

# Temporal and Information Flow Based Event Detection From Social Text Streams

Qiankun Zhao Prasenjit Mitra Bi Chen

College of Information Sciences and Technology, Pennsylvania State University, University Park  
qkzhao@ist.psu.edu pmitra@ist.psu.edu bchen@ist.psu.edu

## Abstract

Recently, *social text streams* (e.g., blogs, web forums, and emails) have become ubiquitous with the evolution of the web. In some sense, social text streams are sensors of the real world. Often, it is desirable to extract real world events from the social text streams. However, existing event detection research mainly focused only on the stream properties of social text streams but ignored the contextual, temporal, and social information embedded in the streams. 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* as the information flow between a group of *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 the information flow between two actors. The content and temporal associations within the flow of information are embedded in the corresponding edge. Events are detected by combining text-based clustering, temporal segmentation, and information flow-based graph cuts of the dual graph of the social networks. Experiments conducted with the Enron email dataset<sup>1</sup> and the political blog dataset from Dailykos<sup>2</sup> show the proposed event detection approach outperforms the other alternatives.

## Introduction

Recently, *social text stream data*, such as weblogs, message boards, mailing lists, review sites and web forums, have become ubiquitous with the evolution of the web. We refer to a collection of text communication data that arrives over time as *social text stream data*, where each piece of text in the stream is associated with some social attributes such as author, reviewer, sender, and recipients. In the last few years, social text stream data has changed the way we communicate daily, the way we market businesses, and even the way political campaigns are conducted. For example, weblogs are now used not only by individual users for sharing information about their daily lives and thoughts, but are also used by large business corporations and political parties to

release their latest products or new proposals. 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) (Kleinberg 2006). In this sense, social text streams are sensors of the real world.

Social text streams generate large amounts of text data from various types of sources. The streams have rich content such as text, social networks, and temporal information. Efficiently organizing and summarizing the embedded semantics has become an important issue. Most text semantic analysis techniques mainly focused on TDT (Topic Detection and Tracking) (Krause, Leskovec, & Guestrin 2006; Yang, Pierce, & Carbonell 1998) for *general text data* (e.g., scientific papers, newswires, and corporation documents). However, the *social text stream data* is substantially different from general text stream data: (1) social text stream data contains rich social connections (between the information senders/authors and recipients/reviewers) and temporal attributes of each text piece; and (2) the content of text piece in the social text stream data is more context sensitive. That is, not only within the text piece content, the meaning of words are dependent from the context; but also the meaning of a text piece is dependent from the social actors (e.g., sender, author, recipient, and commenter) and other temporally correlated text pieces (e.g., the temporal and content information of previous text pieces).

In this work, we propose to detect events from social text streams by exploring features in three dimensions: textual content, social, and temporal. In the literature, there are some existing works on semantic analysis of text stream data and social network data (McCallum & Huang 2004; Mei *et al.* 2006; Yang, Pierce, & Carbonell 1998). For example, different text content analysis techniques have been proposed to classify huge amount of emails into different topics or identify entities (McCallum & Huang 2004). Event detection algorithms have been proposed for newswires (Yang, Pierce, & Carbonell 1998) and blogs (Mei *et al.* 2006). Different communities can be extracted from the email communication network as well (Leuski 2004). However, most of them ignored either the social network information, or the temporal properties of the stream data.

However, the embedded temporal and information flow pattern (communication pattern), and the social network relations in social text stream data can be used to improve the

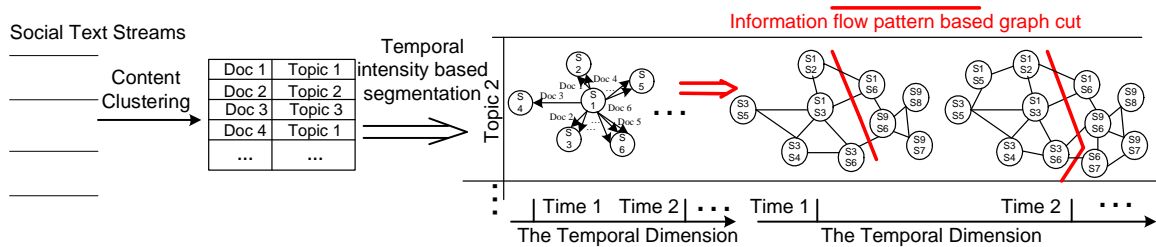


Figure 1: Overview of Temporal and Information Flow based Event Detection From Social Text Streams

quality of event detection by distinguishing events that cannot be separated purely based on the textual content. For example, similar events that share the same keywords (e.g., *the Emily hurricane event*, *the hurricane movie*, and *the Dennis hurricane event*, or various of *the New Year events* in different communities such as Chinese, Indian, and Western) cannot be distinguished until the temporal, information flow pattern, and social dimension are taken into account. The above lists of *hurricane events* can be detected by combining the information flow pattern and content of social text streams, whereas the *New Year events* can be distinguished by combining both the information flow pattern, social network community extraction and text content analysis.

By integrating the temporal, information flow pattern, and social network information, not only can we extract events more accurately, at a finer granularity, but we can also explore the relations between the detected events further along the content, the temporal, and the social dimensions. Such knowledge can be useful in various applications for distinguish ambiguous concepts and events. However, discovering such knowledge is challenging due to the fact that: (1) in the same time period, the same pair of social actors may communicate about more than one event, which cannot be efficiently distinguished using only the content as the contexts are missing; (2) at different time periods, either the same or different pairs of social actors may communicate with similar content but for either the same or different events. That is, in the above two cases, by taking the temporal information and social information as extra features of each text message(email or blog entry) and clustering them cannot separate similar events. As a result, it is a challenge to utilize the embedded social and temporal information with each piece of text content efficiently such that events with similar content can be distinguished using the social and temporal dimensions.

In this paper, we propose to detect events from social text stream data by exploring the embedded social network, temporal, and information flow patterns along with the text content. Specifically, we define an *event* as the information flow between a group of *social actors* on a specific *topic* over a certain *time period*. Rather than combining the temporal and social dimensions with the content dimension, which cannot fully reflect the contexts if used as additional features, we propose to utilize the temporal, information flow patterns, and social network within and across the content-based event detection results to further refine the quality of detected events.

The overview of our event detection approach is presented in Figure 1. First, the social text streams are represented as a graph of text pieces connected by content-based document similarity and they are clustered into a set of topics using graph cut algorithm (Shi & Malik 2000). As a result, each email or blog entry belongs to a topic. For a given topic, the social network is constructed based on the embedded social relations such as the *sender-recipient* and *author-commenter* relations. Then, for topic-based social network, the graph is partitioned into a sequence of graphs based on the *intensity* along the temporal dimension. That is, each graph in the temporal dimension, for a given topic, represents a communication peak (intensive discussion) that corresponds to a specific aspect or a smaller event. *Note that, the communication in the email dataset is the email message itself, whereas in the blog dataset, the communication is the comment made between bloggers.* After that, each graph in a specific time window with respect to a specific topic is converted into its dual graph and the dual graph is further partitioned into a set of smaller graphs based on the dynamic time warping (Keogh 2002) based information flow pattern similarity between social actor pairs using graph cut algorithm (Shi & Malik 2000). The intuition is to separate different communities for a specific topic. Lastly, the output of each event will be represented as a graph of social actors connected via a set of emails or blog comments during a specific time period about a specific topic.

Experimental results with the Enron dataset and the Dailymail blog dataset show that: (1) exploring the social and temporal dimensions with content can improve the event detection quality compared with the pure content-based approach, (2) the proposed stepwise event detection approach outperforms the approach that simply combines the all the dimensions together. Although we use email and blog data because of its easy availability, our work is easily extendible to other types of social text streams such as web forums, user groups, and USENET newsgroups.

In summary, the contributions of this paper are as follows: (1) we introduce the concept of *social text stream data* and propose the first approach to explore the temporal, text content, information flow pattern, and social network for event detection with an extended definition of event; (2) we proposed a stepwise event detection approach to utilize the content, temporal, information flow, and social network information and conducted extensive experiments.

The rest of this paper is organized as follows. Section 2 shows the proposed event detection approach. Experimental

results are presented in Section 3. In Section 4, we reviewed some related works and Section 5 concludes this paper.

## Event Detection From Social Text Stream

In this section, we present the details of our proposed event detection approach. First, the content-based clustering phase is presented, then, the temporal intensity based segmentation is discussed. After that, the information flow pattern is defined and our extended definition of an *event* is presented based on the content-based clustering results, temporal segmentation, and information flow over social network. Last, the event detection algorithm is explained.

### Content-Based Clustering

A collection of social text stream data can be represented as  $D = \langle (p_1, t_1, s_1), (p_2, t_2, s_2), \dots, (p_n, t_n, s_n) \rangle$ , where  $p_i \in P = \{p_1, p_2, \dots, p_{|P|}\}$  is a piece of text content that flows between the pair of *social actors*  $s_i$  at time point  $t_i$ ;  $s_i$  is a pair of social actors  $\langle a_i, r_i \rangle$  where  $a_i$  is the actor who initiates the information flow and  $r_i$  is the actor who receives/comments on  $p_i$ ;  $t_i$  is the timestamp the corresponding text piece was created.

Taken each piece of text  $p_i$  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  $\vec{p}_i = \langle w_1, w_2, \dots, w_n \rangle$ , where  $w_i$  is the weight of the  $i$ th word  $word_i$  in the vector. Here the weight of each word in the text piece is quantified as the *TF.IDF*, which is defined as:

$$w_i = \frac{|word_i|}{\sum_k word_k} \cdot \log \frac{|P|}{|p_i \supset word_i|}$$

where  $|word_i|$  the frequency of  $word_i$  in a text piece,  $k$  is the total number of words,  $|P|$  is the number of text pieces/emails. Note that the vector representation of each text pieces is constructed after removing stop words, stemming, and removing the signature of emails/blogs and quoted segments.

Given two text pieces,  $p_i$  and  $p_j$ , the content-based similarity is defined as the cosine similarity:

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

Then, the text stream corpus can be modeled as a graph, where each node is a text piece and each edge is the similarity between text pieces. Text pieces are clustered into different topics using the graph cut algorithm (Shi & Malik 2000). The graph cut based clustering 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, each piece of text belongs to a topic cluster in the graph cut-based result. Hereafter, we refer to each topic cluster as a topic.

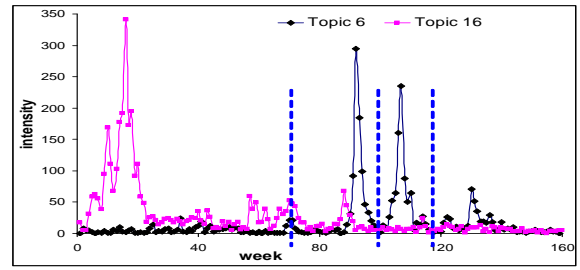


Figure 2: Topic Intensity Over Time and Segmentation

### Temporal Intensity-based Segmentation

For a given topic in the content-based text stream data clustering result, in real datasets, we observe that the *intensities* of the topic discussion change over time. Here the *intensity* of a topic at a time window is defined as the total number of text pieces created within the time window under the corresponding topic. For example, Figure 2 shows the intensities of 2 different topics extracted from the DailyKos dataset over 3 years on a weekly basis. Topic 16 is about *dean, howard, kerry, candidacy, and dnc*, topic 6 is about *plame, leak, CIA, Libya and fitzgerald*. It can be observed that under the same topic, the temporal dimension can be segmented to meaningful time intervals, which correspond to smaller events within this topic.

Given a sequence of intensities of a special topic  $\langle i_1, i_2, i_3, \dots, i_n \rangle$ , we adopt the adaptive time series model proposed by (Lemire 2007) to segment the sequence into a sequence of  $k$  intervals  $\langle I_1, I_2, I_3, \dots, I_k \rangle$  based on a sequence of segmentation indexes  $z_0, \dots, z_{k+1}$  such that  $\forall i_j \in I_m, z_m \leq i_j \leq z_{m+1}$ . The assumption is that elements within each interval  $I_m$  are generated from a model such as constant, linear, or quadratic. In this work, we assume the quadratic model of element distribution in the temporal dimension and the segmentation error is computed from  $\sum_i^k Q(I_m)$  where function  $Q$  is the regression error. Formally,  $Q(I_m) = \min_p \sum_{r=z_{m-1}}^{z_m-1} (p(r) - i_r)^2$  where the minimum is over the polynomials  $p$  of a given degree (Lemire 2007). We use the quadratic function  $p(x) = ax^2 + bx + c$  where  $a, b$ , and  $c$  are found by regression. By minimizing the segmentation error, the time series can be segmented into homogeneous segments. For example, the intensity sequence for topic 6 in Figure 2 is segmented into 4 segments by the dotted lines.

### Information Flow Pattern

As a result of the temporal segmentation, each topic is represented as a sequence of social network graphs over the temporal dimension. Within each social network graph, the weights of edges between different social actors denote the *communication intensity* of the corresponding social actors. Here communication intensity, denoted as  $CI_i^m(b_i, b_j)$ , refers to the number of communication text pieces between two social actors  $b_i$  and  $b_j$  under topic  $m$  within the  $n$ th time window. For instance, blogger  $b_1$  posted 4 comments to blogger  $b_2$  under topic 2 in the first week, then the communication intensity  $CI_1^2(b_1, b_2)$  is 4. Note that the commu-

nication intensity is directed. Based on the communication intensity, the *information flow pattern* can be defined:

**Definition 1. (Information Flow Pattern)** Given two social actors  $b_i$  and  $b_j$ , for a given topic  $m$ , the information flow pattern between them, denoted as  $F^m(b_i, b_j)$ , is defined as a vector of communication intensities.  $F^m(b_i, b_j) = F^m(b_j, b_i) = \langle CI_1^m(b_i, b_j), CI_2^m(b_i, b_j), \dots, CI_k^m(b_i, b_j), CI_1^m(b_j, b_i), CI_2^m(b_j, b_i), \dots, CI_k^m(b_j, b_i) \rangle$ .

That is, for a given topic, an *information flow-based social network* can be constructed between different social actors, where each vertex is a social actor and each edge represents the information flow pattern between the two social actors it connects. Given three social actors,  $b_1$ ,  $b_2$ , and  $b_3$ , if they are interested in the same aspect or subtopic under topic  $m$ , their information flow patterns should be similar to each other. That is,  $F^m(b_1, b_2)$ ,  $F^m(b_1, b_3)$ , and  $F^m(b_3, b_2)$  should be similar to each other.

From the real blog and email datasets, for two pairs of similar social actors, we observed that their information flow patterns have the same overall trend, but they have different scales and one of them have some offsets (delay in the temporal dimension) to the other. As a result, we define the similarity between two information flow patterns using the dynamic time warping concept (Keogh 2002). Given two information flow patterns  $F^m(b_i, b_j)$  and  $F^m(b_i, b_l)$ , in the dynamic time warping approach, a  $2k \times 2k$  matrix  $A$  is built, where element  $a_{pq}$  denotes the distance between the  $p$ th element in  $F^m(b_i, b_j)$  and the  $q$ th element in  $F^m(b_i, b_l)$  (typically the Euclidean distance is used). A warping path,  $\mathcal{W}$ , is a set of continuous elements in the matrix that defines a mapping between the two information flow patterns. The  $t$ th element of  $\mathcal{W}$  is defined as  $w_t = (p, q)_k$ . Then the dynamic time warping based similarity between two information flow patterns  $F^m(b_i, b_j)$  and  $F^m(b_i, b_l)$  is defined as:

$$Sim_F^m(F^m(b_i, b_j), F^m(b_i, b_l)) = 1 - \min \frac{1}{2k} \sqrt{\sum_{z=1}^{2k} w_z}$$

To identify different relations between different social actor pairs, we propose to convert the information flow-based social network graph  $G$  into its dual form  $G'$ . That is, in  $G' = (V', E')$  each vertex represents a pair of bloggers  $b_i$  and  $b_j$  and each edge represents the similarity between the two pairs of bloggers it connects. Here, the similarity between two pairs of bloggers, under a specific topic, is defined as the dynamic time warping based similarity between the corresponding information flow patterns.

Then, the graph cut algorithm is applied to the dual graph to partition social actors into different groups under a specific topic (Shi & Malik 2000). Note that there are two advantages of performing the graph cut on the dual graph: (i) in the original information flow based social network, each edge is a vector, which existing graph cut or clustering algorithms does not support; (ii) graph cut on the dual graph allows one social actor to belong to multiple clusters.

### Event Definition and Detection Algorithm

In traditional 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 with respect to the above mentioned content-based similarity (Yang & Shi 2006). We extend that definition of an event by introducing the social and temporal contexts of text streams.

**Definition 2. (Event)** Given a social text stream corpus denoted as  $D = \langle (p_1, t_1, s_1), (p_2, t_2, s_2), \dots, (p_n, t_n, s_n) \rangle$ , an *event* is defined as a subset of triples  $\mathbb{M} = \{(p_i, t_i, s_i), (p_{i+1}, t_{i+1}, s_{i+1}), \dots, (p_l, t_l, s_l)\}$  such that: (1)  $\forall p_i, p_j \in P_{\mathbb{M}} = \{p_i, p_{i+1}, \dots, p_{|\mathbb{M}|}\}$  belongs to the same topic cluster based on the content-based text clustering results; (2)  $\forall$  any timestamp in  $\langle t_i, t_{i+1} \dots t_j \rangle$  is within the same time interval  $I_n$ , which is one of the time segments in the temporal intensity-based segmentation results; and (3) each pair of social actors  $s_t \in S_{\mathbb{M}} = \{s_i, s_{i+1}, \dots, s_l\}$  belongs to the same cluster among the graph cut results on the dual graph of the information flow pattern based graph.

That is, an *event* is represented as a set of text pieces with semantical, temporal, social and information flow pattern constraints. Different from previous event detection approaches, which only explore one or two of the above constraints, our event detection not only identify events at a finer granularity and in a more meaningful manner, but also events in our detection results can be organized along the three dimensions to fully explore their relations, which can be useful in many applications.

Based on the above definition of event, content-based text clustering, temporal intensity based segmentation, and information flow pattern based partition, our event detection algorithm is shown in Algorithm 1. The input is a collection of social text stream data and the goal is to extract a list of events. First, each text piece is represented as a vector of words using the vector space model (Wong & Raghavan 1984). Then, these text pieces are clustered into topics  $C$  based on the *TF.IDF* similarity using the graph cut approach (Lines 1–4). Within each cluster/topic, the temporal intensity of blog entries or email messages is constructed and text pieces within this topic are partitioned into a sequence of time intervals  $I$  based on the trends of intensity in the time dimension using the adaptive time series segmentation technique (Lemire 2007). After that, the information flow pattern between two social actors under this topic is calculated as a vector, and the information flow based social network  $G$  is constructed. Based on the information flow pattern and the dynamic time warping similarity (Keogh 2002), social actors are clustered into different groups  $\mathcal{G}$  by converting the information flow pattern. Then, within each time segment  $I_l$ , the social actors are partitioned into smaller groups with respect to  $\mathcal{G}$ . As a result, the triple  $\langle c_i, I_l, G'_j \rangle$  contains a set of text pieces and social actors that talks about the same topic over a certain time interval with similar information flow patterns is added into the event list. This process iterates till all the events are detected and finally returned.

## Performance Study

### Dataset

There are different examples of social text stream data such as blogs, discussion forum, and emails. In this paper, the Enron Email dataset (Klimt & Yang 2004) and the Dailymkos blog dataset are used. The raw Enron corpus con-

### Algorithm 1 Event Detection from Social Text Streams

**Input:**

A social text stream:  $D = \langle (p_1, t_1, s_1), \dots, (p_n, t_n, s_n) \rangle$

**Output:**

A set of detected events:  $E$

**Description:**

```

1: for all  $p_i \in D$  do
2:   generate the word vector representation  $\vec{p}_i$  using TF.IDF
3: end for
4: clustering  $\vec{p}_1, \dots, \vec{p}_n$  content into clusters:  $C = c_1, \dots, c_m$ 
5: for all  $c_i \in C$  do
6:   construct the temporal intensity for  $c_i$  over time
7:   segment  $c_i$  into a set of intervals  $I = I_1, \dots, I_k$ 
8:   calculate the information flow pattern for each actor pair
9:   for all  $I_l \in I$  do
10:    construct the information flow based social network  $G$ 
11:    convert  $G$  into its dual graph  $G'$ 
12:    partition  $G'$  into  $\mathcal{G} = \{G'_1, \dots, G'_t\}$  using  $Sim^i_{\mathcal{F}}$ 
13:    for all  $G'_j \in \mathcal{G}$  do
14:       $E = E \cup \langle c_i, I_l, G'_j \rangle$ 
15:    end for
16:   end for
17: end for
18: return  $E$ 

```

tains 619,446 messages belonging to 158 users with an average of 757 messages per user from year 1998 to year 2002. Each message is a plain text file and these messages are organized based on the network references(email addresses). Emails that have multiple recipients are taken as multiple emails between any pair of sender and recipient. The Dailykos was crawled from the dailykos political blog website and it contains 249543 blog entries created by 18994 authors and 6026547 comments to these entries from 34975 individuals from October 12, 2003 to October 28, 2006. Among these 18994 authors the number of average blog entries is 13 in 1111 days and the average number of comments for each blog entry is 24.

### Evaluation

To evaluate the performance of our event detection approach, we manually labelled 30 events as the ground truth by looking into the email dataset, blog dataset, and corresponding real world events in news sources. 15 example events are presented in Table 1, where each event is represented as a triple of time interval, keywords, and social actors. Among these 15 examples, the first 11 is from the Enron emails and the rest is from the Dailykos blogs.

To measure the quality of our event detection results, we use the average micro- $F_1$  and average macro- $F_1$ . Given a detected event and a labelled event, let  $a$  is the number of true positive blog/email posts,  $b$  is the number of false positive blog/email posts,  $c$  is the number of false negative blog/email posts, and  $d$  is the number of true negative blog/email posts. Then, given  $n$  events, the average micro- $F_1$  and average macro- $F_1$  are define as:

$$\begin{aligned}
 MicF_1 &= \frac{2 \times \sum_{i=1}^n a_i}{2 \times \sum_{i=1}^n a_i + \sum_{i=1}^n (b_i + c_i)} \\
 MacF_1 &= \frac{1}{n} \times \sum_{i=1}^n \frac{2 \times a_i}{2 \times a_i + b_i + c_i}
 \end{aligned}$$

Time Interval	Key terms				#Social actors
Year 2000,	exert	senate	hear	utility	21
Week 42-46	market	margaret	send	favor	12
	contract	barwatt	change	fuel	7
Year 2001,	vote	bill	assemble	passage	54
Week 2-6	left	mind	update	summary	9
	mdq	gpm	west	receipt	6
Year 2001,	chri	joni	germany	weselack	37
Week 8-14	home	georg	margarita	ebiz	11
	uglier	dreaming	plan	reliant	7
Year 2001,	todd	strauss	avena	wisconsin	12
Week 28-32	legisl	activity	update	AY	9
week 31-35	prison	torture	ghraib	detain	122
week 50-54	Iran	nuclear	Korea	weapon	76
week 92-96	hillary	clinton	2008	rodham	99
week 135-139	cell	stem	disease	flu	58

Table 1: Examples of Labelled Events

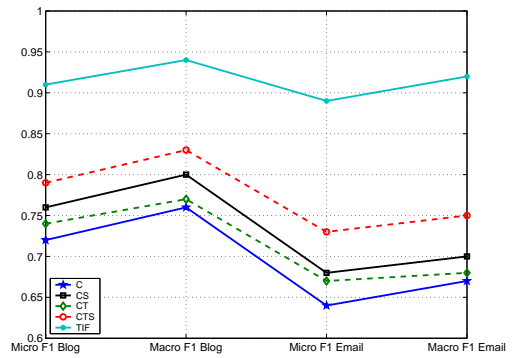


Figure 3: Performance Comparison Using  $F_1$  Measure

For each labelled event, the corresponding event with the highest  $F_1$  score in our detection results is selected and the average micro- $F_1$  and average macro- $F_1$  in the blog dataset and email dataset are presented in Figure 3. Five sets of experimental results are presented,  $C$  denotes the content-based event detection approach,  $CT$  denotes the content and temporal based event detection by taking the timestamps as an extra feature,  $CS$  denotes the content and social based event detection by taking the social dimension as an extra feature,  $CTS$  denotes the content, temporal, and social based event detection approach by taking the temporal and social dimensions as two extra features,  $TIF$  denotes our proposed temporal and information flow pattern based event detection approach. It can be observed that taking the temporal and social dimensions into the detection account can improve the  $F_1$  score significantly. However, the improvement of these approaches that takes temporal and social dimensions as extra features is not as significant as our proposed approach that utilizes temporal and social information in a stepwise manner.

Table 2 shows the average values of micro  $F_1$  and macro  $F_1$  on the two datasets with respect to the 30 labelled events for the five event detection approaches. It is obvious that our approach is substantially better than the content-based approach and other alternatives.

	<i>C</i>	<i>CS</i>	<i>CT</i>	<i>CTS</i>	<i>TIF</i>
Micro $F_1$	0.70	0.75	0.73	0.79	<b>0.90</b>
Macro $F_1$	0.74	0.80	0.78	0.82	<b>0.93</b>

Table 2: Average Event Detection Results

## Related Work

In the literature there are two lines of related work: text stream data classification and social network analysis of text streams. In (McCallum & Huang 2004), a set of benchmark experiments about email classification has been conducted on Enron and SRI corpora. The authors concluded that email classification is a challenging task by comparing most existing email classification approaches (Cohen 1996; Kiritchenko & Matwin 2001). Other works mainly focused on exploring different features (Carvalho & Cohen 2005; Krause, Leskovec, & Guestrin 2006). As for blog data, approaches have been proposed based on simple counts of entries, links, keywords, and phrases from the content point of view (Kumar *et al.* 2003). Prior research addressed name extraction, name disambiguation, social network extraction, and role discovery in the email social network (Leuski 2004) in email corpus. For blog data, prior works addressed burst detection (Kumar *et al.* 2003), trend detection (Chi, Tseng, & Tatemura 2006), blog spam (Kolari *et al.* 2006), theme pattern and life cycle extraction (Mei *et al.* 2006), social network analysis (Kumar, Novak, & Tomkins 2006; Qamra, Tseng, & Chang 2006), and structural and topic evolution/flow pattern extraction (Metzler *et al.* 2005; Qi & Candan 2006; Song *et al.* 2006).

## Conclusion and Future Works

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 that can be exploited to produce better results than existing state-of-the-art event detection approaches. Not only did we produce finer granularity events, but also explored the social, temporal, content dimensions to give better summarizations about the extracted events can be further explored to reason the transition and hidden relations between them.

## 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.

## References

Carvalho, V. R., and Cohen, W. W. 2005. On the collective classification of email "speech acts". In *SIGIR*, 345–352.

Chi, Y.; Tseng, B. L.; and Tatemura, J. 2006. Eigen-trend: trend analysis in the blogosphere based on singular value decompositions. In *CIKM*, 68–77.

Cohen, W. W. 1996. Learning rules that classify E-mail. In *AAAI Spring Symposium on MLIR*, 124–143.

Keogh, E. J. 2002. Exact indexing of dynamic time warping. In *VLDB*, 406–417.

Kiritchenko, S., and Matwin, S. 2001. Email classification with co-training. In *CASCON*, 8.

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

Klimt, B., and Yang, Y. 2004. The enron corpus: A new dataset for email classification research. In *ECML*.

Kolari, P.; Java, A.; Finin, T.; Oates, T.; and Joshi, A. 2006. Detecting Spam Blogs: A Machine Learning Approach. In *AAAI*.

Krause, A.; Leskovec, J.; and Guestrin, C. 2006. Data association for topic intensity tracking. In *ICML*.

Kumar, R.; Novak, J.; Raghavan, P.; and Tomkins, A. 2003. On the bursty evolution of blogspace. In *WWW*, 568–576.

Kumar, R.; Novak, J.; and Tomkins, A. 2006. Structure and evolution of online social networks. In *KDD*, 611–617.

Lemire, D. 2007. A better alternative to piecewise linear time series segmentation. In *SDM*.

Leuski, A. 2004. Email is a stage: discovering people roles from email archives. In *SIGIR*, 502–503.

McCallum, A., and Huang, G. 2004. Automatic categorization of email into folders: Benchmark experiments on enron and sri corpora. Technical Report IR-418, CIIR, University of Massachusetts Amherst.

Mei, Q.; Liu, C.; Su, H.; and Zhai, C. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW*.

Metzler, D.; Bernstein, Y.; Croft, W. B.; Moffat, A.; and Zobel, J. 2005. The recap system for identifying information flow. In *SIGIR*, 678–678.

Qamra, A.; Tseng, B.; and Chang, E. Y. 2006. Mining blog stories using community-based and temporal clustering. In *CIKM*, 58–67.

Qi, Y., and Candan, K. S. 2006. Cuts: Curvature-based development pattern analysis and segmentation for blogs and other text streams. In *HYPERTEXT*, 1–10.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE TPAMI* 22(8):888–905.

Song, X.; Tseng, B. L.; Lin, C.-Y.; and Sun, M.-T. 2006. Personalized recommendation driven by information flow. In *SIGIR*, 509–516.

Wong, S. K. M., and Raghavan, V. V. 1984. Vector space model of information retrieval: a reevaluation. In *SIGIR*.

Yang, C. C., and Shi, X. 2006. Discovering event evolution graphs from newswires. In *WWW*.

Yang, Y.; Pierce, T.; and Carbonell, J. 1998. A study of retrospective and on-line event detection. In *SIGIR*, 28–36.