Event Detection with Burst Information Networks

Tao Ge[1,2]*, Lei Cui[3], Baobao Chang[1,2], Zhifang Sui[1,2], Ming Zhou[3]
[1]Key Laboratory of Computational Linguistics, Ministry of Education,
School of Electronics Engineering and Computer Science, Peking University, Beijing, China
[2]Collaborative Innovation Center for Language Ability, Xuzhou, Jiangsu, China
[3]Microsoft Research, Beijing, China
getao@pku.edu.cn, lecu@microsoft.com, chbb@pku.edu.cn
szf@pku.edu.cn, mingzhou@microsoft.com

Abstract

Retrospective event detection is an important task for discovering previously unidentified events in a text stream. In this paper, we propose two fast centroid-aware event detection models based on a novel text stream representation – Burst Information Networks (BINets) for addressing the challenge, following the *D2N2K (Data-to-Network-to-Knowledge)* paradigm. The BINets are time-aware, efficient and can be easily analyzed for identifying key information (centroids). These advantages allow the BINet-based approaches to achieve the state-of-the-art performance on multiple datasets, demonstrating the efficacy of BINets for the task of event detection.

1 Introduction

Retrospective Event Detection (RED) (sometimes called topic detection) is a core task for text stream analysis, which aims to detect events that are previously unknown to the system (Wayne, 1998; Rajaraman and Tan, 2001) and is useful for many applications such as text stream summarization and evolutionary analysis of events in both news and social streams.

docid	time	text
d_1	Jan 12, 2010	A 7.0 magnitude quake hits the impoverished Caribbean nation of Haiti, killing more than 200,000 people, injuring an estimated 300,000.
d_2	Feb 27, 2010	A huge magnitude 8.8 earthquake strikes near the coast of south-central Chile, shaking buildings, causing blackouts and killing at least 147 people.
d_3	Apr 14, 2010	A 7.1-magnitude earthquake struck Tibetan Autonomous Prefecture of Yushu in southern Qinghai Province on April 14, 2010, killing at least 400 people and injuring more than 10,000.

Table 1: Documents discussing different earthquake events.

Most previous event detection approaches tend to use document- or keyword-based clustering models. Another solution proposed in recent years is to build a keyword graph to model the co-occurrence of keywords for detecting keyword communities as events (Sayyadi and Raschid, 2013). Even though these methods can achieve fair performance in small datasets, they have either of the following limitations:

- No time-awareness: many event detection models do not take into account time information. As a result, it is very likely that the documents that talk about different events (as Table 1 shows) are grouped into one cluster just because their lexical similarity is high.
- Inefficient: clustering-based methods tend to be time-consuming. For example, the time complexity of GAC (group average clustering) – the most commonly used clustering method in event detection – is $O(n^2 \log n)$. The computational challenge makes them difficult to work on a large dataset.

*This work was done when the first author was visiting Microsoft Research Asia

- Deviation of cluster centroids: it is likely that the clusters obtained by the methods are not event-centric, which has an adverse effect on the result, as illustrated in Figure 1.

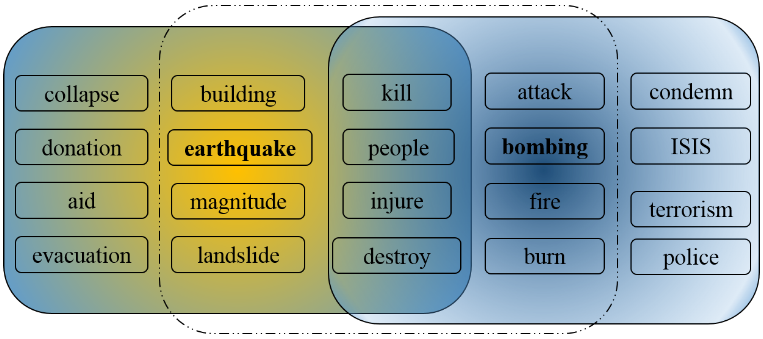

Figure 1: Deviation of cluster centroids: If clusters are not constructed around the centroid of the events (e.g., the dashline cluster is constructed around non-centroids such as *people*, *kill* and *injure* instead of *earthquake* or *bombing*), the performance will be adversely affected.

To offer a better solution to event detection without the above limitations, we propose to use a novel text stream representation: Burst Information Networks (BINets) (Ge et al., 2016a; Ge et al., 2016b). In contrast to the keyword graph which is based on word co-occurrence, a BINet is constructed based on burst co-occurrence. In a BINet (Fig. 2), a node is a burst of one word, which can be represented by the word with one of its burst periods, and an edge between two nodes indicates how strongly they are related (i.e., how frequently they co-occur). Since the nodes in a BINet contains temporal information (e.g., burst period), a BINet is time-aware in which nodes in a community are both topically and temporally coherent. Hence, we can say each community in a BINet corresponds to an event. Based on the BINet representation, we propose two fast centroid-aware event detection models. We show that the BINet-based models are efficient, allowing it to work on a large dataset, and the clusters obtained by the models center around the key information of events. Experiments on multiple datasets show that the BINet-based approaches achieve the state-of-the-art performance in terms of both accuracy and efficiency.

The contributions of this paper are:

- We propose to use BINets – a novel text stream representation for event detection, which is time-aware, can be efficiently built and support event-centric clustering, addressing the typical limitations of previous models.
- We propose two fast centroid-aware algorithms for event detection based on the BINet representation, which not only solve the centroid deviation problem but also are more efficient than traditional approaches.
- We construct and release a dataset for evaluating event detection models on a large text stream during a long time span.

2 Burst Information Networks

2.1 Burst Detection

A word's burst refers to a sharp increase of word frequency during a period. It usually indicates key information, important events or trending topics in a text stream as Figure 3 shows and is useful for many applications. In this paper, we detect a word's burst using the method of Zhao et al. (2012) which is a variant of (Kleinberg, 2003) and models burst detection as a burst state sequence decoding problem where a word w's burst state $s_t(w)$ at time t could be 1 or 0 to indicate if the word bursts or not at t. Specially, if a word w bursts at every time epoch during a period, we call this period a burst period of w and w has a burst during this period. In Figure 3, *earthquake* has 2 burst periods (i.e., Jan 12 - Jan 31, and Feb 27 - Mar 7), which correspond to two famous earthquake events (i.e., 2010 Haiti earthquake and 2010 Chile earthquake).

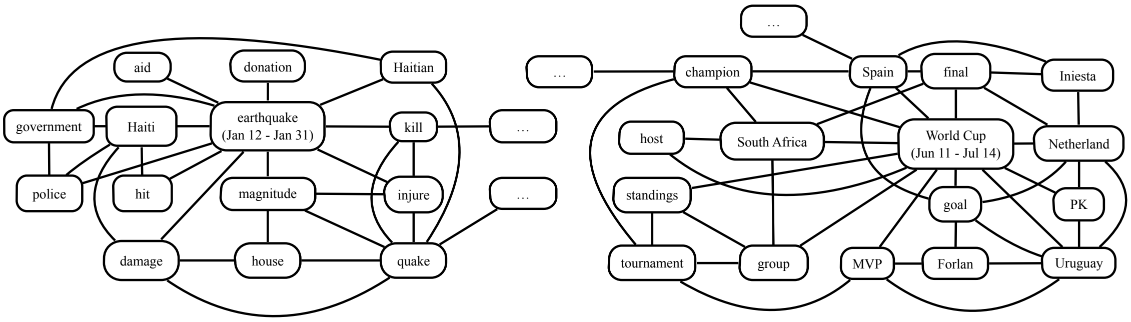

Figure 2: An example of Burst Information Network.

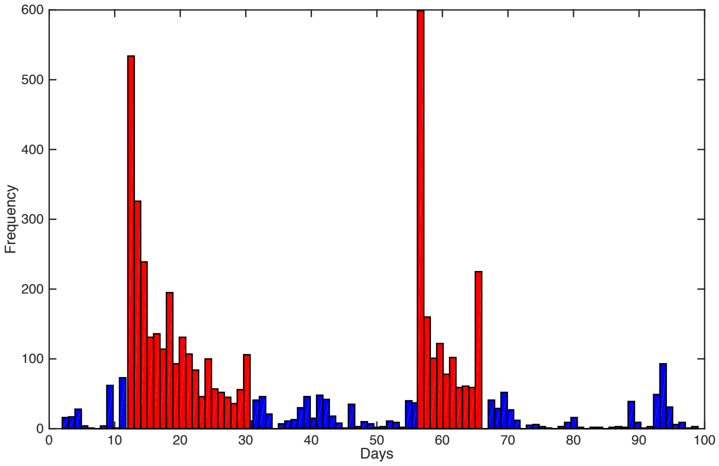

Figure 3: The frequency of *earthquake* during the first 100 days in 2010. There are two burst periods (red) for *earthquake* during the period, corresponding to two strong earthquake events happening in Haiti and Chile respectively.

Formally, we define $\mathcal{P}_i(w)$ as the ith burst period of the word w. It is a time interval, during which the word w bursts at every time epoch:

$$\mathcal{P}_i(w) = [t_i^s(w), t_i^e(w)]$$
$$\forall t \in \mathcal{P}_i(w) \;\; s_t(w) = 1$$

where $t_i^s(w)$ and $t_i^e(w)$ denotes the starting and ending time of the ith burst period of w, and $s_t(w)$ denotes the burst state of w at time t.

2.2 Burst Information Network Construction

A BINet represents associations between key facts in a text stream, which has been proven to be effective in multiple knowledge mining tasks (Ge et al., 2016a; Ge et al., 2016b). The basic component of a BINet is burst elements which are nodes of the information network:

A Burst Element is a burst of a word. It can be represented by a tuple: $\langle w, \mathcal{P}_i(w) \rangle$ where w denotes the word and $\mathcal{P}_i(w)$ denotes one burst period of w. Though a word may have multiple burst periods, a burst element has only one burst period. A word during its different burst periods will be regarded as different burst elements.

There are two main advantages using burst elements as nodes to build the information network:

- A burst element not only includes semantic information but also incorporates the temporal dimension. Nodes in a community are topically and temporally coherent while nodes that are topically or temporally distant cannot be adjacent, which makes it reasonable to consider a community in a BINet corresponds to an event.
- Since a burst element denotes a burst word during one of its burst period, its sense is likely to be consistent. Multiple bursts of a word will be considered as different burst elements. Therefore,

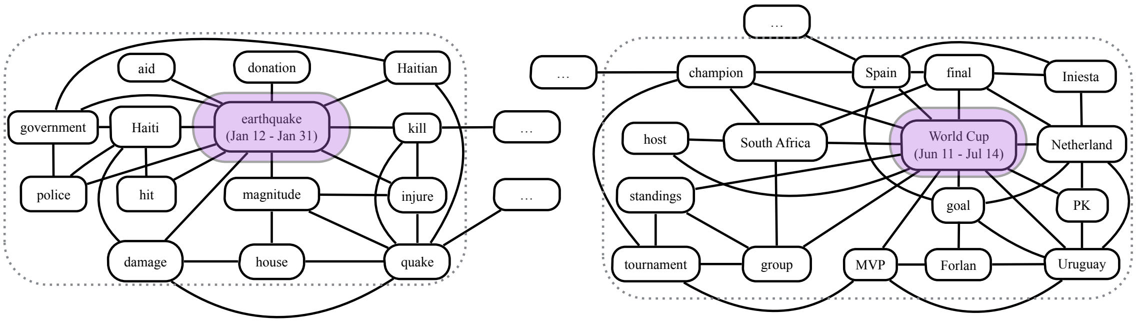

Figure 4: Node based detection model. Shaded nodes denote key nodes.

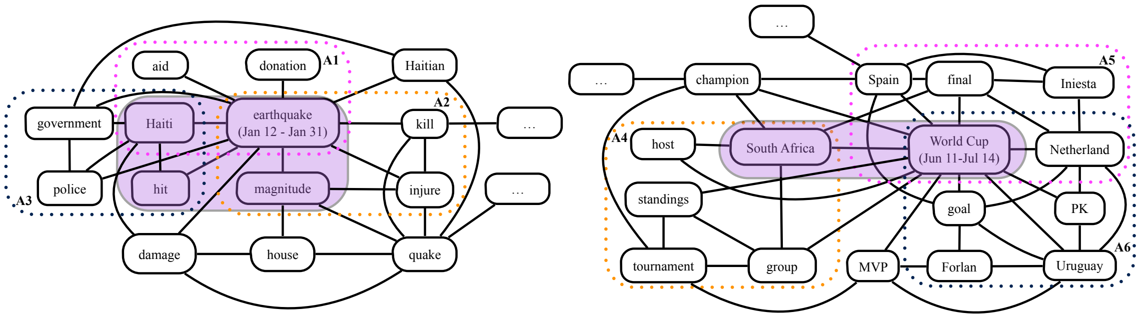

Figure 5: Area based detection model. Shaded areas denote key areas.

nodes in a BINet tends to be less ambiguous.

Formally, a BINet is defined as $G = \langle V, E \rangle$. Each node $v \in V$ is a burst element and each edge $e \in E$ denotes the association between burst elements. Intuitively, if two burst elements frequently co-occur, then they should be highly weighted. We define $\omega_{i,j}$ as the weight of an edge between v_i and v_j, which is equal to the number of documents where v_i and v_j co-occur.

3 Event detection based on the BINet

3.1 Motivation

The goal of event detection is to organize a text stream into multiple document sets, in each of which the documents coherently discuss the same event. The traditional clustering methods are usually inefficient and not time-aware. Moreover, they tend to suffer from the problem of deviation of cluster centroids, as illustrated in Figure 1. In Figure 1, *earthquake* and *bombing* are centroids (i.e., key information) of an earthquake event and a bombing event respectively. If clusters are constructed around the centroids (e.g., solid line clusters), the performance will be good; while if clusters center around non-centroid nodes (e.g., the dashline cluster centers around *kill* and *people*), the results will be poor.

To address the limitations above, we propose to model event detection problem as community detection on the BINet in which each community is both topically and temporally coherent, corresponding to one event. Instead of using popular community detection algorithms in social network analysis whose time complexity is high, we propose two fast centroid-aware event detection model: node-based detection model (NDM) and area-based detection model (ADM). Both of the approaches first identify the key nodes (or key areas) on the BINet, which indicate the centroid (i.e., key information) of events in the text stream, and then construct clusters that center around the key nodes (or key areas). The difference of the models is that NDM attempts to detect a bunch of node communities as clusters while ADM detects the overlapping document areas to form document clusters, as Figure 4 and Figure 5 depict. In some sense, NDM and ADM correspond to the keyword- and document-based clustering model respectively. In the following sections, we will present the details of NDM and ADM.

Community	The word of nodes with top PageRank value	Event
1	Iraq, war, Iraqi, US-led, Baghdad	Iraq war in 2003
2	flu, a/h1n1, health, virus, influenza	2009 A/H1N1 flu pandemic
3	earthquake, quake, Sichuan Province, Sichuan, quake-hit	Sichuan earthquake in 2008
4	Beijing, Olympic_Games, gold, medal, team	2008 Beijing Olympics
5	financial, crisis, global, economy, economic	financial crisis in 2008

Table 2: Example of the communities detected by our approach. Each community corresponds to one event and nodes with the top PageRank values tend to be keywords that are the most situable to describe the events.

3.2 Centroid-aware event detection models

3.2.1 Node-based detection model

The goal of node-based detection model (NDM) is to detect node communities on the BINet each of which corresponds to one event. To guarantee that detected communities center around the key nodes that correspond to key information (i.e,. centroid) of events in the text stream, we first identify the key nodes on the BINet.

Owing to the BINet representation, it is easy to identify the key nodes through the analysis of the network. Among a variety of ways to identify the influential nodes in a network, we simply adopt the Pagerank algorithm (Page et al., 1997). For a node v, its PageRank value $pr(v)$ is computed as follows:

$$pr(v) = d \sum_{v' \in N(v)} \hat{\omega}_{v,v'} \times pr(v') + \frac{1-d}{|V|}$$

where $|V|$ is the number of nodes in the BINet, $N(v)$ denotes the set of nodes adjacent to v, d is the damping factor and is set to 0.85, $\hat{\omega}_{v,v'} = \frac{\omega_{v,v'}}{\omega_{v',*}}$, which is the normalized weight of the edge between v and v'.

Intuitively, a node with a high PageRank value is usually important and likely to be the key node that indicates the key information of an event. Therefore, we rank nodes in the BINet by their PageRank value and choose the node which has the highest PageRank value and does not belong to any community as a key node to construct a community $\mathcal{E}$ around it with its closely related nodes:

$$\mathcal{E} = \{v\} \cup \{u | \hat{\omega}_{v,u} > \sigma_N\}$$

where v is the node with the highest PageRank value and does not belong to any community, $\hat{\omega}_{v,u}$ is the normalized weight of the edge between v and u, and σ_N is the threshold for selecting closely related nodes.

By repeating the process, we can detect multiple communities on the BINet efficiently, each of which centers around a key node. Table 2 shows some communities detected by this approach from 1995-2010 Xinhua news in English Gigaword. One can observe that each community corresponds to one event and nodes with the top PageRank values in a community tend to be key information of the events. We summarize the algorithm in Algorithm 1.

For NDM, we need to infer a document's event after community detection. For a document d, we infer the probability that d discusses the event e_k as follows:

$$P(e_k|d) = \frac{\sum_{v_k \in V_k(d)} pr(v_k)}{\sum_{v \in V(d)} pr(v)} \tag{1}$$

where $V(d)$ denotes the set of nodes that the words of d corresponds to in the BINet, $V_k(d) \subset V(d)$ denotes a subset of $V(d)$ that are in the community of the event e_k, and $pr(v)$ is the PageRank value of a node v. In Eq (1), the PageRank values of nodes in $V(d)$ can be considered as weights. The nodes with high PageRank values are highly weighted because they tend to indicate important topical and event information.

Algorithm 1 Node-based detection model

```
Input: Ranked list of nodes by PageRank value: $\mathcal{L}$, BINet: $G = \langle V, E \rangle$;
Output: A list of event communities: $\mathcal{C} = [\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_k]$
while $\|\mathcal{L}\| > 0$ do
    $v \leftarrow \mathcal{L}[0]$ (the first element in $\mathcal{L}$)
    $\mathcal{E} \leftarrow \{v\} \cup \{u | \hat{\omega}_{v,u} > \sigma_N\}$
    $\mathcal{L} \leftarrow \mathcal{L} - \mathcal{E}$
    $\mathcal{C}.append(\mathcal{E})$
end while
```

3.2.2 Area-based Detection Model

A document area is the area (i.e., a set of nodes) on the BINet a document corresponds to. For example, $A3$ in Figure 5 is the area that the document written during the Haiti earthquake about *Haiti*, *government* and *police* corresponds to on the BINet. The idea of area-based detection model (ADM) is discovering the document areas that massively overlap on the BINet to construct clusters so that the documents whose areas are in the same cluster are about the same event. In contrast to NDM in which each item in a cluster is a node, the items in a cluster obtained by ADM is document areas on the BINet.

To guarantee that the clusters center around the centroids of events, we first identify key nodes on the BINet, as NDM does. In ADM, however, we treat a key area as the centroid of an event, which is different from NDM that treats a key node as an event centroid. To identify the key areas on the BINet, we first define the PageRank score of an area A as the normalized sum of the PageRank value of the nodes in it:

$$pr(A) = \frac{\sum_{v \in A} pr(v)}{\sqrt{|A|}}$$

Then, we repeatedly choose the area which has the highest PageRank score and does not belong to any cluster as a key area to construct a cluster with the areas that massively overlap it:

$$\mathcal{E} = \{A\} \cup \{A' | f(A, A') > \sigma_A\} \tag{2}$$

where σ_A is the threshold to construct cluster, $f(A, A')$ is a score to indicate how much A overlaps A' and it is computed as follows:

$$f(A, A') = \frac{|A \cap A'|}{|A \cup A'|} \tag{3}$$

We summarize the algorithm of ADM in Algorithm 2. As NDM, ADM detects events in a greedy manner; hence, the detection process is fast. However, in contrast to NDM, ADM allows one area to belong to multiple communities, which means that one document could belong to multiple events.

Algorithm 2 Area-based detection model

```
Input: Ranked list of documents areas: $\mathcal{L}$, BINet: $G = \langle V, E \rangle$;
Output: A list of event communities: $\mathcal{C} = [\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_k]$
while $\|\mathcal{L}\| > 0$ do
    $A \leftarrow \mathcal{L}[0]$ (the first element in $\mathcal{L}$)
    $\mathcal{E} \leftarrow \{A\} \cup \{A' | f(A, A') > \sigma_A\}$
    $\mathcal{L} \leftarrow \mathcal{L} - \mathcal{E}$
    $\mathcal{C}.append(\mathcal{E})$
end while
```

4 Experiments and Evaluation

We conduct experiments to evaluate the performance of our approach. We first evaluate our approach on the TDT4 dataset to compare other event detection approaches. Then, we apply our approach on a larger corpus (2009 – 2010 news corpus) to test its scalability and performance.

For preprocessing, we remove stopwords and conduct lemmatization and name tagging using Stanford CoreNLP toolkit (Manning et al., 2014) before the construction of a BINet.

4.1 Evaluation on TDT4

The TDT4 collection is a well known dataset for comparing methods for event detection. The English part of the dataset includes approximately 29,000 news documents from news agencies such as CNN and BBC from October 2000 to Janurary 2001 (spanning 4 months), while only 1,884 documents[1] are annotated to be related to 71 human identified events (topics). As the setting adopted by previous work (Li et al., 2005; Sayyadi and Raschid, 2013), we use the annotated subset as gold standard for evaluating the performance of our models.

As most of the previous work (Yang et al., 1998; Li et al., 2005) addressing the event detection challenge, we use Micro-Precision, Micro-Recall, Micro-F1 as well as Macro-F1 to evaluate the performance. We compare our approach to the following models whose effectiveness on the TDT4 corpus has been verified by previous work:

- Allan[2] (Allan et al., 1998): A popular online event detection model, which is often used as a baseline to compare event detection models.
- GAC (Yang et al., 1998): A classical but effective approach for event detection using group average clustering.
- KeyGraph (Sayyadi and Raschid, 2013): Betweenness score based community detection approach on KeyGraph. It is notable that the evaluation measures used in Sayyadi and Raschid (2013) are somewhat different from those in this paper and other work – they used Macro-precision, Macro-recall[3] and Macro-F1. We only report its Macro-F1 in Table 3.
- Probabilistic model (Li et al., 2005): A time-aware probabilistic graphical model for event detection. It is the state-of-the-art approach on TDT4 dataset.

Models	Micro-P	Micro-R	Micro-F1	Macro-F1
Allan	0.64	0.57	0.60	0.62
GAC	0.83	0.63	0.72	0.75
KeyGraph	-	-	-	0.69
Probabilistic Model	0.85	0.67	0.75	0.78
BINet-NDM	0.79	0.69	0.74	0.75
BINet-ADM	0.81	0.70	0.75	0.77

Table 3: Performance of various event detection approaches on TDT4.

Table 3 shows the results[4] on the TDT4 dataset. The BINet approaches perform well on the dataset: Both NDM and ADM outperform the classical baselines (i.e., Allan, GAC and Keygraph). The ADM performs better than NDM and even achieves the comparable performance to the state-of-the-art approach by (Li et al., 2005) because the centroid in ADM is a key area that contains more information than a key node in NDM. The reasons for the good performance are two-fold: First, the BINet-based approach is time-aware, which avoid many unnecessary mistakes made by the baseline models that only take into account text content; Second, the BINet-based models are centroid-aware, which guarantee that

[1] Among these 1,884 documents, there are 38 documents belonging to more than one event.

[2] This baseline is often referred as kNN in literature (Li et al., 2005; Sayyadi and Raschid, 2013). However, to avoid the ambiguity with the popular kNN classification model, we simply refer it as Allan.

[3] For reference, the Macro Precision and Recall reported in Sayyadi and Raschid (2013) are 0.82 and 0.59 respectively.

[4] The results of *Allan*, *GAC* and *Probabilistic Model* are from Li et al. (2005) while the results of *KeyGraph* come from Sayyadi and Raschid (2013).

Model	Micro-P	Micro-R	Micro-F1	Macro-F1	Running time
GAC	-	-	-	-	>2 hours
KeyGraph	-	-	-	-	>2 hours
Probabilistic model	-	-	-	-	>2 hours
B-GAC	0.81	0.65	0.72	0.67	7189s (896s)
BINet-NDM	0.85	0.68	0.76	0.69	3591.98 (1350.08s)
BINet-ADM	0.84	0.71	0.77	0.71	3610.03s (1368.13s)

Table 4: Performance and running time of various event detection models on the 2-year news stream. We do not report the precision, recall and f-score for the models that cannot get results within 2 hours. The number in the round bracket is the running time of the model when it is run in 8-way parallel. The running time is measured on a workstation with Intel Xeon 3.5 GHz CPU and 64GB RAM.

the generated clusters center around centroids of events and avoid the problem of deviation of cluster centroids.

4.2 Evaluation on a 2-year news stream

Even though TDT4 is a widely used dataset for event detection, it has several limitations: First, the period of TDT4 dataset is short (only 4 months) as Li et al. (2005) claimed. In TDT4 dataset, hardly can we see multiple events of the same type in the TDT4 dataset (e.g., there is only one flood event in TDT4 dataset). Therefore, even if we just use content-based clustering methods regardless of time information, the performance is not bad. Second, the data size of the TDT4 corpus is so small compared with a real text stream that many stream-based features such as burst cannot function as well as in a real stream. To test the performance of our detection models on a real text stream, we construct a dataset using 2009 – 2010 news from English Gigaword (APW and XIN sections) as a text stream where there are 584,414 news articles in total. We construct a BINet on this dataset, which contains 46,254 nodes and 514,682 edges. For evaluation, we select 83 events that happened during 2009 – 2010 and annotate their relevant documents in the text stream. The selected events are all important events and have their corresponding Wikipedia pages. The annotation process is similar to (Li et al., 2005): we use the Wikipedia title of the events to search the candidate documents using Lucene and then manually identify if the returned documents are actually relevant to the events. Since this annotation process does not guarantee finding all the relevant documents to an event, we call the annotations silver standard[5]. In total, there are 2,584 documents that are annotated as relevant to those 83 events.

Table 4 shows the results of various approaches on the 2-year news stream. Due to the size of the dataset, most traditional event models cannot finish the detection task within two hours since their time complexity is too high. The B-GAC model proposed by (Zhao et al., 2012) is the only one that can finish the task within 2 hours because it adopts the split-merge-clustering strategy that splits[6] the data into multiple small pieces by time for clustering and then merges the clusters. Though such a strategy can alleviate the issue of the scalability, the split of data will affect the global overview of the text stream and have an adverse effect on finding the centroids of events. In contrast, our BINet-based approaches can finish detecting events within 1 hour without splitting the stream and achieve the best result owing to their awareness of both time[7] and event centroids.

We compare the time complexity of our centroid-aware event detection models to other commonly used event detection approaches, as shown in Table 5 where n is the number of documents, K is the event number (K in our BINet-based approaches depends on the selection of σ_N and σ_A), $|W|$ is the size of vocabulary, and $|V|$ and $|E|$ are the number of nodes and edges on the BINet respectively. The running time of NDM and ADM consist of four parts: burst detection, BINet construction, PageRank analysis,

[5] The annotation data can found at http://getao.github.io

[6] We split the 2-year news stream into 8 small pieces, each of which is a 3-month news stream so that it can get the result within 2 hours.

[7] For our BINet-based approaches, only burst detection part is run in parallel in 8-way parallel setting, which is different from B-HAC that splits the stream and clusters documents in parallel.

Models	**Time complexity**
GAC	$O(n^2 \log n)$
B-GAC	$O(n^2 \log n)$
Keygraph	$O(nK + \lvert W\rvert^3)$
NDM	$O(nK + \lvert V\rvert \log \lvert V\rvert + \lvert E\rvert + \lvert W\rvert T)$
ADM	$O(n(\log n + L) + \lvert V\rvert + \lvert E\rvert + \lvert W\rvert T)$

Table 5: Time complexity of various event detection models.

and event detection. The first three parts are the same for NDM and ADM, which are preliminary steps for event detection. The time complexity of burst detection algorithm is $O(T)$ for one word where T is the time span of the stream and it can be conducted in parallel for different words because the burst detection processes for different words are independent. The time complexity of the BINet construction and PageRank analysis is $O(n)$ and $O(|V| + |E|)$ respectively. For the event detection part, the time complexity of community detection of NDM is $O(|V| \log |V| + |E| + nK)$. The former term is the time cost for sorting nodes by PageRank value, and the second and the third term are the cost for constructing node communities and assigning events to documents respectively. The time complexity of the detection part in ADM is somewhat different. Its time complexity is $O(n \log n + nL)$. As NDM, the first term is the time for sorting document areas by PageRank value. The second term is the time cost for computing Eq (3) in which L is the average number of times that a document (area) is taken for computing Eq (3) and is affected by the selection of threshold parameter σ_A. In the worst case, $L = K$; while in the best case, $L = 1$, meaning that a document is taken for computing Eq (3) only once. In most cases, L is a small number. The running time of those parts of NDM and ADM is shown in Table 6. Note that, for the part of the PageRank computation, the time is measured by running the PageRank algorithm for 1,000 iterations.

	BINet-NDM	**BINet-ADM**
Burst detection	2,562.17s (320.27s)	2,562.17s (320.27s)
BINet construction	304.56s	304.56s
PageRank computation	716s	716s
Event detection	9.25s	27.3s
Total	3591.98s (1350.08s)	3610.03s (1368.13s)

Table 6: The running time of 4 parts of our BINet-based event detection approaches. The number in the round bracket is the running time of the model when it is run in 8-way parallel.

5 Related Work

Event detection is one of the most popular research topics in recent years and has been extensively studied for the decades (Yang et al., 1998; Swan and Allan, 2000; Allan, 2002; Fung et al., 2005; He et al., 2007; Sayyadi et al., 2009; Zhao et al., 2012; Sayyadi and Raschid, 2013; Ge et al., 2015). They are based on either document- or keyword-based clustering, which usually suffer from either unawareness of time, high expensive computation cost or deviation of cluster centroids. In contrast, our approach is time-aware, centroid-aware and so efficient that it can be run on a large text stream.

In addition, there is much work (Sakaki et al., 2010; Lee et al., 2011; Diao et al., 2012; Aggarwal and Subbian, 2012; Wang et al., 2013; Dong et al., 2015) studying event detection problem in social media. They usually use more or less social media features such as spatio-temporal information, which are not in the same setting with our task.

6 Conclusion and Future Work

This paper proposes to use a novel text stream representation – Burst Information Networks to address the retrospective event detection challenge. Based on the BINet, we propose two fast centroid-aware event detection models that can effectively overcome the limitations of the previous event detection models and achieve the state-of-the-art performance on both TDT4 and a long-span text stream.

In the future, we plan to study events in a text stream more deeply based on the BINet representation. Since a BINet can offer a global overview of events in the stream level, we plan to use the BINets to derive an event's type, extract its schema and even fill its slots after we detect its corresponding regions on the BINet. Hopefully, this framework could work for endless event knowledge mining if it could be used for monitoring the massive text streams.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. This work is supported by the National Key Basic Research Program of China (No.2014CB340504), the Research Fund for the Doctoral Program of Higher Education (20130001110027) and the National Natural Science Foundation of China (No. 61375074, 61273318). The contact author is Zhifang Sui.

References

Charu C Aggarwal and Karthik Subbian. 2012. Event detection in social streams. In *SDM*.

James Allan, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *SIGIR*.

James Allan. 2002. *Topic detection and tracking: event-based information organization*, volume 12. Springer Science & Business Media.

Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. 2012. Finding bursty topics from microblogs. In *ACL*.

Xiaowen Dong, Dimitrios Mavroeidis, Francesco Calabrese, and Pascal Frossard. 2015. Multiscale event detection in social media. *Data Mining and Knowledge Discovery*, 29(5):1374–1405.

Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S Yu, and Hongjun Lu. 2005. Parameter free bursty events detection in text streams. In *VLDB*.

Tao Ge, Wenzhe Pei, Heng Ji, Sujian Li, Baobao Chang, and Zhifang Sui. 2015. Bring you to the past: Automatic generation of topically relevant event chronicles. In *ACL*.

Tao Ge, Lei Cui, Baobao Chang, Sujian Li, Ming Zhou, and Zhifang Sui. 2016a. News stream summarization using burst information networks. In *EMNLP*.

Tao Ge, Qing Dou, Xiaoman Pan, Heng Ji, Lei Cui, Baobao Chang, Zhifang Sui, and Ming Zhou. 2016b. Aligning coordinated text streams through burst information network construction and decipherment. *arXiv preprint arXiv:1609.08237*.

Qi He, Kuiyu Chang, and Ee-Peng Lim. 2007. Using burstiness to improve clustering of topics in news streams. In *ICDM*.

Jon Kleinberg. 2003. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

Chung-Hong Lee, Hsin-Chang Yang, Tzan-Feng Chien, and Wei-Shiang Wen. 2011. A novel approach for event detection by mining spatio-temporal information on microblogs. In *ASONAM*.

Zhiwei Li, Bin Wang, Mingjing Li, and Wei-Ying Ma. 2005. A probabilistic model for retrospective news event detection. In *SIGIR*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*.

Larry Page, S Brin, R Motwani, and T Winograd. 1997. Pagerank: Bringing order to the web. Technical report, Stanford Digital Libraries Working Paper.

Kanagasabi Rajaraman and Ah-Hwee Tan. 2001. Topic detection, tracking, and trend analysis using self-organizing neural networks. In *PAKDD*.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*.

Hassan Sayyadi and Louiqa Raschid. 2013. A graph analytical approach for topic detection. *ACM Transactions on Internet Technology (TOIT)*, 13(2):4.

Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. 2009. Event detection and tracking in social streams. In *ICWSM*.

Russell Swan and James Allan. 2000. Automatic generation of overview timelines. In *SIGIR*.

Xun Wang, Feida Zhu, Jing Jiang, and Sujian Li. 2013. Real time event detection in twitter. In *WAIM*.

Charles Wayne. 1998. Overview of tdt.

Yiming Yang, Tom Pierce, and Jaime Carbonell. 1998. A study of retrospective and on-line event detection. In *SIGIR*.

Wayne Xin Zhao, Rishan Chen, Kai Fan, Hongfei Yan, and Xiaoming Li. 2012. A novel burst-based text representation model for scalable event detection. In *ACL*.