Event Detection and Visualization for Social Text Streams

Qiankun Zhao Prasenjit Mitra
College of Information Sciences and Technology
Pennsylvania State University, University Park, PA, 16802
{qzhao, pmitra}@ist.psu.edu

Abstract

In this paper, we propose to detect events from social text streams by exploring the content as well as the temporal, and social dimensions. We define the term *event* in the *social text streams*(e.g., blogs, emails, and Usenets) as a set of relations between *social actors* on a specific *topic* over a certain *time period*. We represent social text streams as multi-graphs, where each node represents a social actor and each edge represents a piece of text communication that connects two actors. The content and temporal associations within each text piece are embedded in the corresponding edge. Then, events are detected by combining text-based clustering, temporal segmentation, and graph cuts of social networks. Moreover, we provide a multi-dimensional visualization tool that visualizes the relations between different events along the three different dimensions. Experiments conducted with the Enron email dataset[1] show the advantages of exploring the social and temporal dimensions along with content, and the usefulness of the visualization tool.

Keywords

Social Network, Event Detection, Text Mining.

1. Introduction

Recently, *social text stream data*, such as weblogs, message boards, mailing lists, Usenet, have become ubiquitous with the evolution of the Web. Social text stream data is defined as a collection of informal text communication data that arrives over time and each piece of text in the stream is associated with some social attributes such as author, reviewer, sender, and recipients. Usually, social text stream data arrives over time and each piece of the stream carries part of the semantics (e.g., information about real world events) [2]. In this sense, social text streams are sensors of the real world.

With massive amount of data and various types of sources in social text streams, together with the rich content such as text, social actors and relations, and temporal information, efficiently organizing and summarizing the embedded semantics has become an important issue [4, 1]. Most text semantic analysis techniques mainly focused on *formal text stream data*. However, the *social text stream data* is substantially different from formal text stream data: (1) social text stream data is more context sensitive (e.g., some emails may be meaningless without contexts), and (2) social text stream data contains rich social connections between the information senders/authors and recipients/reviewers.

Given the distinguishing features of social text stream data and the rich information embedded in them, we propose to detect events from them by exploring features in the three dimensions: textual content, social, and temporal. Different from existing event detection approach, our event detection results provide more comprehensive summarizations such as (1) *people that are involved*; (2) *content of the event*; and (3) the *time period* this event happened. Specifically, within the three components (social, temporal, content), the following are important: (a) which are the key emails that initiated the discussion? (b) how were these people organized into groups; and (c) how this event evolved from one phase to another phase. In addition, the relations between different events in the three dimensions (temporal, social, and content) can be explored to provide a better summarization and navigation experience over the social text stream data.

Such knowledge can be useful in many application ranging from terrorism activity detection and monitoring, to business intelligence. Discovering such knowledge is challenging due to the following reasons. Existing topic detection approaches may integrate the social or temporal contexts. However, they consider the temporal and social information as two extra dimensions along with several email content-based features. Existing approaches may not be efficient to extract such comprehensive and context-based events from the social text stream data because: (1) fusing the temporal and social features along with content features may diminish the importance of the temporal information and the social network information, and (2) relations between different events along the three dimensions are difficult to extract from the fused feature sets. We observe that: (a) Users may communicate on more than one topic/event via emails during the same time period; (b) An event or a set of similar events may cover more than one time periods both consecutive periods and non-consecutive periods; (c) An event may be discussed by different groups of email addresses that may or may not have social overlaps at the same or different time periods; (d) Sometimes, a stream of emails may drift from one event to another unintentionally and smoothly even without changing the subject.

In this paper, we propose to extract, summarize, and visualize events from social text stream data by exploring the embedded social and temporal information with the email content. Note that *event* here is defined as a group of social actors that communicate with each other on a specific topic

[1] http://www.cs.cmu.edu/∼enron/

over a certain time period. The three dimensions are combined to improve the quality of event detection, which cannot be detected from any single or two combinations of them, by exploring the correlation within and cross dimensions. Moreover, relations between different events are explored for a better summarization and visualization.

Experimental results with Enron dataset show that: (1) exploring the social and temporal dimensions with content can improve the event detection quality, and (2) the 3 dimensional visualization makes the navigation more efficient and brings more insights about these events. Although we use email data, our work is easily extendible to weblog data and data from web forums and other social text streams.

The following are our major contributions: (i) We introduce the concept of *social text stream data* and propose the first approach to explore the social network, text content, and temporal information embedded in the social text stream together; (ii) A new definition of event is provided with respect to the social, temporal and content dimensions. We describe a graph-based algorithm for event detection in social text stream data; (iii) We propose a comprehensive visualization method to fully display the relations among the set of events extracted. Our method provides viewpoints from different angles (temporal, social, and content) and can be used for further reasoning and modeling.

2. Problem statement

Social text stream data is characterized by the following properties: (1) pieces of text data stream over time; and (2) social actors (identified by their electronic or virtual IDs) are involved in the text stream and information flows from one actor to another. Different from any other text corpus or stream data, which only have text or/and temporal dimensions, social text stream data have three dimensions: text content, temporal information, and social relations.

A collection of social text stream data can be represented as $D = <(p_1,\ t_1,\ s_1),\ (p_2,\ t_2,\ s_2),\ \cdots,\ (p_n,\ t_n,\ s_n)>$, where $p_i \in P = \{p_1, p_2, \cdots, p_{|P|}\}$ is a piece of text content that flows between the set of *social actors* s_i at time point t_i; s_i is a pair of social actors $< a_i, r_i >$ where a_i is the actor who initiates the information flow and r_i is the actor who receives/comments on p_i. In this paper, p_i represents an email; a_i and r_i represent the sender and recipient of the email; t_i is the timestamp when this email is sent.

In traditional text stream segmentation or event detection approaches, an event is defined as a set of content pieces that are similar to each other but different from content pieces in other event sets [5]. We extend the existing definition of an event by introducing the social and temporal contexts of text streams.

Definition 1. (**Event**) Given a social text stream corpus denoted as $D = <(p_1,\ t_1,\ s_1),\ (p_2,\ t_2,\ s_2),\ \cdots,\ (p_n,\ t_n,\ s_n)>$, an *event* is defined as a subset of triples $\mathbb{M} = \{(p_1,\ t_1,\ s_1),\ (p_2,\ t_2,\ s_2),\ \cdots,\ (p_l,\ t_l,\ s_l)\ \}$ such that: (1) $\forall p_i, p_j \in P_{\mathbb{M}} = \{p_1, p_2, \cdots, p_{|\mathbb{M}|}\}$ are semantically coherent to a topic $\mathcal{T}_{\mathbb{M}}$; (2) $\forall$ consecutive time stamps $t_i, t_j \in T_{\mathbb{M}} = \{t_1, t_2, \cdots, t_{|\mathbb{M}|}\}$, $\delta(t_i, t_j) < \omega_{\mathbb{M}}$, where $\omega_{\mathbb{M}}$ is a time interval; and (3) all actors $a_i, r_i \in S_{\mathbb{M}} = \{s_1, s_2, \cdots, s_{|\mathbb{M}|}\}$ form a fully connected graph via $P_{\mathbb{M}}$.

That is, an *event* is represented as a set of text pieces (here emails) with semantical, social, and temporal constraints. Note that these events can be extracted at different granularities. That is from the temporal, content, and social dimensions, events can be detected at different levels of details.

Definition 2. (**Basic event**) Given an event $\mathbb{M}$, denoted as $\mathbb{M} = \{(p_1,\ t_1,\ s_1),\ (p_2,\ t_2,\ s_2),\ \cdots,\ (p_l,\ t_l,\ s_l)\ \}$, $\mathbb{M}$ is a *basic event* if and only if there exist no event $\mathbb{M}' \neq \mathbb{M}$ such that $l(\mathcal{T}_{\mathbb{M}}) \geq l(\mathcal{T}_{\mathbb{M}'})$, $\omega_{\mathbb{M}} \geq \omega_{\mathbb{M}'}$, and $S_{\mathbb{M}} \supseteq S_{\mathbb{M}'}$. Here $l(\mathcal{T}_{\mathbb{M}})$ and $l(\mathcal{T}_{\mathbb{M}'})$ are the level of the topic in the hierarchical topic structure $H_{\mathcal{T}}$.

Note that the hierarchy of topics $H_{\mathcal{T}}$ and meaningful time intervals are extracted automatically from the social text stream data. From the definition, it can be observed that *basic events* are basic/smallest elements of events. That is, events can be categorized into two classes: *basic events* and *composite events*. A *composite event* consists of a set of *basic events* that share some common properties in any of the semantic, social, or temporal dimensions.

In general, a basic event $\mathbb{M}$ is denoted as $\{\mathcal{T}_{\mathbb{M}},\ \omega_{\mathbb{M}},\ S_{\mathbb{M}}\}$. Given a collection of basic events, there are different ways to combine them to form events at different granularities. For example, two basic events $\mathbb{M}$ and $\mathbb{M}'$ can be merged into a composite event, given that they share any of the three elements in the event (e.g., $\mathcal{T}_{\mathbb{M}}$ and $\mathcal{T}_{\mathbb{M}'}$ are siblings in $H_{\mathcal{T}}$, or $T_{\mathbb{M}} = T_{\mathbb{M}'}$, or $S_{\mathbb{M}} = S_{\mathbb{M}'}$). Based on this observation, rather than using flat representation, we propose to build an *event hierarchy*. Formally, a *event hierarchy* is defined as follows.

Definition 3. (**Event hierarchy**) Given a set of basic events $\{\mathbb{M}_1,\ \mathbb{M}_2,\ \cdots,\ \mathbb{M}_n\}$, the *event hierarchy* $H_{\mathbb{M}}$ is a hierarchical structure such that: (1) each leaf node is a basic event; (2) each internal node is a composite event and consists of its children events; (3) for any $\mathbb{M}_i$, $\forall\ \mathbb{M}_j \in Anc(\mathbb{M}_i)$, $l(\mathcal{T}_{\mathbb{M}_j}) \geq l(\mathcal{T}_{\mathbb{M}_i})$, $\omega_{\mathbb{M}_j} \geq \omega_{\mathbb{M}_i}$, and $S_{\mathbb{M}_j} \supseteq S_{\mathbb{M}_i}$, where $Anc(\mathbb{M}_i)$ is the list of ancestor events of $\mathbb{M}_i$ in the event hierarchy.

The event hierarchy has the following advantages: (1) it contains all valid events at different granularities besides the basic events; (2) it reflects the relations between different events in different dimensions; and (3) using the event hierarchy, a *multi-context summarization* can be obtained for each event. Here *multi-context summarization* refers to summarization of event with respect to related events in the three dimensions.

In general, the problem of **event detection and visualization** for **social text stream data** (**STSD**) can be defined as follows: (1) automatically extract the set of basic events, $\{\mathbb{M}_1, \mathbb{M}_2, \cdots, \mathbb{M}_n\}$, for a given STSD; (2) generate multi-granularity events by exploring the three dimensions of the basic events; (3) construct the event hierarchy $H_{\mathbb{M}}$ and provide events visualization from different dimensions.

3. Event detection & visualization

The work flow of event detection from STSD is as follows. Given the STSD collection, first, the contents of text pieces in the data stream are analyzed to extract a hierarchy of topics. Within each topic, there may be many semantically similar events that cannot be detected. Rather than combining the content and temporal information together to extract these events, text pieces within each topic is further partitioned into smaller clusters based on the traffic flow of emails over time. The number of events is automatically determined by the traffic pattern under this topic. Then, a hierarchy among these events is constructed with respect to the text content and temporal dimensions. After that, the social network is explored by represent the STSD as a multi-graph, where each node is a social actor, each edge represents an email communication between two actors. As a result, the semantic communities among these email addresses can be extracted using graph cut algorithm. Here each email address may belong to different communities at the same time, and it may change from one community to another over time as well. To capture such dynamic patterns, for each event extracted using the content and temporal information, a set of communities is extracted. Moreover, the set of basic events is extracted by combining basic elements in the three dimensions. By exploring the relations in these basic events, a hierarchy of events and visualization can be constructed.

3.1 Hierarchical topic clustering

Taken each piece of text in the social text stream data as plain text, it can be represented as a sequence of words. Each text piece is denoted as a vector of words $\vec{p_i} =< w_1, w_2, \cdots, w_n >$, where w_i is the weight of the ith word $word_i$ in the vector. Here the weight of each word in the text piece is quantified as the $TF \cdot IDF$. Given two text pieces, p_i and p_j, the content-based similarity is defined as the cosine similarity:

$$Sim_{content}(p_i, p_j) = \frac{\vec{p_i} \cdot \vec{p_i}}{|\vec{p_i}||\vec{p_j}|}$$

Then, a graph representation is used for the text corpus, where each node is an email and the edge is content similarity between two emails. To extract the topic hierarchy, the graph cut algorithm is to minimize the following function:

$$\sum_{r=1}^{k} \frac{cut(P_r, P - P_r)}{\sum_{p_i, p_j \in P_r} Sim_{content}(p_i, p_j)}$$

where $cut(P_r, P - P_r)$ is defined as the sum of similarities of these edges that are removed to partition P into P_r and $P - P_r$.

$$cut(P_r, P - P_r) = \sum_{u \in P_r, t \in P - P_r} w(u, t)$$

As a result, a hierarchy of topics is extracted purely based on the content of the social text stream.

3.2 Temporal-improved event detection

However, within a topic, there are different events that cannot be effectively distinguished by using only the content. The intuition to utilize the temporal information is that email communications about the same event are expected to fall into a specific time interval in the history. That is, the traffic pattern (highs, lows, and peaks) of emails within a topic can be used to extract events in a finer granularity.

Given the sequence of emails within a topic, $< p_1, p_2, \cdots, p_{n+1} >$, the gap between any two consecutive emails can be constructed as $< g_1, g_2, \cdots, g_n >$. Basically, if the sequence of messages are generated by the two hidden events $\mathbb{M}_1$ and $\mathbb{M}_2$, the gap between two consecutive emails should follow this function: $f_0(g_i) = \alpha_0 \cdot e^{-\alpha_0 \cdot g_i}$ when it is in state $\mathbb{M}_1$; $f_1(g_i) = \alpha_1 \cdot e^{-\alpha_1 \cdot g_i}$ when it is in state $\mathbb{M}_2$. $\mathbb{M}_1$ has the probability to stay in $\mathbb{M}_1$ and change to $\mathbb{M}_2$. To automatically extract different events, extending the two-state model, we adopt the infinite-state model [3]. Given a sequence of $n+1$ messages and the gap sequence $g =< g_1, g_2, \cdots, g_n >$, to find a sequence of state q=$< q_1, q_2, \cdots, q_n >$, to minimize the cost function:

$$C(q|g) = (\sum_{t=0}^{n-1} \tau(i_t, i_{t+1})) + (\sum_{t=1}^{n} - \ln f_{i_t}(g_t))$$

As a result, a hierarchy of temporal segments can be extracted using the above infinite-state model. That is for each topic in $H_{\mathcal{T}}$, a list of temporal improved events $\{\mathbb{M}_1, \mathbb{M}_2, \cdots, \mathbb{M}_l\}$ can be extracted.

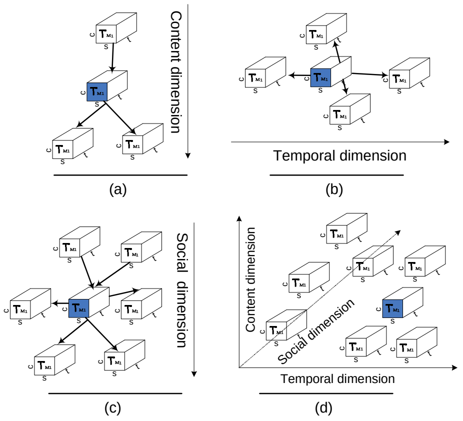

Fig. 1: *Construction of event hierarchy*

3.3 Social-improved event detection

Even with the improvement of temporal information, it is possible that different groups of social actors (communities) talk about the same event at the same time. However, the purpose and point of view from different communities can be different and these communities may or may not be somehow socially connected. For instance, even the same keyword or topic may mean different real thing in two different communities. To further differentiate such events, we propose to utilize the social network information.

Given the event hierarchy extracted in the previous phase, the social text stream data is represented as a multi-graph $G = (N, E)$, where N is a set of nodes and any $n_i \in N$ corresponds to an email address; E is a set of edges and any $e_i \in E$ corresponds to an email or a group of emails between the two connected address. That is, emails that belong to the same temporal-improved event between a pair of addresses can be represented as a single edge here. For each edge, there are two attributes: *event* and *timestamp*. Note that there can be multiple edges between the same pair of nodes.

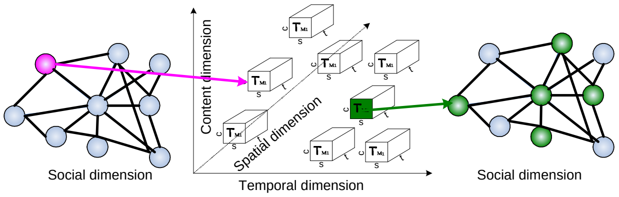

Fig. 2: *Social-based visualization*

In general, a set of event independent communities can be extracted from the multi-graph. That is, each node is represented as a vector of events and weight of each event can be the frequency of events and inverted actor frequency using the IR concepts. Then, the similarity between any two nodes can be obtained as the cosine similarity of the two vectors. Moreover, given a temporal-improved event, the event dependent communities can be extracted as well. That is, for a specific temporal based event, there will be a set of different social communities. The temporal based event with respect to each social community will correspond to the concept of basic event we introduced.

3.4 Event visualization

Topic	Term1	Term2	Term3	Term4	Term5
E1	agreement	attach	doc	doc	execute
E2	california	pacific	logo	assembly	reference
E3	crawler	hourahead	intervent	manual	failure
E4	fantasy	injury	nfl	sports	league
E5	company	billion	govern	nation	million

Table 1: *Examples of basic event*

Once the basic events are detected, there are two issues for the event summarization: *event hierarchy construction*, which explores the relations between events, and *multi-context event visualization* that visualize events from different prospectives.

Given the set of basic events, which can be represented as a set of triples $\{\mathcal{T}_{\mathbb{M}}, \omega_{\mathbb{M}}, S_{\mathbb{M}}\}$, the event hierarchy can be constructed by merging these basic event from any one or two of the three dimensions. An example of the relations between events from different prospectives is shown in Figure 1.
Once the event hierarchy is constructed, given an event, a multi-context visualization can be obtained. That is, by exploring the parent and children events in the hierarchy, context-based visualization can be presented. For example, as shown in Figure 2, given a social actor, we can visualize all the events that he/she was involved in a 3-D environment.

Topic	Term1	Term2	Term3	Term4	Term5
C1	California	electro	energy	power	util
C2	ferc	commission	propos	file	rate
C3	agreement	execute	attach	document	copi

Table 2: *Examples of composite event*

4. Performance study

In this paper, the Enron Email dataset is used. The raw Enron corpus contains 619,446 messages belonging to 158 users with an average of 757 messages per user. Each message is a plain text file and these messages are organized based on the network references(email addresses). For each email address, there are different folders such as *sent* and *inbox*.

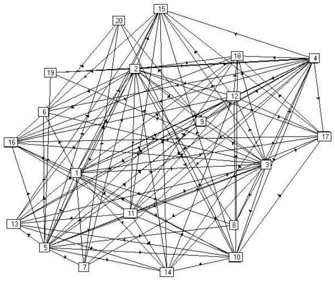

Fig. 3: *Content-temporal-social-based community*

Table 1 presents *5* basic events that are extracted **based only on the email content**. For each cluster, a list of discriminative keywords is selected based on the TF.IDF weights. Table 2 presents *5* composite events in the content-based hierarchy. It can be observed that basic events are in a finer granularity than these composite as the basic events contain a smaller number of emails that discussing only about one aspect of the topic represented by their parent nodes in the hierarchy. Existing social text stream mining approach usually use flat representation of events, whereas our hierarchical representation can explore events at different granularity and capture the relations between them.

Time Interval	Key terms				
Year 2000, Week 34–38	dalla	mail	california	sierra	resourc
	roll	price	clinton	assembl	bill
Year 2000, Week 42–46	exert	senat	utilit	summari	hear
	market	margaret	send	favor	add
	contract	barwatt	pool	fuel	change

Table 3: *Temporal-improved events*

Table 3 shows the **temporal-improved events extracted via segmentation of content-based cluster**. That is, each segment of the content-based event is supposed to represent a smaller event. It can be observed that: (1) temporal improved events can distinguish segments within topic-based events that belong to two different aspects of the topics/events that share similar keywords, (2) there are more than one event in a specific time window.

Figures 3 presents an example community extracted from the email stream based on the **content-temporal-social relation**. Note that the community in Figures 3 is a subcommunity of general content-based community during a certain time period. In the experimental results we observed that social communities extracted using only email content cannot accurately reflect how these people communicated.

5. Conclusions

In this paper, we proposed the concept of *social text stream data*, which is becoming more popular and useful. By exploring the temporal and social information, together with the text content, we showed that social text stream data contains much richer semantics and our proposed event detection and visualization approach can produce much better results than existing state-of-the-art event detection approaches.

Acknowledgement

This research was partially sponsored by the Laboratory Directed Research and Development Program of the Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC., for the U.S. DOE under Contract No. DE-AC05-00OR22725. This manuscript has been authored by sub-contractors for UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. Authors would like to thank Anyi Chen of Pennsylvania State University for the help of data preparation.

References

[1] S. D. Afantenos, *et al.* An introduction to the summarization of evolving events: Linear and non-linear evolution. In *LNCS*, pages 91–99, 2005.

[2] J. Kleinberg. *Data Stream Management: Processing High-Speed Data Streams*, chapter Temporal Dynamics of On-Line Information Streams. Springer, 2006.

[3] Jon Kleinberg. Bursty and hierarchical structure in streams. In *KDD*, pages 91–101, 2002.

[4] A. Krause, J. Leskovec, and C. Guestrin. Data association for topic intensity trackingn. In *ICML*, 2006.

[5] Mehran Sahami and Timothy D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *WWW*, pages 377–386, 2006.