

A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs

Qiaozhu Mei[†], Chao Liu[†], Hang Su[‡], and
ChengXiang Zhai[†]

[†]: University of Illinois at Urbana-Champaign

[‡]: Vanderbilt University

Weblog as an emerging new data...



hurricane katrina

[Advanced Blog Search](#)
[Preferences](#)

Search Blogs

Search the Web

Blog Search

Results 1-10 of about 1,093,993 for 'hurricane katrina' (0.72 seconds)

Sorted by relevance [Sort by date](#)

Related Blogs: [Hurricane Katrina Relief](#) - Knox Students in New Orleans

[Read President Bush's Commencement Address To MGCCC Graduates](#)

12 hours ago by Margaret Saizan

This is my 10th visit to Mississippi since **Hurricane Katrina** hit. I've seen firsthand the devastation in Gulfport and Gautier, Poplarville and Pascagoula, and Pass Christian, Bay Saint Louis and Biloxi. This was the first city in your ...

[Hurricane Katrina - http://www.hurricane-katrina.org/](http://www.hurricane-katrina.org/)

[[More results from Hurricane Katrina](#)]



An Example of Weblog Article

Location Info.

General	
Nickname	JenMahan
Gender	Female
Age	27
Occupation	Chief Cook and Bottle Washer
Location	North Carolina
Interests	Bloggng reading cake decorating
More about me	First of all I am Colleen & Wesley's sive. If you have this in place, you will be so much better off in the aftermath of a disaster. Much more likely to survive, and sustain less friend. I love my family and God. I

Social	
--------	--

Home | Profile | Blog | Photos | Lists

Blog

Previous entry: A new use for ... Next entry: More quizzes! ...

August 30

Talking about Hundreds of deaths feared from Katrina - Hurricane Katrina - MSNBC.com

Oh my God. I feel so terrible for these people. The entire city, under water. The life lost, the history and the beautiful landmarks, washed away. If you would like to ensure that your family is prepared as possible if something like this were to hit your town, please follow the guidelines in the links state your family's personal Disaster Recovery Plan. I was a planner by profession before I stayed at home with my children, I would be more to help anyone create their plan, I will answer questions and give advice - just leave me a comment, and your email address if you want me I back directly.

ask as long as you think it might to make your plan. Please start on it today. It is something you will not have time to think straight once you are in the throes of the situation. I encourage you to take the time to talk to those affected by the tragedy of Hurricane Katrina and earnest encouragement.

for just getting started - manageable steps. http://www.flylady.net/pages/FLYingLessons_Prepared.asp

Hundreds of deaths feared from Katrina - Hurricane Katrina - MSNBC.com

Skip

Add a comment | Read comments (3)

4:03 PM | Permalink | Trackbacks (0) | Blog it | Letters To The Editor

Permalink

<http://spaces.msn.com/auntiejen21/blog/cns!45D2FED4032E>

Comments

There are no words...but in NO...what they're saying is that things could have been worse...

Published By JacquelynnM (<http://spaces.msn.com/members/JasminesMomma/>) - August 31 8:38 AM

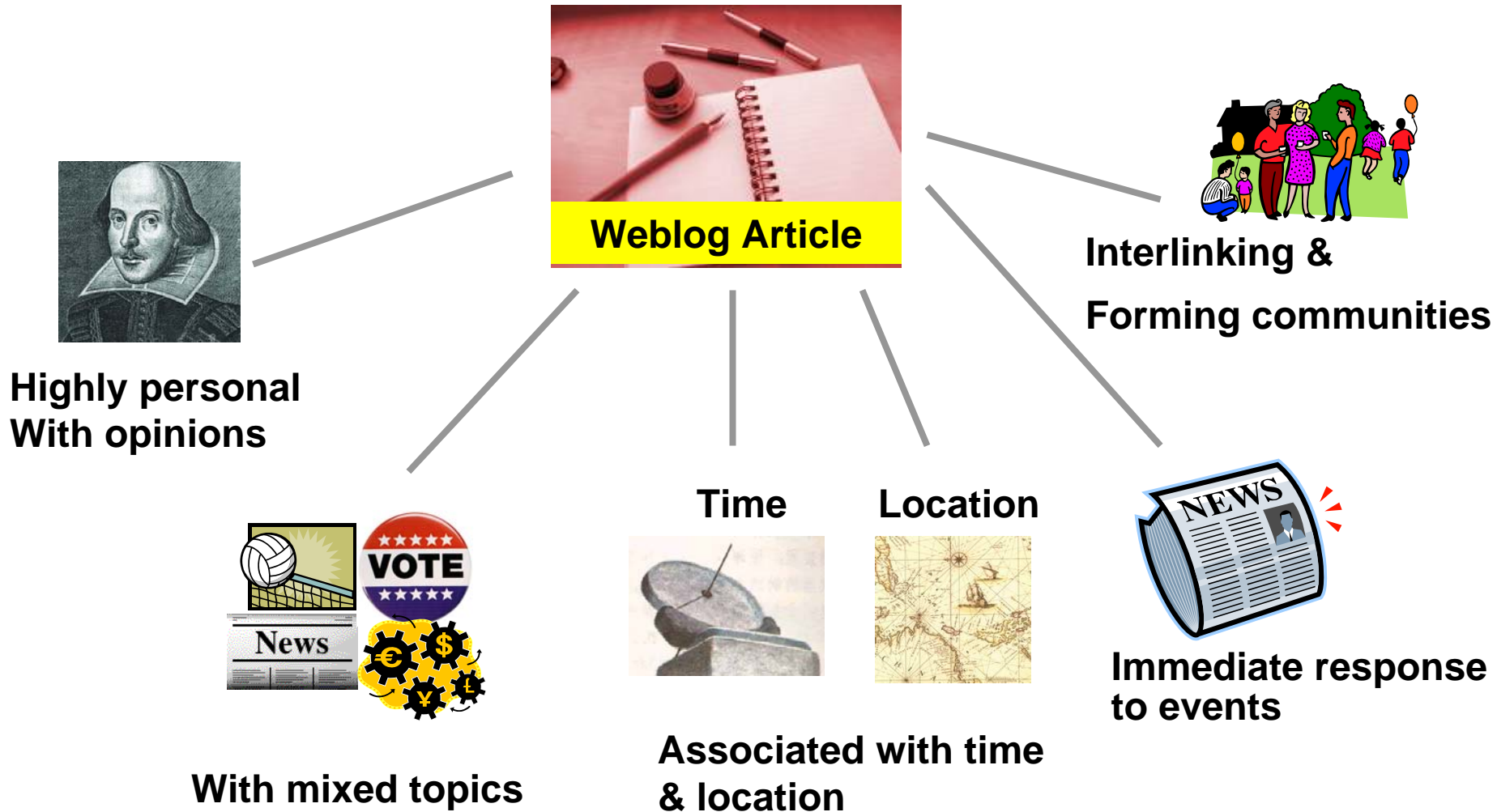
Hurricanes are no fun for anyone! I was glued to the tv through much of Katrina. I feel terrible for the people that lost thier homes and loved ones. Its almost a surreal feeling to watch footage from Katrina. What part of NC are you from Jen? I used to live in Jacksonville for 2 years and then Havelock for another 6 years. My ex-husband's were both Marines. I went through hurricanes Bertha, Fran, Bonnie, and Floyd and a few smaller ones and tropical storms. I can't imagine what it was like for the people in the gulf. Anyways Jen, I added you on my favorite blog list. I like your blog alot! Thanks for sharing it with me!

Kristi

Blog Contents

The time stamp

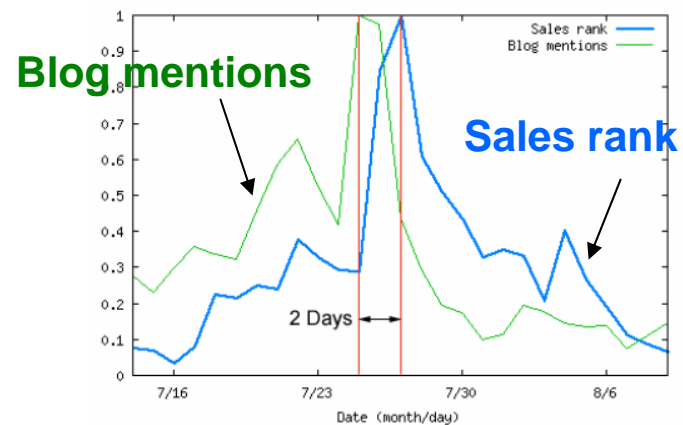
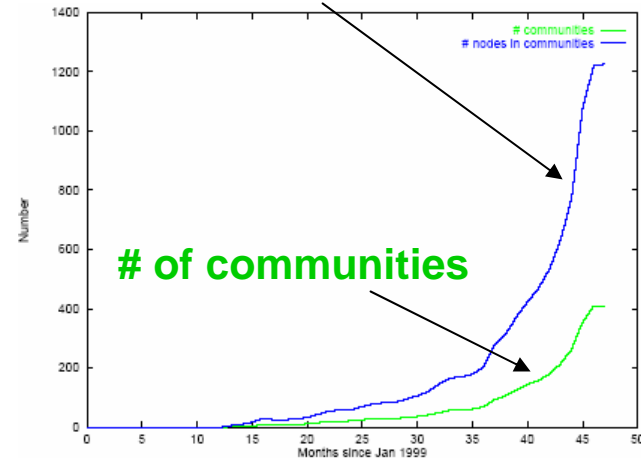
Characteristics of Weblogs



Existing Work on Weblog Analysis

- Interlinking and Community Analysis
 - Identifying communities
 - Monitoring the evolution and bursting of communities
 - E.g., [Kumar et al. 2003]
- Content Analysis
 - Blog level topic analysis
 - Information diffusion through blogspace
 - Use topic bursting to predict sales spikes
 - E.g., [Gruhl et al. 2005]

of nodes in communities



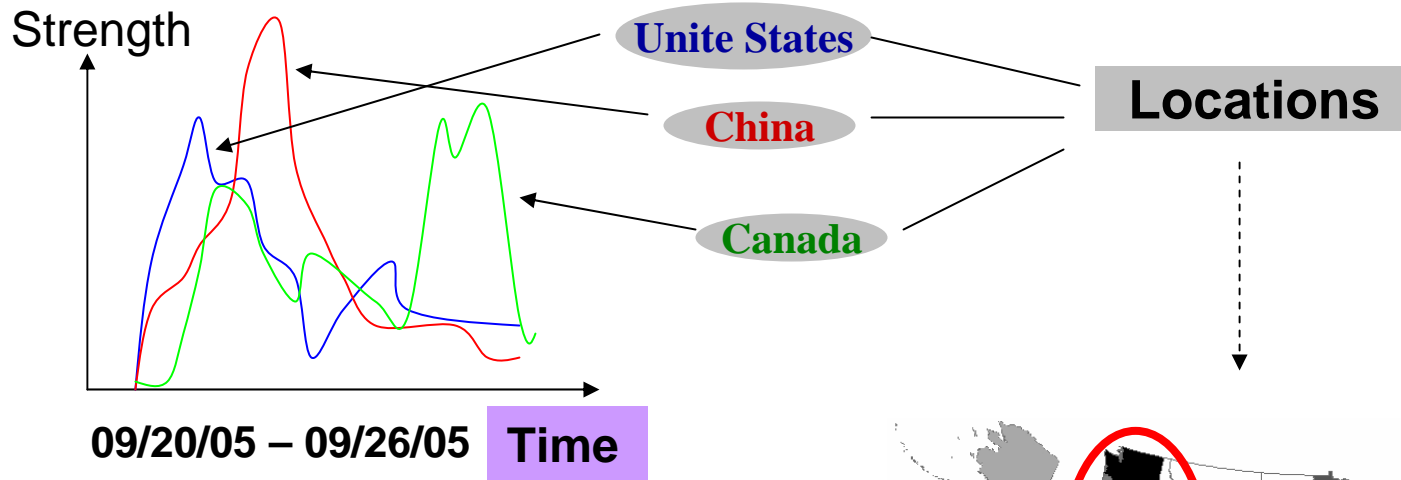
How to Perform Spatiotemporal Theme Mining?

- Given a collection of Weblog articles about a topic with time and location information
 - Discover multiple themes (i.e., subtopics) being discussed in these articles
 - For a given location, discover how each theme evolves over time (generate a theme life cycle)
 - For a given time, reveal how each theme spreads over locations (generate a theme snapshot)
 - Compare theme life cycles in different locations
 - Compare theme snapshots in different time periods
 - ...

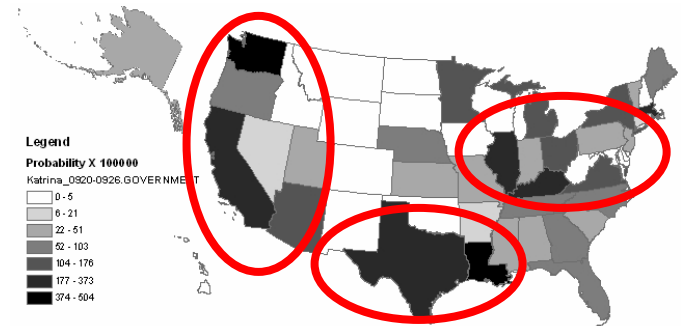
Spatiotemporal Theme Patterns

Theme life cycles

Discussion about “Release of iPod Nano”
in articles about “iPod Nano”



Discussion about “Government Response” in
articles about Hurricane Katrina



A theme snapshot

Applications of Spatiotemporal Theme Mining

- Help answer questions like
 - Which country responded first to the release of iPod Nano? China, UK, or Canada?
 - Do people in different states (e.g., Illinois vs. Texas) respond differently/similarly to the increase of gas price during Hurricane Katrina?
- Potentially useful for
 - Summarizing search results
 - Monitoring public opinions
 - Business Intelligence
 - ...

Challenges in Spatiotemporal Theme Mining

- How to represent a theme?
- How to model the themes in a collection?
- How to model their dependency on time and location?
- How to compute the theme life cycles and theme snapshots?
- All these must be done in an unsupervised way...

Our Solution: Use a Probabilistic Spatiotemporal Theme Model

- Each theme is represented as a multinomial distribution over the vocabulary (language model)
- Consider the collection as a sample from a mixture of these theme models
- Fit the model to the data and estimate the parameters
- Spatiotemporal theme patterns can then be computed from the estimated model parameters

Probabilistic Spatiotemporal Theme Model

Choose a theme θ_i

Draw a word from θ_i

Theme θ_1 price 0.3
oil 0.2..

oil

Theme θ_2 donate 0.1
relief 0.05
help 0.02 ..

donate

...

...

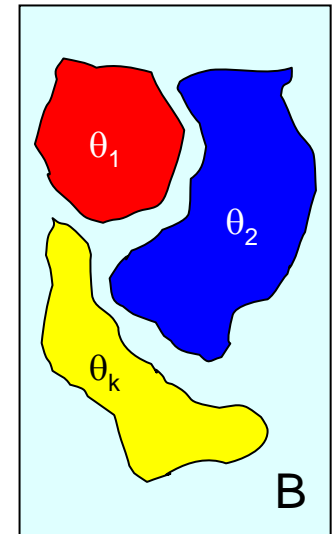
Theme θ_k city 0.2
new 0.1
orleans 0.05 ..

city

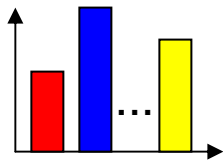
Background B Is 0.05
the 0.04
a 0.03 ..

the

Time = t
Location = l



Document d



Probability of choosing theme $\theta_i = \lambda_{TL} P(\theta_i | t, l) + \lambda_{TL} P(\theta_i | d)$

λ_{TL} = weight on spatiotemporal theme distribution

The “Generation” Process

- A document d of location l and time t is generated, word by word, as follows
 - First, decide whether to use the background theme θ_B
 - With probability λ_B , we’ll use the background theme and draw a word w from $p(w/\theta_B)$
 - If the background theme is not to be used, we’ll decide how to choose a topic theme
 - With probability λ_{TL} , we’ll sample a theme using the “shared spatiotemporal distribution” $p(\theta/t, l)$
 - With probability $1 - \lambda_{TL}$, we’ll sample a theme using $p(\theta/d)$
 - Draw a word w from the selected theme distribution $p(w/\theta_i)$
- Parameters
 - $\{p(w/\theta_B), p(w/\theta_i), p(\theta/t, l), p(\theta/d)\}$ (will be estimated)
 - $\lambda_B = \text{Background noise}$; $\lambda_{TL} = \text{Weight on spatiotemporal modeling}$ (will be manually set)

The Likelihood Function

$$\log p(C) = \sum_{d \in C} \sum_{w \in V} \frac{c(w, d)}{\sum_{w \in V} c(w, d)} \times \log \left[\lambda_B P(w | B) + (1 - \lambda_B) \sum_{j=1}^k p(w | \theta_j) \left((1 - \lambda_{TL}) p(\theta_j | d) + \lambda_{TL} p(\theta_j | t_d, l_d) \right) \right]$$

Count of word w
in document d

Generating w
using the background theme

Generating w
using a topic theme

Choosing a topic theme
according to the document

Choosing a topic theme
according to the
spatiotemporal context

Parameter Estimation

- Use the maximum likelihood estimator
- Use the Expectation-Maximization (EM) algorithm
- $p(w/\theta_B)$ is set to the collection word probability

E Step

$$\left\{ \begin{aligned} p(z_{d,w} = j) &= \frac{(1 - \lambda_B) p^{(m)}(w | \theta_j) [(1 - \lambda_{TL}) p^{(m)}(\theta_j | d) + \lambda_{TL} p^{(m)}(\theta_j | t_d, l_d)]}{\lambda_B p(w | B) + (1 - \lambda_B) \sum_{j'=1}^k p^{(m)}(w | \theta_{j'}) [(1 - \lambda_{TL}) p^{(m)}(\theta_{j'} | d) + \lambda_{TL} p^{(m)}(\theta_{j'} | t_d, l_d)]} \\ p(y_{d,w,j} = 1) &= \frac{\lambda_{TL} p^{(m)}(\theta_j | t_d, l_d)}{(1 - \lambda_{TL}) p^{(m)}(\theta_j | d) + \lambda_{TL} p^{(m)}(\theta_j | t_d, l_d)} \end{aligned} \right.$$

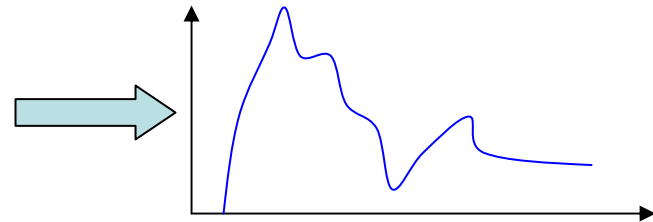
M Step

$$\left\{ \begin{aligned} p^{(m+1)}(\theta_j | d) &= \frac{\sum_{w \in V} c(w, d) p(z_{d,w} = j) (1 - p(y_{d,w,j} = 1))}{\sum_{j'=1}^k \sum_{w \in V} c(w, d) p(z_{d,w} = j') (1 - p(y_{d,w,j'} = 1))} \\ p^{(m+1)}(\theta_j | t, l) &= \frac{\sum_{d: t_d=t, l_d=l} \sum_{w \in V} c(w, d) p(z_{d,w} = j) p(y_{d,w,j} = 1)}{\sum_{d: t_d=t, l_d=l} \sum_{j'=1}^k \sum_{w \in V} c(w, d) p(z_{d,w} = j') p(y_{d,w,j'} = 1)} \\ p^{(m+1)}(w | \theta_j) &= \frac{\sum_{d \in C} c(w, d) p(z_{d,w} = j)}{\sum_{w' \in V} \sum_{d \in C} c(w', d) p(z_{d,w'} = j)} \end{aligned} \right.$$

Probabilistic Analysis of Spatiotemporal Themes

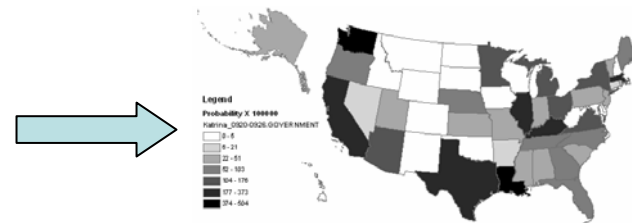
- Once the parameters are estimated, we can easily perform probabilistic analysis of spatiotemporal themes
 - Computing theme life cycles given location

$$p(t | \theta_j, \tilde{l}) = \frac{p(\theta_j | t, \tilde{l}) p(t, \tilde{l})}{\sum_{\tilde{t} \in T} p(\theta_j | \tilde{t}, \tilde{l}) p(\tilde{t}, \tilde{l})}$$



- Computing theme snapshots given time

$$p(\theta_j, l | \tilde{t}) = \frac{p(\theta_j | \tilde{t}, l) p(\tilde{t}, l)}{\sum_{\tilde{l} \in L} \sum_{j'=1}^k p(\theta_{j'} | \tilde{t}, \tilde{l}) p(\tilde{t}, \tilde{l})}$$



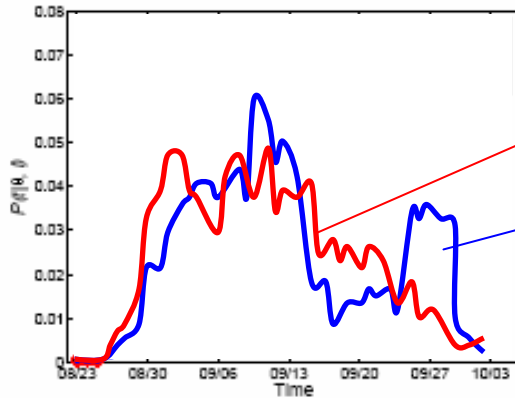
Experiments and Results

- Three time-stamped data sets of weblogs, each about one event (broad topic):

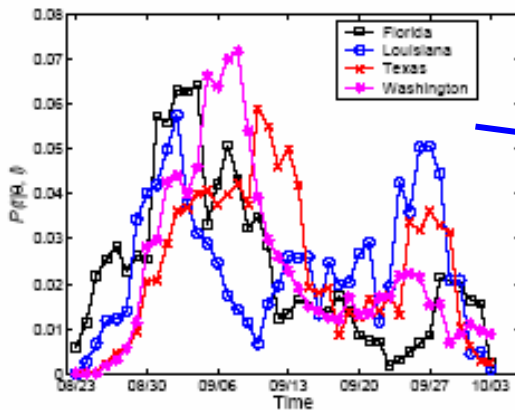
Data Set	# docs	Time Span(2005)	Query
Katrina	9377	08/16 -10/04	Hurricane Katrina
Rita	1754	08/16 - 10/04	Hurricane Rita
iPod Nano	1720	09/02 - 10/26	iPod Nano

- Extract location information from author profiles
- On each data set, we extract a set of salient themes and their life cycles / theme snapshots

Theme Life Cycles for Hurricane Katrina



(a) Theme life cycles in Texas
(Hurricane Katrina)



(b) Theme "New Orleans" over states
(Hurricane Katrina)

Oil Price

New Orleans

price 0.0772
oil 0.0643
gas 0.0454
increase 0.0210
product 0.0203
fuel 0.0188
company 0.0182
 ...

city 0.0634
orleans 0.0541
new 0.0342
louisiana 0.0235
flood 0.0227
evacuate 0.0211
storm 0.0177
 ...

Theme Snapshots for Hurricane Katrina

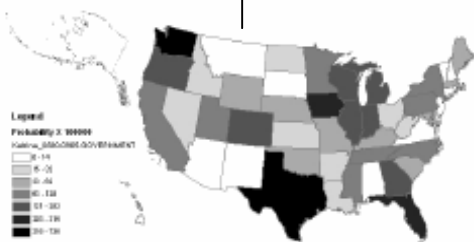
Week2: The discussion moves towards the north and west

Week1: The theme is the strongest along the Gulf of Mexico

Week3: The theme distributes more uniformly over the states



(a) Week1: 08/23-08/29



(b) Week Two: 08/30-09/05



(c) Week Three: 09/06-09/12

Theme 1	
Government Response	
bush	0.0716374
president	0.0610942
federal	0.0514114
govern	0.0476977
fema	0.0474692
administrate	0.0233903
response	0.0208351
brown	0.0199573
blame	0.0170033
governor	0.0142153



(d) Week Four: 09/13-09/19

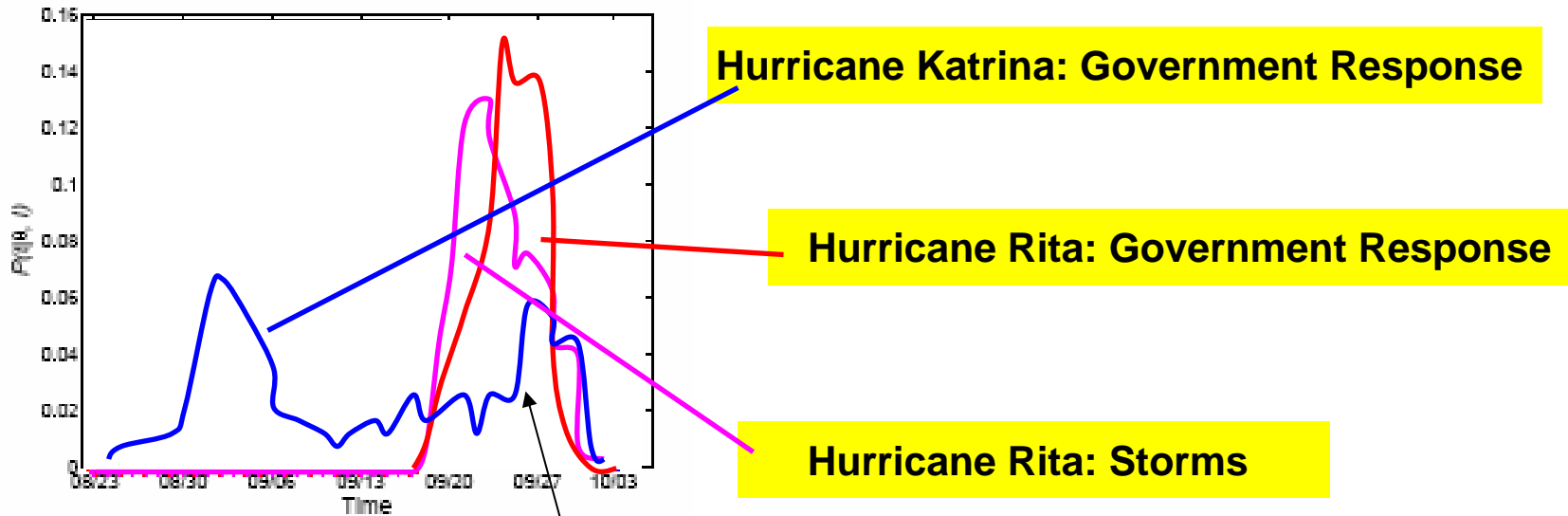


(e) Week Five: 09/20-09/26

Week4: The theme is again strong along the east coast and the Gulf of Mexico

Week5: The theme fades out in most states

Theme life cycles for Hurricane Rita



(d) Theme life cycles in Louisiana
(Hurricane Rita)

**A theme in Hurricane Katrina is inspired again by
Hurricane Rita**

Theme Snapshots for Hurricane Rita

Both Hurricane Katrina and Hurricane Rita have the theme “Oil Price”



(a) Week One of Rita: 09/17/-09/23



(b) Week Two of Rita: 09/24-09/30



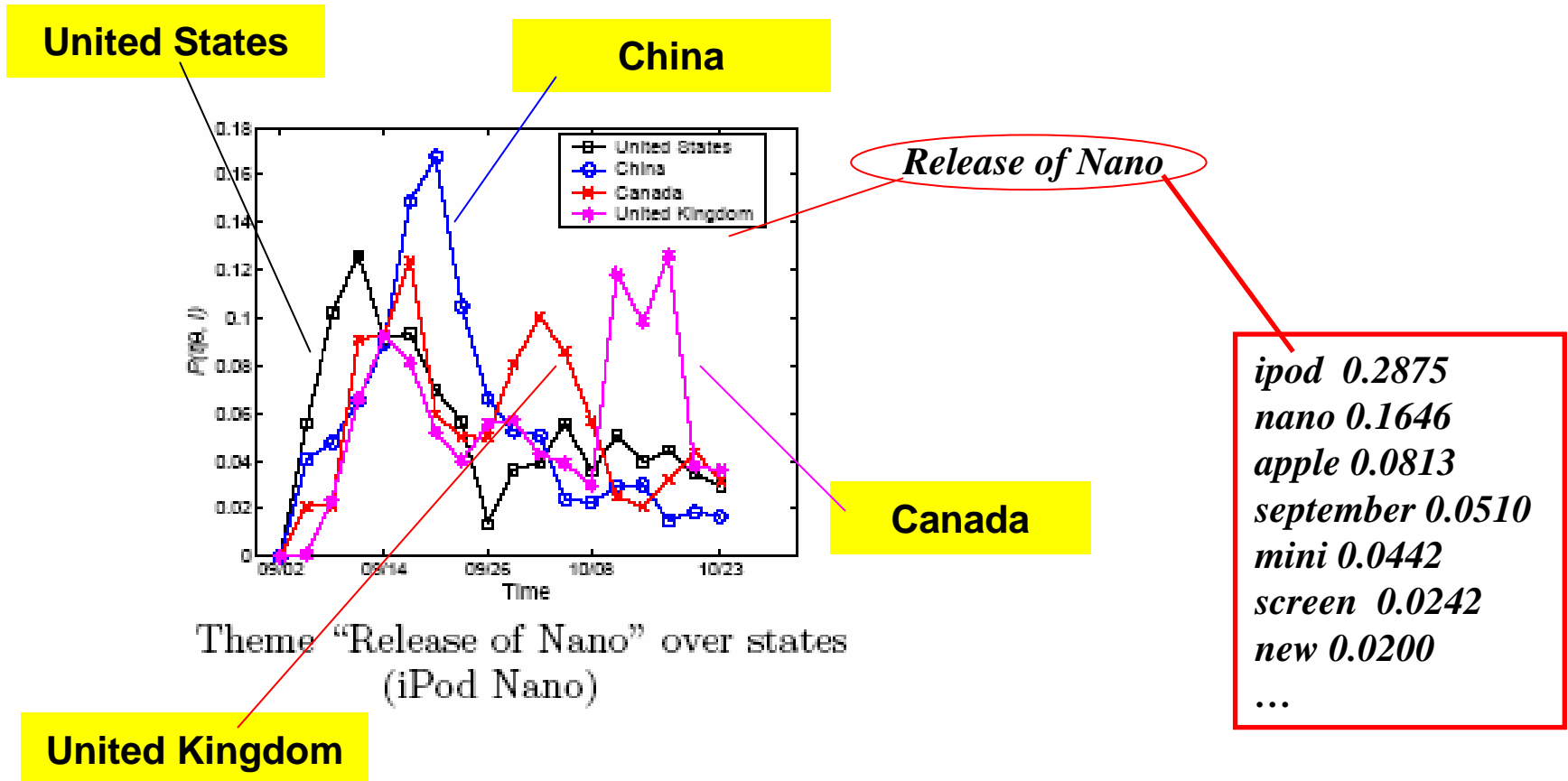
(c) Week Four of Katrina: 09/13-09/19



(d) Week Five of Katrina: 09/20-09/26

The spatiotemporal patterns of this theme at the same time period are similar

Theme Life Cycles for iPod Nano



Contributions and Future Work

- Contributions
 - Defined a new problem -- spatiotemporal text mining
 - Proposed a general mixture model for the mining task
 - Proposed methods for computing two spatiotemporal patterns -- theme life cycles and theme snapshots
 - Applied it to Weblog mining with interesting results
- Future work:
 - Capture content dependency between adjacent time stamps and locations
 - Study granularity selection in spatiotemporal text mining

Thank You!