics in DTM and the generated sequential summaries contain more temporal information. Since the three measures account for sequential information, we observe the advantage of DTM in sequential summarization. In all these approaches, the ST+SE-PA approach shows the best results. This indicates the advantages of subtopic modeling in the ST+SE-PA approach. Meanwhile, in the ST+SE-PA approach we only use a small set of tweets (only tweets within the peak areas), which improves the efficiency of topic modeling.

B. *Influence of parameters on semantic-based sequential summarization*

In this section we evaluate the impact of coefficients δ (the number of subtopics selected in Eigen approach) and Δt (window size in ST+SE(-PA)) on semantic-based approaches. In Fig. 6 (a) and (b), when Δt is around 0.5~1 hour or δ is around 0.5~0.6, an overall best performance in coverage is achieved. When $\Delta t \sim (0.5 \sim 1)\text{hour}$, the average detected subtopic number in ST+SE is 5.8~4.2 and when $\delta \sim (0.5 \sim 0.6)$, the average subtopic number in Eigen is 4.1~5.6. These subtopic numbers are quite similar to the average sub-summaries number 4.62. When the topic number K is around this value, DTM may infer out a more reasonable subtopic distribution. Consequently with this setting, we can achieve the best result.

In the Eigen approach, when we set a small value to δ, it will generate very few subtopics on average 2~3 subtopics. Too few sub-summaries cannot contain enough information and lead to the lower coverage. With the increase of δ, we can generate more sub-summaries and, have more opportunities to cover more information, and thus achieve better coverage. After the subtopic number K exceeds the average subtopic number, the subtopics may contain more redundant information and increase the difficulty of subtopic matching and subtopic sorting, which influences the performances in both coverage and novelty. Similarly, when the subtopic number is small, it is easy for the system-generated summaries and the human-generated summaries to keep consistent in the time sequence, while when the subtopic number exceeds the average subtopic number, it will be difficult to sort the subtopics and reduce the correlation. In the ST+SE approach, when we increase Δt, the detected subtopic number declines, which is opposite to the increase of δ in Eigen approach. The trends of curves in the Eigen approach and the ST+SE approach are generally opposite to each other. These analyses suggests that the essential factor in sequential summarization is the determination of subtopic number. By comparing the ST+SE and ST+SE-PA approaches, we can also draw the conclusion that only using the tweets within peak areas in DTM achieves better results. This can be easily explained by the facts that 1) the tweets in the peak areas can reveal the subtopics in a more reasonable way and 2) this process can be regarded as a process of filtering noise tweets (filtering out the tweets outside peak areas). With the ST+SE-unfixed approach, DTM can generate a more accurate subtopic distributions and in turn improve the performances.

C. *Influence of parameters on stream-based sequential summarization*

In stream-based sequential summarization, two kinds of coefficients (window size Δt in peak area detection and the weights $\alpha_1, \alpha_2, \alpha_3$ in sub-summary generation) have impact on performance. In this section, we try different values for these coefficients. The evaluation results are illustrated in Fig. 7 When we only compare the three weights (that is to set $\alpha_{1,2,3}$ equal to 1), the endorsement measurements remarkably outperform the other two measurements in coverage, which suggests the crowds' endorsement in Twitter can greatly reflect the significance of a tweet. Looking at the curves of ($\alpha_{1,3} = 0.5$ and $\alpha_2 = 0$),

($\alpha_{2,3} = 0.5$ and $\alpha_1 = 0$) and ($\alpha_{1,2,3} = 1/3$), though their performances are competitive, we find that the global and local relevance cannot help the endorsement measurement to achieve a better performance. It is also shown in curves ($\alpha_{1,2} = 0.5$ and $\alpha_3 = 0$), ($\alpha_{2,3} = 0$ and $\alpha_1 = 1$), and ($\alpha_{1,3} = 0$ and $\alpha_2 = 1$) that the global and local relevance measurements do not work effective in coverage. This may tell us that the global and local relevance are different from the endorsement measurement and the linear combination may not be proper in sub-summary generation. Different from coverage, we find that the local relevance measurement ($\alpha_{1,3} = 0$ and $\alpha_2 = 1$) plays an important role in the novelty evaluation. This is because if sub-summaries care more about their local information of subtopics, they tend to include more novel information and increase novelty. Besides when we increase the window size Δt, the number of peak areas declines. There is little opportunity for these approaches to include new information and thus the performances of novelty decline. Similar to coverage, the endorsement is the most important factor in correlation where the curve of ($\alpha_{1,2} = 0$ and $\alpha_3 = 1$) is usually in the highest position. The reason for this can be explained by that the endorsement can represent the change of user focus to some extent. These changes are usually caused by the development of the topic. If we highlight this weight, the generated sequential summary tends to be consistent with the order of topic development and achieves a higher performance in correlation.

TABLE II

EVALUATION OF SUBTOPIC SEGMENTATION

Approach	Cosine Similarity
Heuristic (ST)	0.841
OPAD (ST)	0.693
ST+SE LDA	0.688
ST+SE DTM	0.681
ST+SE-PA DTM	0.677

TABLE III

HUMAN EVALUATION OF SEQUENTIAL SUMMARY

Sequential Summary	Heuristic-baseline	ST+SE-PA	Human-generated
Readability	2.16	2.34	4.97
Sequence	2.11	2.51	4.2
Novelty	2.16	3.03	4.95

D. *Evaluation of subtopic segmentation*

The researchers in topic segmentation usually use the metrics of Precision and Recall to examine the performances of their systems. However in our task, it is difficult for human editors to review the tweet streams and identify all the subtopics. To a certain extent, the subtopics cluster a tweet stream into segments and each segment tends to focus on one subtopic. Similar to [27], we compute the average cosine similarity between adjacent subtopics as an indicator of segmentation performance and the lower similarity means the better performance. We compare the stream/semantic- based approaches in Table II, where Δt equals to 3 and they share the same numbers of the subtopics for each topic, except that the topic number is still set to 5 for the heuristic (ST) approach. We can see that the stream/semantic approaches achieve comparable results and are superior to the heuristic (ST) approach. The ST+SE-PA DTM approach slightly outperforms the others. This improvement in topic segmentation also contributes to the sequential summarization.

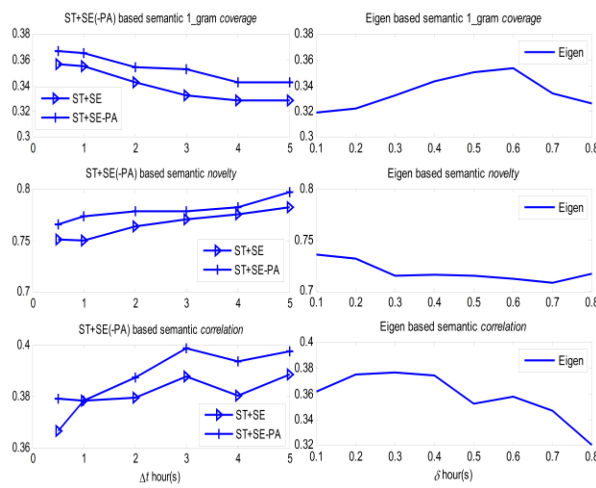

Fig. 6 (a) Evaluation of ST+SE(-PA), (b) Eigen approaches

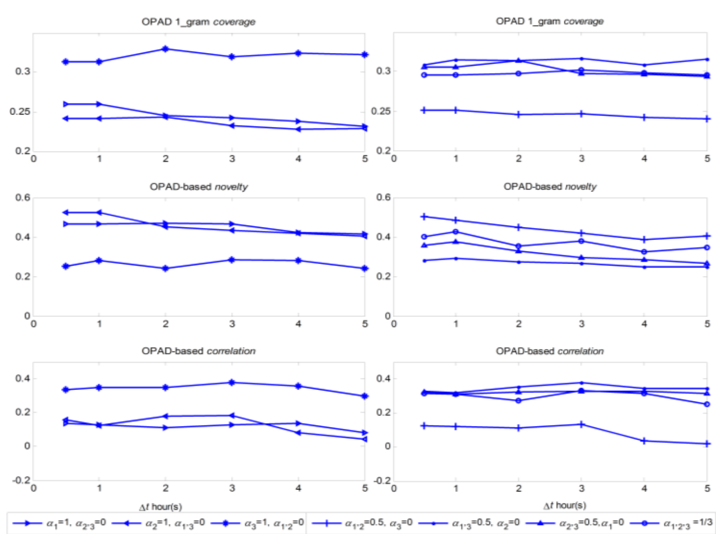

Fig. 7 Evaluation of stream-based approach

E. Human evaluation of sequential summarization

Two human evaluators are invited to rate the heuristic baseline, our approach (ST+SE-PA when $\Delta t = 3\text{hours}$) and human-generated sequential summaries in terms of **readability**, **sequence** and **novelty** using a 5-point scale (5 is the highest one). The readability criterion examines the writing quality of the sequential summary to see if it can be understand easily. The sequence criterion measures the consistence of the sub-summary order with the development of the topic. The novelty criterion evaluates how much new information contained in the successive sub-summaries. For the readability, the human evaluators rate each sub-summary in a sequential summary and the average score is regarded as the overall readability score. For sequence and novelty, the human evaluators rate each sequential summary. As seen in Table III, the comparable readability of heuristic-baseline and ST+SE-PA can result from that both approaches select significant tweets as sub-summaries. The readability is highly rely on the tweet writing style. Thus the readability evaluations of heuristic-baseline and ST+SE-PA are similar. On the contrary, human summarizers write sequential summaries by reorganizing and rewriting some of tweets after looking through the whole tweet streams. Thus it is no surprise that the readability of Human-generated sequential summaries has the highest score. The superior of ST+SE-PA over the heuristic-baseline approach in sequence and novelty evaluations reconfirms the effectiveness of the ST+SE-PA approach. We also find that the quality of human-generated sequential summaries is the highest in all the three criterions, which also suggests that there is still room for further improvement.

VII. CONCLUSION

In this paper we introduce a new application, namely sequential summarization for Twitter trending topics. The two proposed approaches identify the subtopics and extract significant tweets to generate sub-summaries. The evaluations in terms of the three measurements, including coverage, novelty and correlation as well as the human evaluation all demonstrate that the stream/semantic combination ST+SE-PA approach is the best option among all the proposed approaches. During the experiments, we draw another two conclusions as follows.

1) Subtopic Number: The subtopic number plays a crucial role in semantic-based approaches. If we can properly determine the subtopic number and then infer the subtopics with this number in LDA or DTM, the performances of the proposed approaches will be improved with no doubt.

2) Crowds Endorsement: Endorsement measurement is effective in selecting tweets in stream-based sub-summary generation. The approaches which adopt endorsement measurement perform well in the sequential-aware evaluations. This means that the crowds endorsement (retweeting in this paper) is an important indicator for the importance of tweets. If a tweet is retweeted by most of the crowd, it usually means that a new subtopic attracts the attention of Twitter users.

The above conclusions drive us to further study the determination of subtopic number and the better ways to model tweet streams, like a more proper window size or a new model to handle the sequential tweets. Besides, we believe that the sequential summary is a new way to reveal the vistas of topics. In the future, we plan to 1) normalize the selected tweets or employ other form of texts, e.g., from the relevant newswires to improve the readability of sequential summaries; 2) develop a system to visualize sequential summaries with a friendly interface.

References

[1] T. SAKAKI, M. OKAZAKI and Y. MATSUO, "Earthquake shakes Twitter users: real-time event detection by social sensors". *In Proc. WWW-10*, pp.851-860. 2010.

[2] B. O'CONNOR, M. KRIEGER and D. AHN, "TweetMotif: Exploratory Search and Topic Summarization for Twitter". *In Proc. ICWSM-10*. 2010.

[3] B. SHARIFI, M.-A. HUTTON and J. KALITA, "Automatic Summarization of Twitter Topics". *In Proc. of the National Workshop on Design and Analysis of Algorithms*. 2010.

[4] B. SHARIFI, M.-A. HUTTON and J.K. KALITA, "Experiments in Microblog Summarization". *In Proc. SOCIALCOM-10*, pp.49-56. 2010.

[5] F. LIU, Y. LIU and F. WENG, "Why is "SXSW" trending? Exploring Multiple Text Sources for Twitter Topic Summarization". *In Proc. ACL-11 Workshop on Language in Social Media*, pp.66-75. 2011.

[6] D. CHAKRABARTI and K. PUNERA, "Event Summarization Using Tweets". *In Proc. AAAI-11*. 2011.

[7] A. MARCUS, M.S. BERNSTEIN, O. BADAR, D.R. KARGER, S. MADDEN, and R.C. MILLER, "Twitinfo: aggregating and visualizing microblogs for event exploration". *In Proc. CHI-11*, pp. 227-236. 2011.

[8] D.M. BLEI and J.D. LAFFERTY, "*Dynamic topic models*". *In Proc. ICML*, pp.113-120. 2006

[9] L. ZHOU and E. HOVY, 2006. "On the Summarization of Dynamically Introduced Information: Online Discussions and Blogs". *In Proc. AAAI-06 Symposium on Computational Approaches to Analyzing Weblogs*, pp.237-242. 2006.

[10] M. HU, A. SUN and E.-P. LIM, "Comments-oriented blog summarization by sentence extraction". *In Proc. CIKM07*. 2007.

[11] S. HARABAGIU and A. HICKL, "Relevance Modeling for Microblog Summarization". *In Proc. AAAI-11 Symposium on Weblogs and Social Media*. 2011.

[12] A.-M. POPESCU and M, PENNACCHIOTTI, "Detecting controversial events from twitter". *In Proc. CIKM-10*, pp. 1873-1876. 2010.

[13] J. NICHOLS, J. MAHMUD and C. DREWS, "Summarizing sporting events using twitter". *In Proc. IUI-12*. 2012.

[14] X. LI and W.B. CROFT, "Improving novelty detection for general topics using sentence level information patterns". *In Proc. CIKM-06*, pp. 238-247. 2006.

[15] CHEN, L. and ROY, "Event detection from Flickr data through wavelet-based spatial analysis". *In Proc. CIKM-09*. 2009.

[16] S. PETROVIC, M. OSBORNE and V. LAVRENKO, "Streaming first story detection with application to Twitter". *In Proc. ACL-10*. 2010.

[17] J. SANKARANARAYANAN, H. SAMET, B.E. TEITLER, M.D. LIEBERMAN and J. SPERLING, "TwitterStand: news in tweets". *In Proc. SIGSPATIAL-09*, pp.42-51. 2009.

[18] H. BECKER, M. NAAMAN and L. GRAVANO, "Selecting Quality Twitter Content for Events". *In Proc. AAAI-11 Symposium on Weblogs and Social Media*. 2011.

[19] D.A. SHAMMA, L. KENNEDY and E.F. CHURCHILL, "Tweet the debates: understanding community annotation of uncollected sources". *In Proc. SIGMM-09 workshop on Social media*, pp.3-10. 2009.

[20] D.M. BLEI, Ng Y. Andrew and M. I. Jordan, "*Latent Dirichlet allocation*". *Journal of Machine Learning Research*, pp.993-1022. 2003

[21] R. KALMAN, “A new approach to linear filtering and prediction problems”. *Journal of Basic Engineering*, pp.35-45. 1960.

[22] J. CARBONELL and J. GOLDSTEIN, “The use of MMR, diversity based reranking for reordering documents and producing summaries”. *In Proc. SIGIR-98*, pp.335-336. 1998.

[23] C.Y. LIN, “ROUGE: a Package for Automatic Evaluation of Summaries”. *In Proc. ACL-04 Workshop on Text Summarization Branches Out*, pp. 74-81. 2004.

[24] G. AGGARWAL, R. SUMBALY and S. SINHA, “Update Summarization”. *Stanford: CS224N Final Projects*. 2009.

[25] M. LAPATA, “Automatic Evaluation of Information Ordering: Kendall's Tau”. *Computational Linguistics*, pp. 471-484. 2006.

[26] W. LI, W.-K. NG, Y. LIU and K.-L. ONG, “Enhancing the Effectiveness of Clustering with Spectra Analysis”. *IEEE Transactions on Knowledge and Data Engineering*, pp.887-902. 2007.

[27] YAJUAN DUAN, ZHUMING CHEN, FURU WEI, MING ZHOU and HEUNG-YEUNG SHUM, “Twitter Topic Summarization by Ranking Tweets Using Social Influence and Content Quality”. *In Proc. Coling*-12, pp. 763-780. 2012.

Dehong Gao is a PhD candidate in the Department of Computing, the Hong Kong Polytechnic University, Hong Kong. His current research interests include opinion mining and social network study.

Wenjie Li is currently an associate professor in department of computing, the Hong Kong Polytechnic University, Hong Kong. She received her PhD degree from department of systems engineering and engineering management in the Chinese University of Hong Kong, Hong Kong, in 1997. Her main research topics include natural language processing, information extraction and document summarization.

Xiaoyan Cai is currently an assistant professor in College of Information Engineering, Northwest A&F University. She was a research associate in department of computing, the Hong Kong Polytechnic University, Hong Kong, from June 2009 to June 2011. Her current research interests include document summarization, information retrieval and machine learning.

Renxian Zhang is a PhD candidate in the Department of Computing, the Hong Kong Polytechnic University, Hong Kong. His current research interests include natural language processing, text mining, and document summarization.

You Ouyang is currently an algorithm engineer in Beijing Miaozhen Systems, China. He received a PhD degree of computer science from the Hong Kong Polytechnic University in 2011. His main research interests include statistical natural language processing and text mining.