CENTER FOR DATA SCIENCE AND BIG DATA ANALYTICS
Strength in Numbers

OAKLAND UNIVERSITY

Detecting the States of Emergency Events Using Web Resources

Vijayan Sugumaran, Ph.D.
Department of Decision and Information Sciences
School of Business Administration
Oakland University
sugumara@Oakland.edu

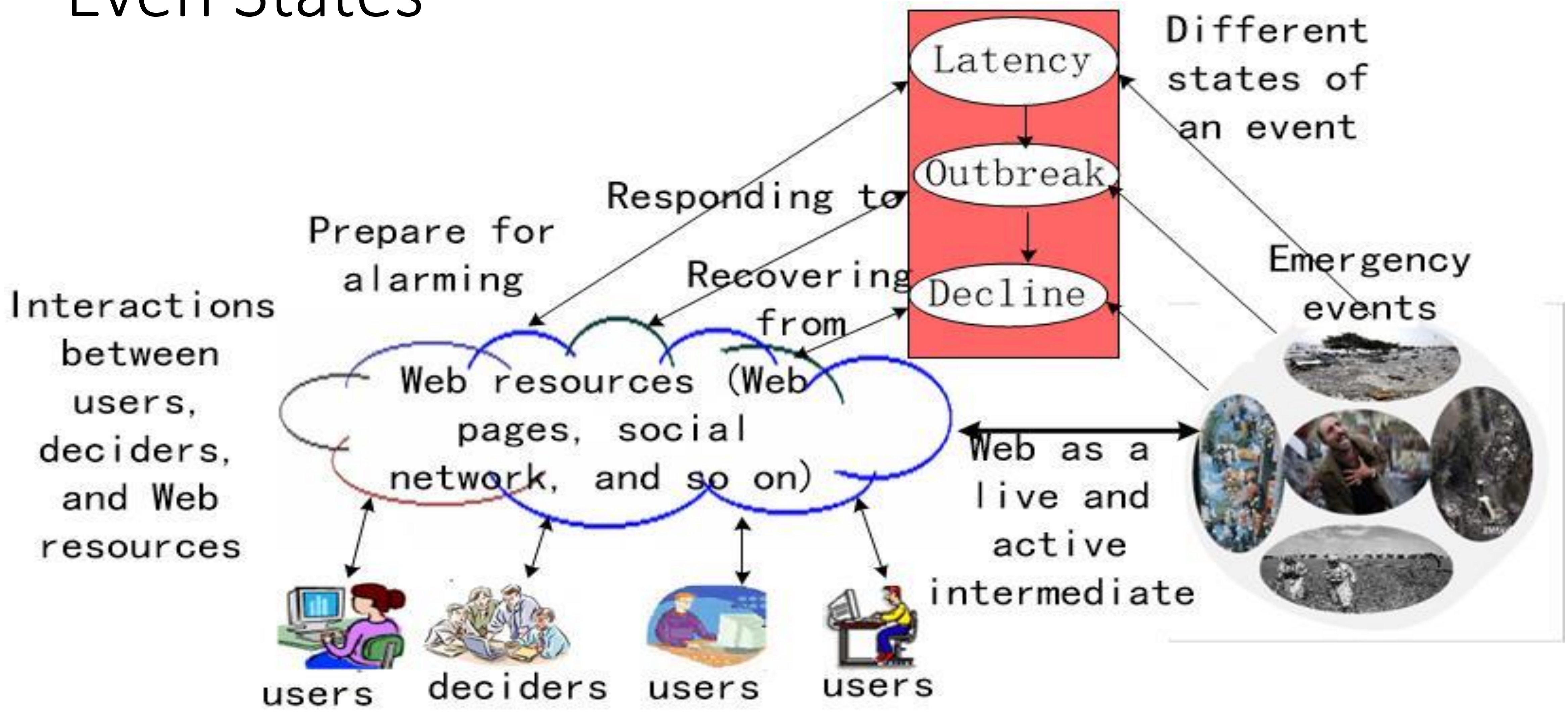
Collaborators

- The Third Research Institute of the Ministry of Public Security, Shanghai, China
- Tsinghua University, Beijing, China
- Shanghai University, Shanghai, China
- Department of Information Systems and Cyber Security, University of Texas at San Antonio, USA
- School of Information Technology & Mathematical Sciences, University of South Australia, Australia

Emergency Events

- Emergency events are inevitable
- Information about the events immediately available on the Web
- Social media sites play the role of information repositories
- Web information is dynamic – keeps up with the evolution of the emergency event
- “Event Evolution” generates large volume of temporal data
- This data can be mined to learn about the events, determine the state of the event, and explore ways to mitigate them

Even States


Latency
Outbreak
Decline
Different states of an event
Responding to
Prepare for alarming
Recovering from
Interactions between users, deciders, and Web resources
Web resources (Web pages, social network, and so on)
Web as a live and active intermediate
Emergency events
users
deciders
users
users

Research Objective

- Develop a new web mining approach for detecting the state of emergency events reported on the web
- For an emergency event, the related web resouces can be found, for example, web news, blogs, and forums
- Based on the content and semantics of these web pages, the temporal features of an event can be identified
- And then, the different states can be identified (latent, outbreak, decline, transition, and fluctuation)

States of Emergency Events

- Latent
  - Fewer web pages with event information
  - Prevention focus
- Outbreak
  - Event occurring
  - Response focus
- Decline
  - Waning of the event
  - Focus is on lessening the effects of the event
- Transition
  - State transition from one to the next
- Fluctuation
  - Variations within a state

Overall Approach

- Develop a set of algorithms for detecting the state of an emergency event reported on the web
- First, the related resources including web pages, keywords of an emergency event are collected using web search engines
- Second, the *outbreak power* and the *fluctuation power* of an emergency event at *timestamp "t"* are computed
- Based on the various temporal values, different states of an emergency event are inferred

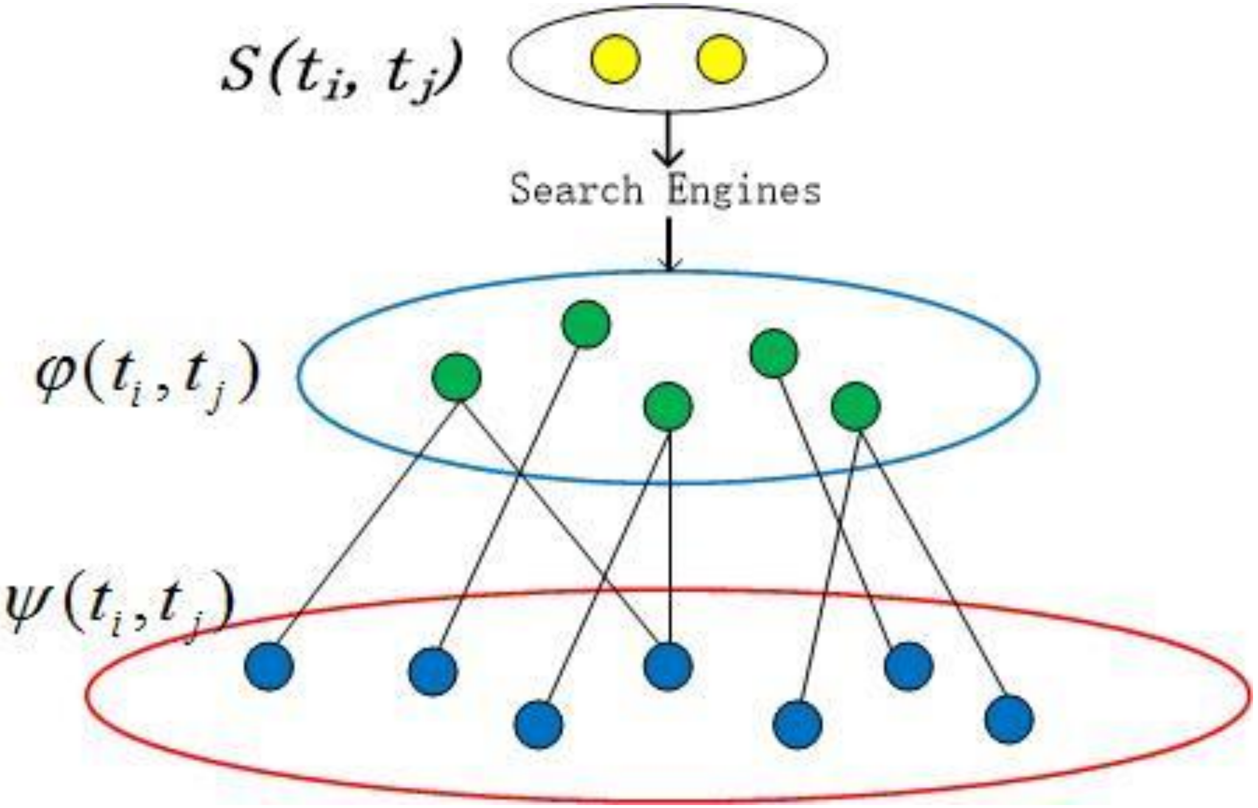
Keywords, Web Pages and Seed Sets

Input: Given an event e and a set of related features (e.g., web pages, event attributes), the starting timestamp is denoted as t_s, and the ending timestamp is denoted as t_e.

Output: A *k-period* S of e represented by $S = \{s_1, s_2, ..., s_k\}$, where s_i is a period of an emergency event. In other words, there are period boundaries $t_1, t_2, ..., t_{k-1}$, $t_s < t_1 < ... < t_{k-1} < t_e$, where $s_1 = (t_s, t_1), s_2 = (t_1, t_2), ..., s_k = s(t_{k-1}, t_e)$.

(1) Use $S(t_i, t_j)$ as the queries to search for related web pages, the returned web pages are denoted as $\varphi(t_i, t_j)$.

(2) Get $\psi(t_i, t_j)$ extracted from $\varphi(t_i, t_j)$, the weight is computed by TF-IDF (term frequency–inverse document frequency) scheme [19].



Temporal Features of Emergency Events

- Five basic temporal features:
  - Number of increased web pages
  - Number of increased keywords
  - Distribution of keywords on web pages
  - Associated relations of keywords, and
  - Similarities of web pages.

Temporal Feature Definitions

Temporal Feature 1. The number of increased web pages from timestamp t_i to t_j, $|\varphi(t_i,t_j)|$. The elements in $\varphi(t_i,t_j)$ do not appear from the starting timestamp t_s to t_i, that is, $\forall d_n \in \varphi(t_i,t_j) \rightarrow d_n \notin \varphi(t_s,t_i)$.

Temporal Feature 2. The number of increased keywords from timestamp t_i to t_j, $|\psi(t_i,t_j)|$. The elements in $\psi(t_i,t_j)$ do not appear from the starting timestamp t_s to t_i, that is, $\forall k_m \in \psi(t_i,t_j) \rightarrow k_m \notin \psi(t_i,t_j)$.

Temporal Feature Definitions

Temporal Feature 3. The distribution of keywords on web pages from timestamp t_i to t_j, $\zeta(t_i, t_j)$. For an emergency event e, the web pages in $\varphi(t_i, t_j)$ can be represented as a vector by the keywords in $\psi(t_i, t_j)$. These vectors can be stored as a matrix:

$$\zeta(t_i, t_j) = \begin{pmatrix} w_{11} & \cdots & w_{1m} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nm} \end{pmatrix}. \qquad (2)$$

Temporal Feature Definitions

Temporal Feature 4. The associated relationships between keywords from timestamp t_i to t_j, $\Gamma(t_i,t_j)$. For an emergency event *e*, the associated relationships of keywords can be stored as a matrix:

$$\Gamma(t_i,t_j)=\begin{pmatrix} f_{11} & \cdots & f_{1m} \\ \vdots & \ddots & \vdots \\ f_{m1} & \cdots & f_{mm} \end{pmatrix}. \tag{3}$$

where f_{ij} means the weight of relation between k_i and k_j, which can be computed by

$$f_{ij}=\frac{\log\left(\dfrac{N(k_i \wedge k_j)*n}{N(k_i)*N(k_j)}\right)}{\log n} \tag{4}$$

where $N(k_i)$ means the number of web pages in $\varphi(t_i,t_j)$ containing k_i; $N(k_i \wedge k_j)$ is the number of web pages in $\varphi(t_i,t_j)$ containing both k_i and k_j.

Temporal Feature Definitions

Temporal Feature 5. The similarities between web pages from timestamp t_i to t_j, $\Xi(t_i,t_j)$. For an emergency event *e*, the similarities between web pages can be stored as a matrix:

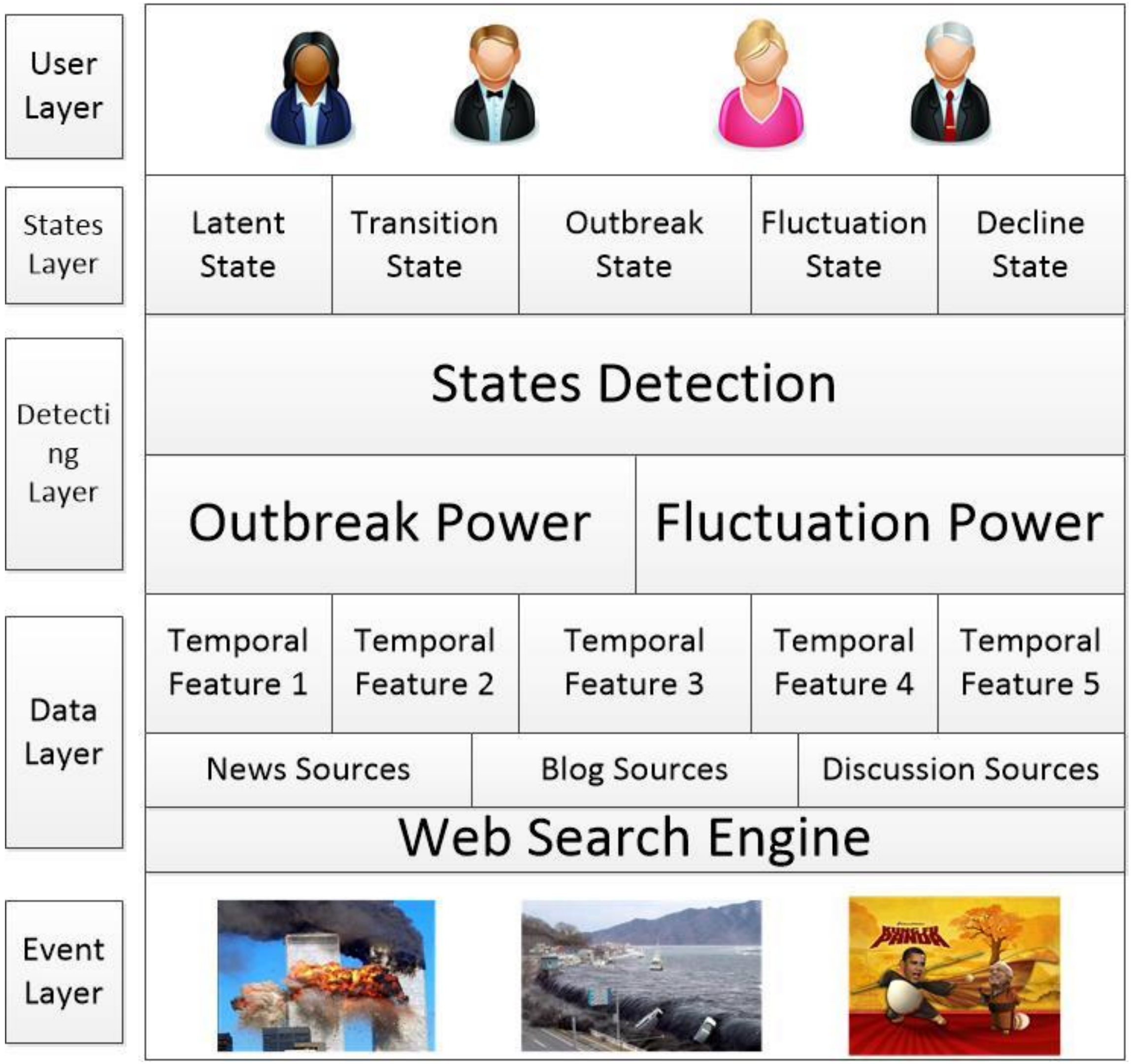
$$\Xi(t_i,t_j) = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}. \tag{5}$$

where a_{ij} means the similarities between d_i and d_j, which can be computed by

$$a_{ij} = \frac{d_i \cdot d_j}{\|d_i\|\|d_j\|}. \tag{6}$$

where $\|d_i\|$ and $\|d_j\|$ denote the mathematical model of vector d_i and d_j.

Proposed Algorithm



Variables and Parameters

Name	Description	Name	Description
emergency event	e	*The distribution of keywords*	$\zeta(t_i,t_j)$
life course of e	L_e	*The relations of keywords*	$\Gamma(t_i,t_j)$
basic features describing e	F_e	*The similarities between web pages*	$\Xi(t_i,t_j)$
seeds set	$S(t_i,t_j)$	*latent state*	LS_e
web pages set	$\varphi(t_i,t_j)$	*decline state*	DS_e
keywords set	$\psi(t_i,t_j)$	*outbreak state*	OS_e
The number of increased web pages	$\lvert\varphi(t_i,t_j)\rvert$	*transition state*	TS_e
The number of increased keywords	$\lvert\psi(t_i,t_j)\rvert$	*fluctuation state*	FS_e
outbreak power	$op(t_i,t_j)$	*representative power of keyword*	$rp(k)$
fluctuation power	$fp(t_i,t_j)$	*confidence of web page*	$cw(d)$

States Detection Algorithm

- Based on the five temporal features, the proposed computation algorithm is divided into three steps:
- **Outbreak power computation**
  - Compute the outbreak power, which reflects the influence degree of an emergency event
- **Fluctuation power computation**
  - Compute the fluctuation power, which reflects the change rate of an emergency event
- **States detection**
  - Based on the outbreak power and fluctuation power, we detect the different states of an emergency event

Computing Outbreak Power

- Degree of influence to the society

Algorithm 1: Computing Outbreak Power

Input: The set of web pages $\varphi(t_i,t_j)$ from time interval (t_i,t_j), the set of keywords on web pages $\varsigma(t_i,t_j)$, the distribution of keywords on web pages $\zeta(t_i,t_j)$

Output: The outbreak power $op(t_i,t_j)$

for each $d_h \in \varphi(t_i,t_j)$ **repeat** // *set the confidence of each web page as an initial state*

$cw(d_h) = \alpha$

for each $\sigma \in \varsigma(t_i,t_j)$ **repeat** // *compute the representative power*

$rp(\sigma) = rp(\sigma) * (1 - cw(d_h))$

$rp(\sigma) = 1 - rp(\sigma)$

for each $\lambda \in \zeta(t_i,t_j)$ **repeat** // *compute the confidence*

$rp(\sigma) = rp(\sigma) + \lambda * rp(\lambda)$

$rp(\sigma) = 1/(1 + e^{-rp(\sigma)})$

for each $d_h \in \varphi(t_i,t_j)$ **repeat** // *iteration computing*

for each $\sigma \in \varsigma(t_i,t_j)$ **repeat**

$cw(d_h) = cw(d_h) + rp(\sigma)$

for each $d_h \in \varphi(t_i,t_j)$ **repeat**

$op(t_i,t_j) = op(t_i,t_j) + (1 - cw(d_h))$

Computing Fluctuation Power

- Change rate of web pages

Algorithm 2: Computing Fluctuation Power

Input: The set of web pages $\varphi(t_{i-1}, t_i)$ from time interval (t_{i-1}, t_i), The set of web pages $\varphi(t_i, t_{i+1})$ from time interval (t_i, t_{i+1})

Output: The fluctuation power $fp(t_i, t_{i+1})$.

```
for each ω ∈ φ(t_i, t_{i+1}) repeat
    for each σ ∈ φ(t_{i-1}, t_i) repeat
        Sim ( ω , σ ); //cosine similarity of two web pages;
    cr(ω) = max(sim(ω, σ)) ; //get maximum similarity;
    fp(t_i, t_{i+1}) = fp(t_i, t_{i+1}) + cr(ω) ;
```

State Detection

- Based on Threshold values

Algorithm 3: States Detection of Emergency Event

Input: The set of states segmentation result $S = \{s_1, s_2, \ldots, s_k\}$, the set of outbreak power $op(t_s, t_e)$ from the starting time t_s to the ending time t_e, the set of fluctuation power $fp(t_s, t_e)$ from the starting time t_s to the ending time t_e

Output: The states detection result of each state

for each $\omega \in op(t_s, t_e)$ **repeat** *//compute average op*

$aop(e) = aop(e) + \omega$

$aop(e) = aop(e) / |op(t_s, t_e)|$

for each $\sigma \in fp(t_s, t_e)$ **repeat** *//compute average fp*

$afp(e) = afp(e) + \sigma$

$afp(e) = afp(e) / |fp(t_s, t_e)|$

for each $\gamma \in S$ **repeat** *// states detection*

If($\gamma == \max(S)$) $\gamma \rightarrow Outbreak\ State$

If($op(\gamma) < aop(e)$ && $fp(\gamma) < afp(e)$), $\gamma \rightarrow lds$

If($op(\gamma) > aop(e)$ && $fp(\gamma) < afp(e)$), $\gamma \rightarrow cts$

If($fp(\gamma) > afp(e)$&& $t > t_e$), $\gamma \rightarrow cfs$

else $\gamma \rightarrow decline\ state$

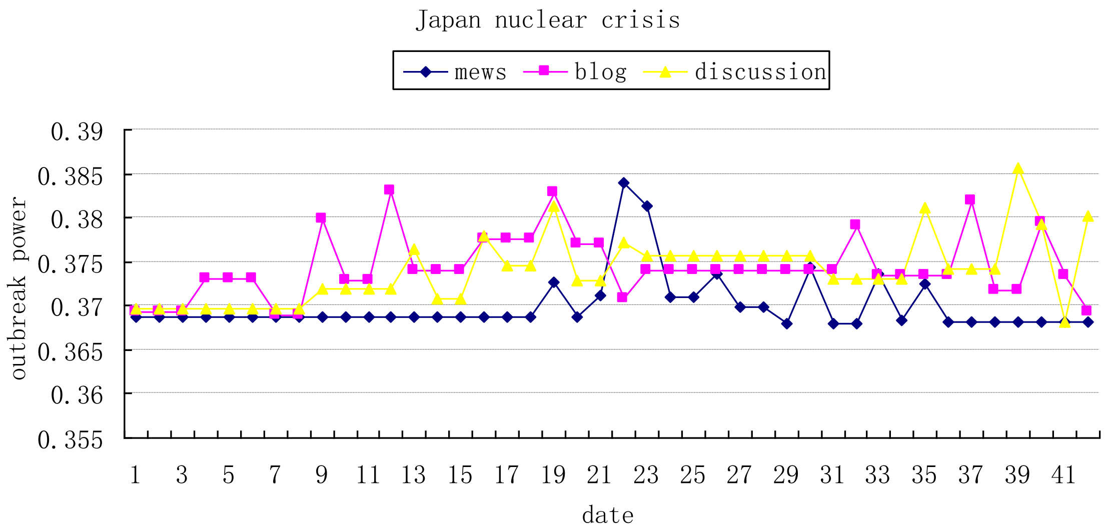
Experiments

- Data Sets
- The events in our experiments are extracted from the “Knowle system”
- Knowle is a news event central data management system
- The core elements of Knowle are news events on the web, which are linked by their semantic relations
- Knowle is a hierarchical data system, which has three different layers, namely: the bottom layer (concepts), the middle layer (resources), and the top layer (events)
- We select 50 events with about 450,000 web pages in our experiments from Knowle system, including political events, accident events, disaster events, and terrorism events
- Knowle provides the seed set, web pages, and keywords of events
- http://wkf.shu.edu.cn/

Initial Results

Table 3. The states detection results of the 50 emergency events with about 450,000 web pages

Event States	Latent state	Outbreak state	Decline state	Transition state	Fluctuation state	All states
Correct Detections	31	103	21	103	108	366
Error Detections	3	45	1	13	18	80
Detection Precision	0.911	0.696	0.955	0.888	0.857	0.821



The outbreak power of "Japan nuclear crisis" from different sources.

Observations

- **Observation 1**. The outbreak power of various information sources is different in most emergency events; i.e., the consistency of temporal feature of various information resources is low.
- **Observation 2**. The date of outbreak state from news source is mostly later than that of blog and bbs information sources.
- **Observation 3**. The outbreak power of blog and bbs information sources is mostly higher after the appearance of the outbreak state compared to that of news sources.
- **Observation 4**. The geographic distribution of social sensors may be related to the outbreak power of an emergency event.

Summary

- All countries, communities, and people are vulnerable to emergency events (e.g. terrorist attacks and natural disasters such as bush fire)
- Most emergency events are reported in the form of web resources (e.g. twitter and other social media feeds)
- Need to quickly process the information related to events
- Developing an approach to detect the different states of emergency events
- Related resources including web pages, keywords of an emergency event are collected using web search engines
- Outbreak power and the fluctuation power of an emergency event at different timestamps are computed
- Based on the various temporal values, different states of an emergency event are inferred
- Future work
  - Further refinement of the algorithms and heuristics
  - Further experimentation
  - Other applications

Papers Published So Far…

- Xu, Z., Luo, X., Liu, Y., Hu, C., Mei, L., Yen, N., Choo, K. K. R., Sugumaran, V. “From Latency, through Outbreak, to Decline: Detecting the States of Emergency Events Using Web Media Big Data,” *IEEE Transactions on Big Data* (forthcoming).
- Xu, Z., Zhang, H., Sugumaran, V. Choo, K. K. R., Mei, L., Zhu, Y. “Participatory Sensing based Semantic and Spatial Analysis of Urban Emergency Events using Mobile Social Media,” *EURASIP Journal on Wireless Communications and Networking*, 2016:44, pp. 1 – 9.
- Xu, Z., Zhang, H., Hu, C., Mei, L., Xuan, J., Choo, K. K. R., Sugumaran, V., Zhu, Y. “Building Knowledge Base of Urban Emergency Events based on Crowdsourcing of Social Media,” *Concurrency and Computation: Practice and Experience*, Vol. 28, No. 15, 2016, pp. 4038 – 4052.