# Interactive Visual Text Analytics for Decision Making

Shixia Liu Microsoft Research Asia

1

### **Text is Everywhere**

- We use documents as primary information artifact in our lives
- Our access to documents has grown tremendously in recent years due to networking infrastructure
  - -WWW

— ...

- Digital libraries

Is This You?

### **Big Question**

What can information visualization provide to help users in understanding and gathering information from text and document collections?

### Outline

### Example tasks in text analytics

- Visually analyzing textual information
  - Dynamic Word Cloud
  - Topic-based Visual Text Summarization
  - TextFlow: Towards Better Understanding of Evolving Topics in Text
- Text Visualization Perspectives

#### How can I find information buried inside the piles of text?

#### Terracotta Army - Wikipedia, the free encyclopedia [2] Mount Li is also where the material to make the terracotta warriors originated. In addition to the warriors, an entire man-made necropolis for the ... en.wikipedia.org/wiki/Terracotta Army - 56k - Cached - Similar pages Museum of Qin Terra Cotta Warriors and Horses The Terka Cotta Warriors and Horses are the most significant archeological excavations of the 20th century. It is a sight not to be missed by any visitor to ... www.travelchinaguide.com/attraction/shaanxi/xian/terra cotta army/ - 17k -Cached - Similar pages Terra Cotta Pit 1 Museum of Qin Terra Cotta Warriors and Horses : Pit 1 ... There are more than 6000 terracotta warriors and horses in Pit No. 1. marshaled into battle line ... www.travelchinaguide.com/cityguides/xian/terracotta.htm - 14k - Cached - Similar page Terracotta Warriors : The Museum Terracotta Warriors Museum, Dorchester, brings together all the wonder of the discovery of the many treasures of the first Emperor of China. www.terracottawarriors.co.uk/ - 14k - Cached - Similar pages Terracotta Warriors - A Fantastic Tourist Attraction in China ... Terracotta Warriors Tours: Private tours to Terracotta Warriors, and other Xi'an ... 1 which contains 6000 life-size terracotta warriors and horses. ... www.chinavista.com/travel/terracotta/warrior01.html - 6k - Cached - Similar pages Terracotta Warriors - A Fantastic Tourist Attraction in China ... Terracotta Warriors Tours: Private tours to Terracotta Warriors. ... Let us go to Xi'an te have a look at the Museum of Qin Terracotta Warriors. ... www.chinavista.com/travel/terracotta/main.html - 6k - Cached - Similar pages ta V viors - a b

#### Information finding

What is in my text?

What's inside the NHTSA Data:	What are the major causes of injuries	What did my customers say about my hotels
450,000+	70,000+ patient	3000+ customer-
documents	emergency room records	posted reviews

**Information Understanding: Text Summarization** 

What is in my text?

Which hotel features do my customers like/dislike	How customers' sentiment have changed toward my hotels	How do customers feel about my new product launch
3000+ customer reviews	3000+ customer- posted reviews	thousands of e- opinion postings

#### **Insight Discovery: Sentiment Analysis**

What is in my text?

What are the correlations of tire problems and highway death in the NHTSA Data:	What are the correlations of patient gender and the cause of injury	Compare the customers' attitude toward our product with theirs for our
	70,000+ patient	competitors
450,000+	emergency room	
documents	records	thousands of e-
		opinion postings

Decision Making and Problem Solving: Text Analysis++

### **Major Challenges**

- Huge amounts of complex information
  - Understanding the meanings of free text is just hard
  - Performing analysis on top of that is harder
- Different people want different things
  - No one-size-fits-all solutions
- People may not know what they want
  - "Tell me something I don't know"
  - "I will tell you when I see it"

#### alderwood

al ex 2 and 15 accession americal americal americal and an experimental and the second and the second americal american americal american americal american americ american american am
bruce chandler buchanan bueno to flass and to the states dependence bush subject of Contract and the states of the
Crum cruz d.j. search wate dan newhouse date published any dynamic and the sector and the sector decorate and the sector and t
ellie olmsen date published was entiatoppa wasarme aster wywar tostval of trees wasarme tast was task salcodo tankie pots tred killan treepons frisbee gay locke george bush george nethercutt genera grave giver giver giver aster and genn
geon tremen god genere grabenhorst gravegide green Way greg gilleland grizzlies geteens haak halloween nameen nameen
honeyford
john panni date published john torch provinger of honor provinger of the second s
bese guiterrez y tos mermanos ochoa paskan josoph villansova perkapopentalevou juni juliet K-12 koth monts Kerry kan Knightistady iam beverzeneae valateriter ver linda looke valanzeterid is universe tos temeroso lower valley universe en energia cala
Iower valley memorial gardens to option territoriate the register metabolic matching and the second se
matters of mariner
mark marro marro married martinez research characteristic research and the married marr
matson native relative modes measuring the sacker mel sacker date published native meyer native melange satisfying to cauto expedies monson montelong o modes moses
matson where notice notice that the sector of the sector of the published where meyer any sector of the sector of
mation where the second
matson endemonent and the sector met saccher date published meyer endemonent means montherapy on the sector and the sect
matison and reason reason reason and a source mel sacher date published - source metal on power moster moster on power moster
matison

onference council forum gathering and meeting party session visit visiting 1-column

accord activity add actition actitional actit after moon age agree air addenwood atterneod's approve area arts access a

ntempt attend back ball bates been an begin bill block board bob bones break brin an stow brother bse budget buen 0 build building bush business bry cabin call candidate can car career carry case cattle center central certs chance chandler change charge charter **child** shorese christian church City class classroom close clown club coach college mittee COMMUNITY complete com CONCEIN concert conference contact continue co copyright cost casts COUNCII country COUNty couple cousin credit cruz curiman cut dance date daughter day de al death debate decide defense defensive de gree department design die difficult disease **district** doctor dridnill drive drug early earn eas The particular states of the second states of the s fun fund funding funds funeral gain game gameral george gilleland girl give goal genzales good grade graduate grandchild grandview granger great grizzly ground group hair half halloween hard harris hawk he ad hear hernandez high highlan Told native NOTE honeyford honor hope hospital agent hegen blog husband idea impotant improve Include increase information interacture INIU aUVE interested involve ing immail SUE (princi JOE) Din Joney Kerry Vidi alge lamb wie laugh haughter interested league learn leave legistatorievei library life type Ine iwe local tong los isse toss for love low lower mabton matter's machine make man manager Marro many martinez matson mayor measure medical meeting mel MemDe memorial mexican mexica meyer middle military million minutes Monday Money month moming moses move murray music national nethercutt newhouse night note nov OCI offensive offer afficer afficer of ficial dimen Opening OpportUnity autoor paige par parentpart party **Pass** past pay **PCOPIE** percent perform performance periodpia pick pioneer place plan **Play** player playoff point police por posteasenoprac precede prepare present president press pretty problem program project pro promote property prosser provide public publics purchase put quarter quarterback raise rate read reading receive recently record red regions regular report representative resident result return revenue read robusco rockford routine rUID rush sacher sates sanches

#### Machines are \*not\* just smart enough.

### Outline

Example tasks in text analytics

### Visually analyzing textual information

- Dynamic Word Cloud
- Topic-based Visual Text Summarization
- TextFlow: Towards Better Understanding of Evolving Topics in Text
- Text Visualization Perspectives

### **Dynamic Word Cloud**

#### Word clouds for content overview

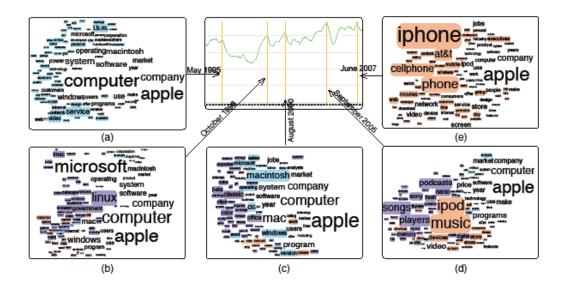
- Aesthetic issues
- Inadequate for temporal patterns

#### Standard time chart: trend

- Inadequate for correlations

### **Our Solution**

- A evolution trend chart + word clouds
  - Measure the evolution
  - Ensure the semantic coherence between clouds



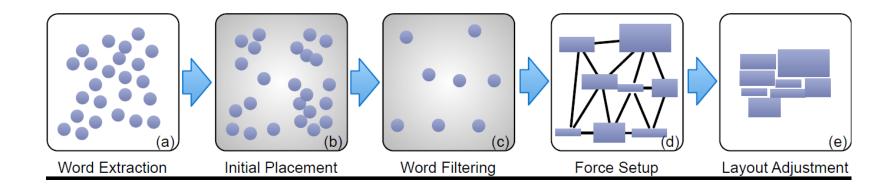
**Evolution Measurement** 

Conditional entropy: measure the amount of information contained by X<sub>i</sub> but not by X<sub>j</sub>

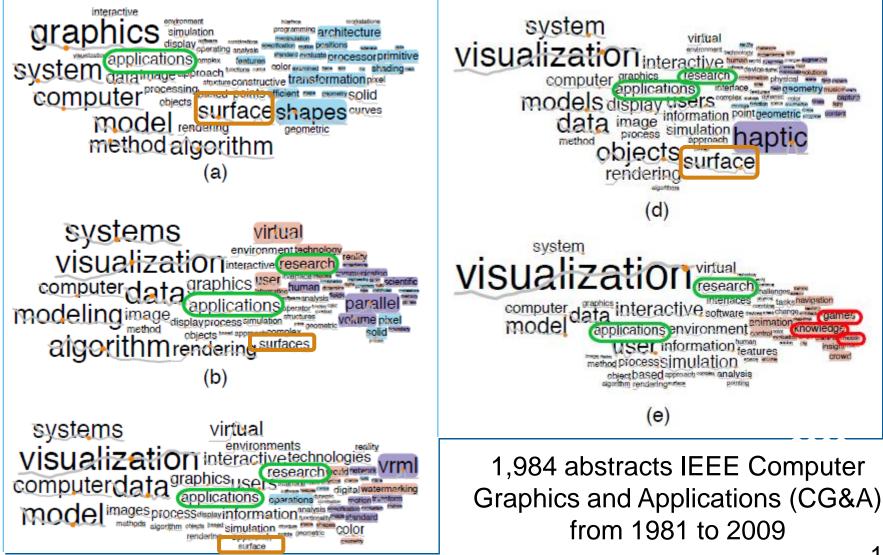
$$S(X_i) = \sum_{j=-w/2}^{w/2} t_j H(X_i | X_{i+j}) = \sum_{j=-w/2}^{w/2} t_j (H(X_i) - H(X_i; X_{i+j}))$$

### Word Cloud Layout

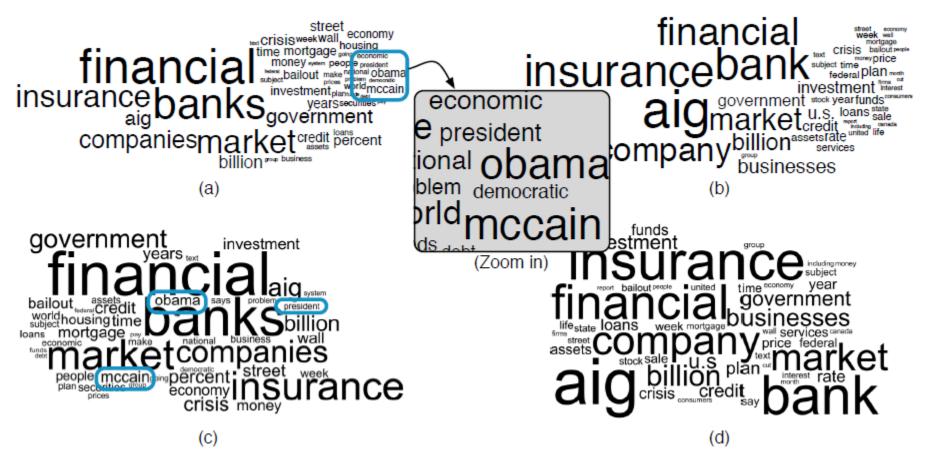
- Geometry meshes to ensure the semantic coherence
  - Semantically related words stay together
  - The same word in different clouds stay at the similar place



### Example: CG&A Abstracts



### **Comparison with Wordle**



13,828 news articles

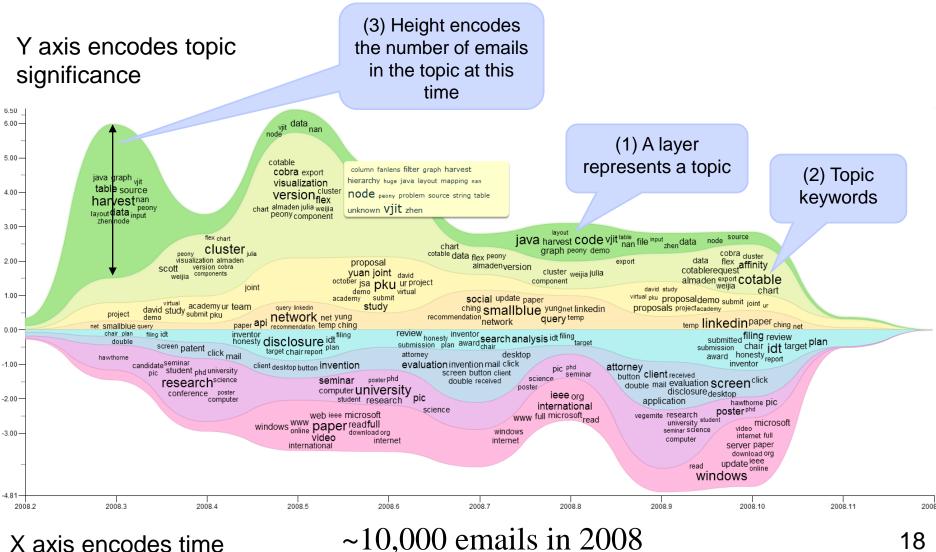
### Outline

Example tasks in text analytics

### Visually analyzing textual information

- Dynamic Word Cloud
- Topic-based Visual Text Summarization
- TextFlow: Towards Better Understanding of Evolving Topics in Text
- Text Visualization Perspectives

### Liu et al. CIKM 09 **Demo: Topic-based Visual Text Summarization**





# Interactive, Time-based Visual Email Summarization

Shixia Liu, Michelle X Zhou, Shimei Pan, Weihong Qian, Weijia Cai, Xiaoxiao Lian

**IBM Research** 

### **Visual Text Summarization: Key Challenges**

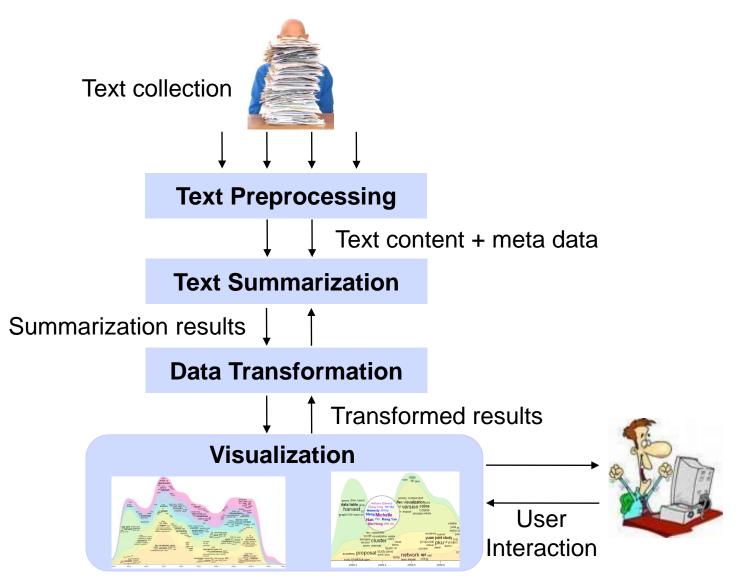
#### How to summarize massive, time-varying text corpora

- Huge amounts of complex information
- Accuracy + extensibility
- Time-varying
- Effectiveness

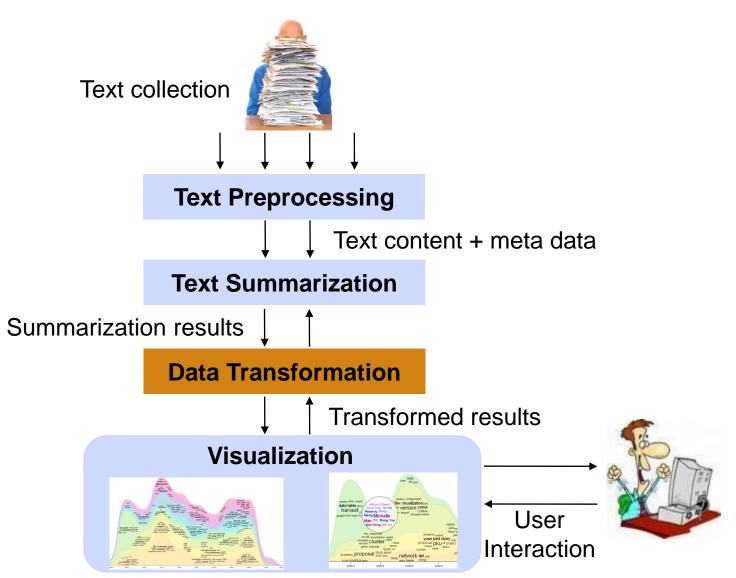
#### How to visually explain text summarization results

- Which visual metaphors to use
- How to switch between visualizations
- How to allow users to provide feedback or articulate their needs
  - Incorrect summarization results or varied user needs

### **TIARA** Overview



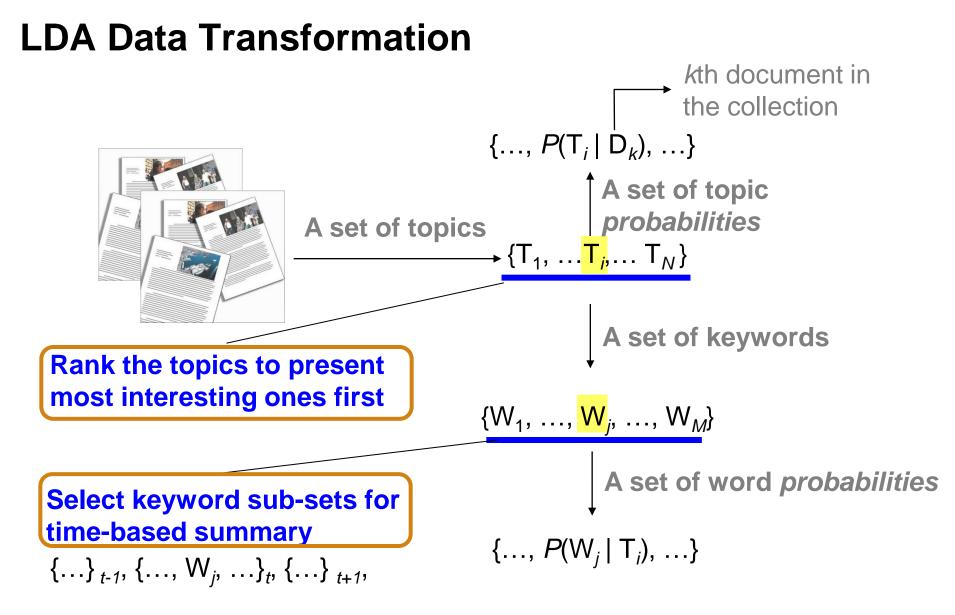
### **TIARA Technical Focus**



#### **Adopted Text Summarization: LDA-based Topic Analysis**

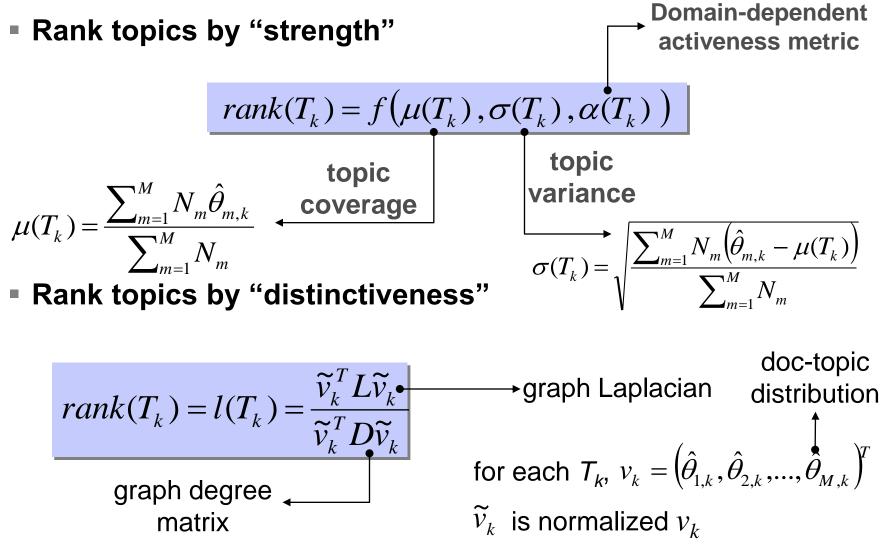
- Latent Dirichlet Allocation (LDA) model [Blei et al. 03]
  - Statistical analysis that requires no extra knowledge about the language or the world (high portability)
  - Keyword-based description of thematic structure of a text collection (high compaction rate for scalability)
  - A finer grained model than clustering
    - One document → multiple topics → finer-grained summarization

kth document in the collection  $\{\ldots, P(T_i | D_k), \ldots\}$ A set of topic probabilities  $\{\mathsf{T}_1,\ldots,\mathsf{T}_i,\ldots,\mathsf{T}_N\}$ A set of topic A set of topics **keywords**  $\{\mathsf{W}_1, \ldots, \mathsf{W}_i, \ldots, \mathsf{W}_M\}$ A set of word probabilities  $\{\ldots, P(W_i | T_j), \ldots\}$ 



#### [Song et al. CIKM 09]

### **Topic Ranking by User Interests**



### **Topic Ranking by User Interests: Experiments**

#### Goal

- Measure which metric produces more "important" topics

### Data sets

- Email
  - 8326 email messages
  - 958,069 word tokens
- News
  - 34,690 documents
  - 11,491,246 word tokens

#### Method

- Users indicate the importance of top-K ranked topics
  - Very important
  - Somewhat important
  - Unimportant

### **Topic Ranking by User Interests: Results**

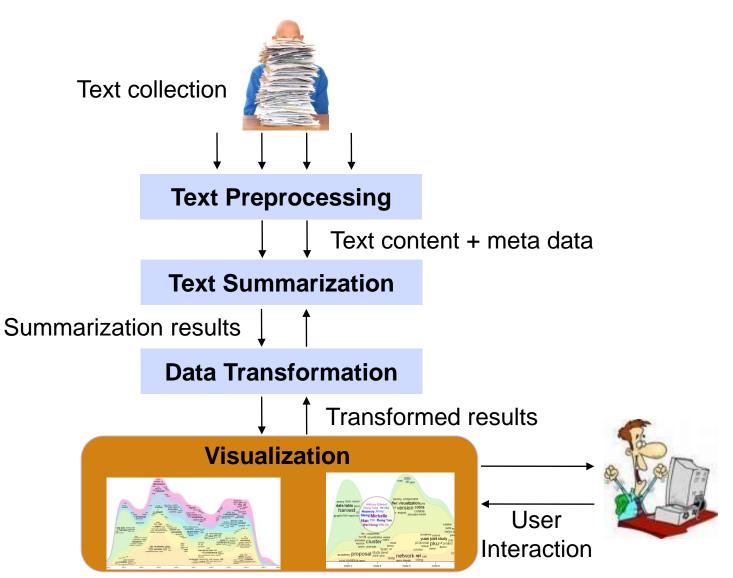
#### Email data (by F1 measure)

Retrieved	Top 5	Top 10
strength	0.760 ± 0.057	0.640 ± 0.035
distinctiveness	0.920 ± 0.069	0.900 ± 0.000
M.I.	0.960 ± 0.057	0.860 ± 0.035
T.S.	0.520 ± 0.056	0.560 ± 0.035

News data (by F1 measure)

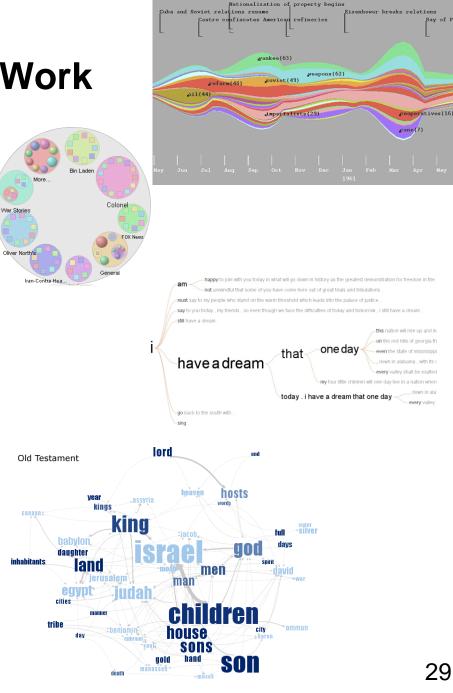
Retrieved	Top 5	Top 10
strength	0.640 ± 0.057	0.68 ± 0.028
distinctiveness	0.760 ± 0.057	0.76 ± 0.035
M.I.	0.760 ± 0.057	0.74 ± 0.035
T.S.	0.720 ± 0.069	0.70 ± 0.045

### **TIARA Technical Focus**



## **Visualizing Text: Existing Work**

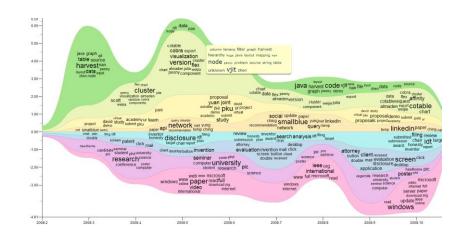
- Visualize text at a high level (e.g., documents)
  - -Issues
    - Inadequate information
- Visualize text at a low level (e.g., words and phrases)
  - -Issues
    - Missing a big picture
    - Scalability
- Few on explaining advanced text analysis results

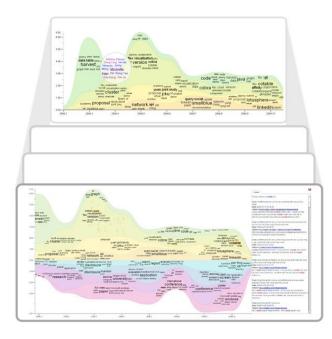


#### [Liu et al. CIKM 09]

### **Our Approach**

- Create visual metaphors to explain text summarization results
- Use interaction to let people
  - Examine data from multiple aspects
- Support both top-down and bottom-up analysis
  - Forest → tree → leaf
  - Leaf → tree → forest
- Design principle
  - Augment common visual metaphors to convey complex summarization results



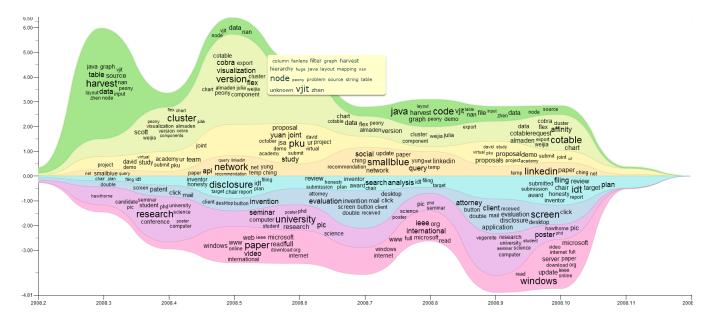


### **Visual Text Summary Metaphor**

#### Data to be visualized:

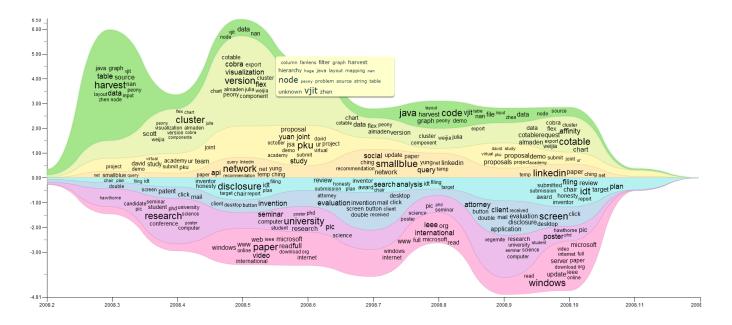
- 1. Topics: { $T_1$ , ...,  $T_i$ , ...,  $T_N$ } and their probabilities
- 2. For each  $T_i$ , Topic keywords by time: ... {..., $w_k^i$ , ..., }<sub>t</sub>, ... and their probabilities over time
- 3. For each  $T_i$ , Topic strength: {...,  $S^i(t)$ , ...} over time

Visual encoding: Augmented stacked graph



### **Enhanced Stacked Graph: Key Steps**

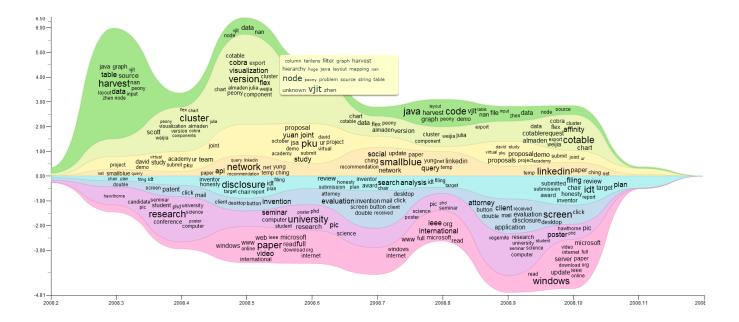
- 1. Computing geometry of layers
- 2. Layer coloring
- 3. Layer ordering
- 4. Layer labeling



### **Enhanced Stacked Graph: Key Steps**

- 1. Computing geometry of layers
- 2. Layer coloring
- 3. Layer ordering
- 4. Layer labeling

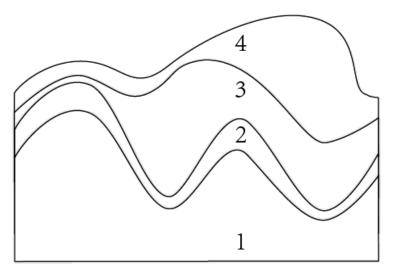
#### Byron\_Infovis08

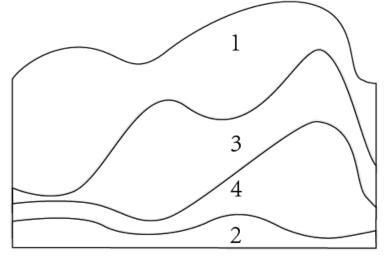


### **Enhanced Stacked Graph: Layer Ordering**

#### Goals

- Minimize distortion
- Maximize usable space
- Ensure semantic coherence



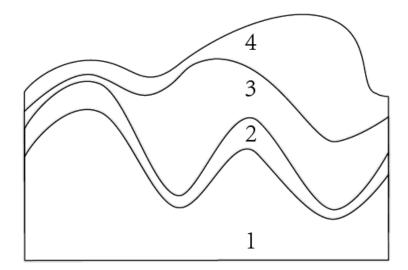


unordered

### Enhanced Stacked Graph: Layer Ordering (cont'd)

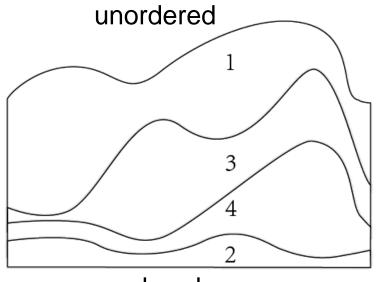
#### Our approach

- Optimization-based approach to
  - Minimize volatility (curvature) of topics
  - Maximize geometric
    complementariness of adjacent topics
  - Ensure semantic proximity of topics

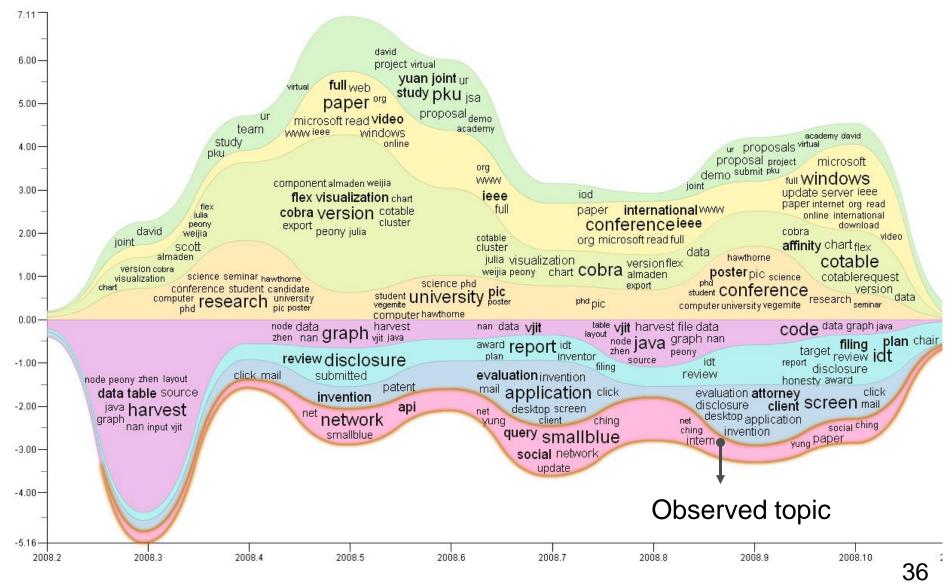




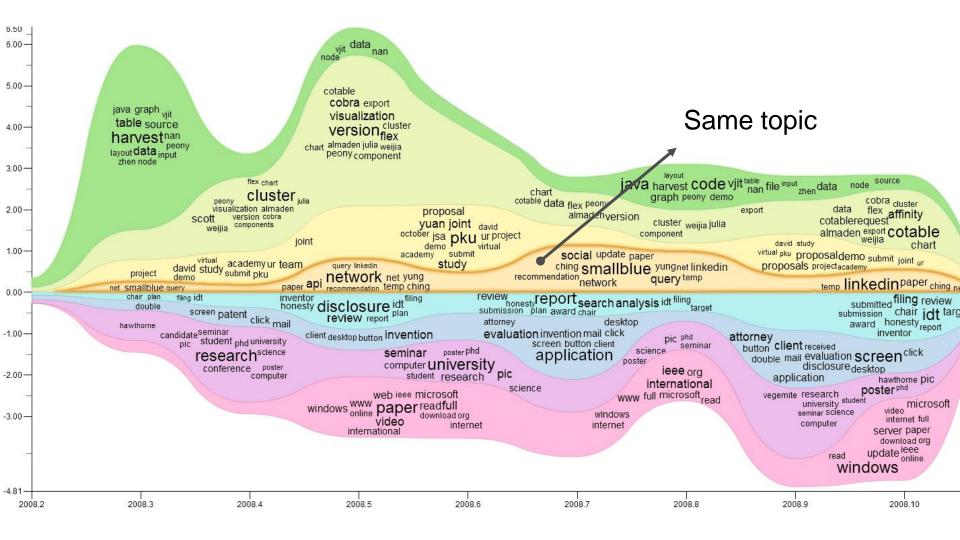
→ 
$$F_{\sigma}(p'_{ij}(t)); p'_{ij}(t) = p_i(t) + p_j(t)$$



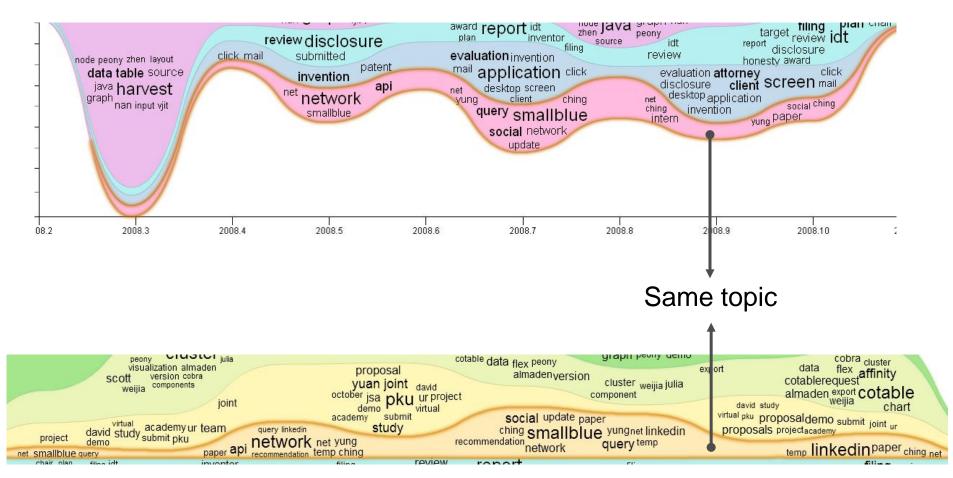
### Enhanced Stacked Graph: Layer Ordering (cont'd)



## Enhanced Stacked Graph: Layer Ordering (cont'd)



# Enhanced Stacked Graph: Layer Ordering (cont'd)



Alternative solution: Interactive reordering

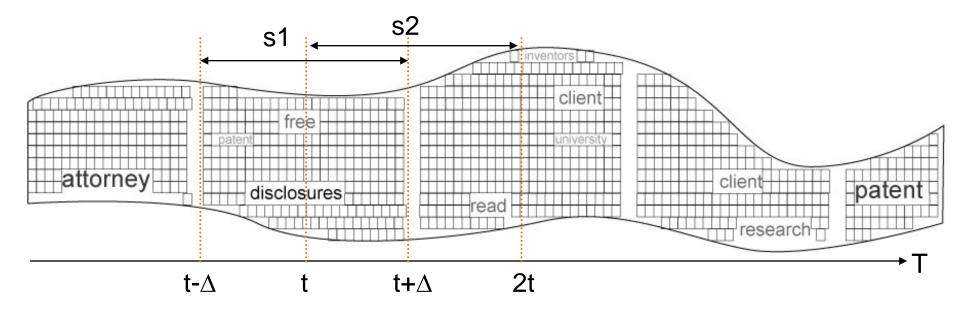
## **Enhanced Stacked Graph: Layer Labeling**

### Goals

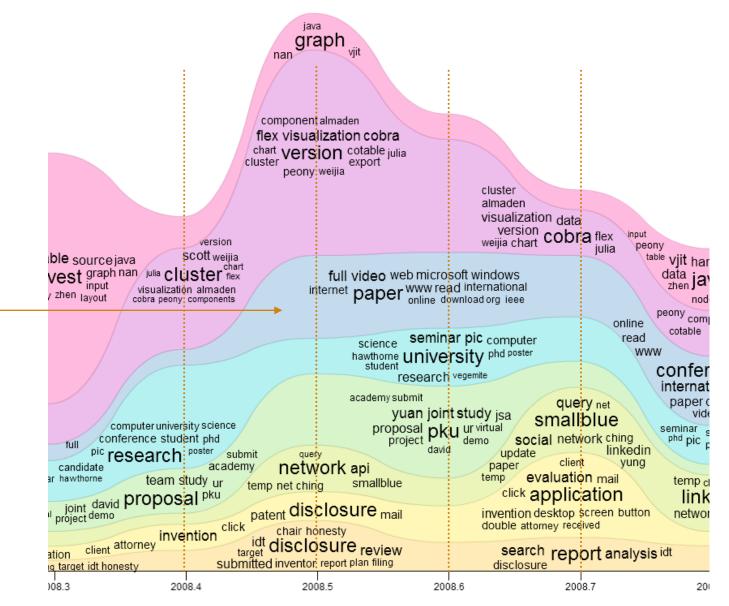
- Temporal proximity
- Informativeness

## Enhanced Stacked Graph: Layer Labeling (cont'd)

- Our approach [Liu et al. CIKM09]
  - Constraint-based space allocation
  - Particle-based layout [Luboschik et al. 08] + wordle

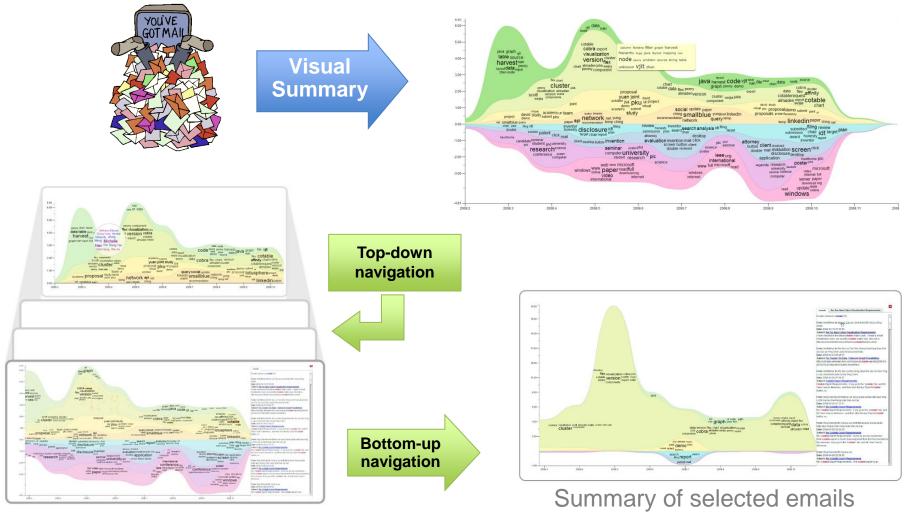


## Enhanced Stacked Graph: Layer Labeling (cont'd)



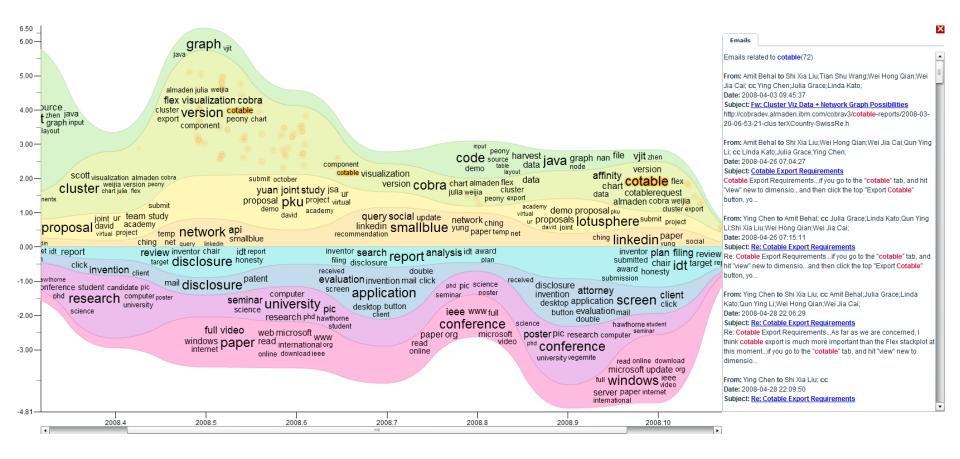
41

## **Top-down and Bottom-up Analysis**



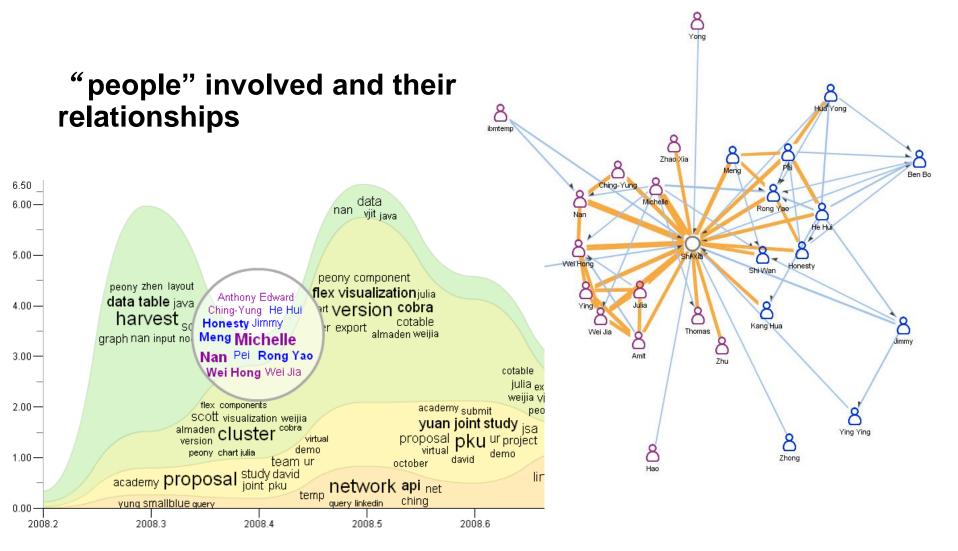
#### **Iterative analysis**

## **Interacting with Visual Summary**



#### Selected keyword "cotable" and relevant email snippets

## **Interacting with Visual Summary**



# **TIARA Preliminary Evaluation**

### Goal

- What kind of tasks can TIARA help users accomplish?
- What factors impact the use of TIARA?
- Data set
  - Emails (~10,000)
- 12 Participants: 6 familiar with the email owner's work
- Tasks
  - Task 1: Analyze emails between two people
  - Task 2: Answer questions about specific projects
  - Task 3: Answer questions characterizing the email owner's work

# **TIARA Preliminary Evaluation**

### Method

- Objective measures
  - answer completion rate, answer error rate, and answer time
- Subjective measures
  - usefulness, usability, and system satisfaction
- Compared TIARA and Themail for Task 1

## **Sample Evaluation Questions**

#### Task1 examining emails between two people

What are the three main concepts mentioned during their June emails?

Which month of 2008 is most active in their email exchanges?

#### Task2 examining emails about a project named "Cobra"

Who were involved in Cobra?

When was this project most active?

What was discussed during the active period?

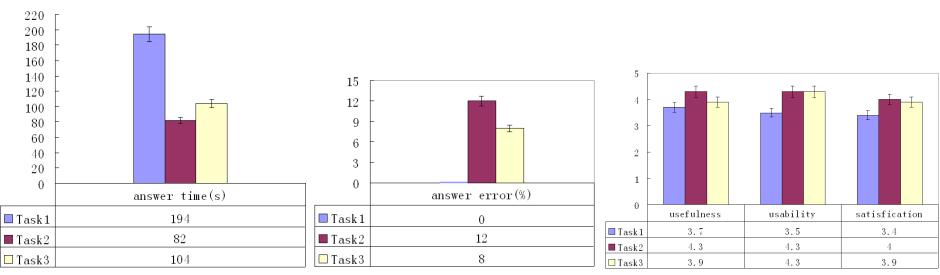
#### Task3 examining emails in general

What was the most active topic in May?

Who were the people involved in this topic?

## **TIARA Evaluation Results**

- Type of tasks impacted the effectiveness of TIARA
  - TIARA helped users complete complex analytic tasks faster
  - Visual summary too coarse for extracting details (e.g., names)
- User's knowledge impacted the effectiveness of TIARA

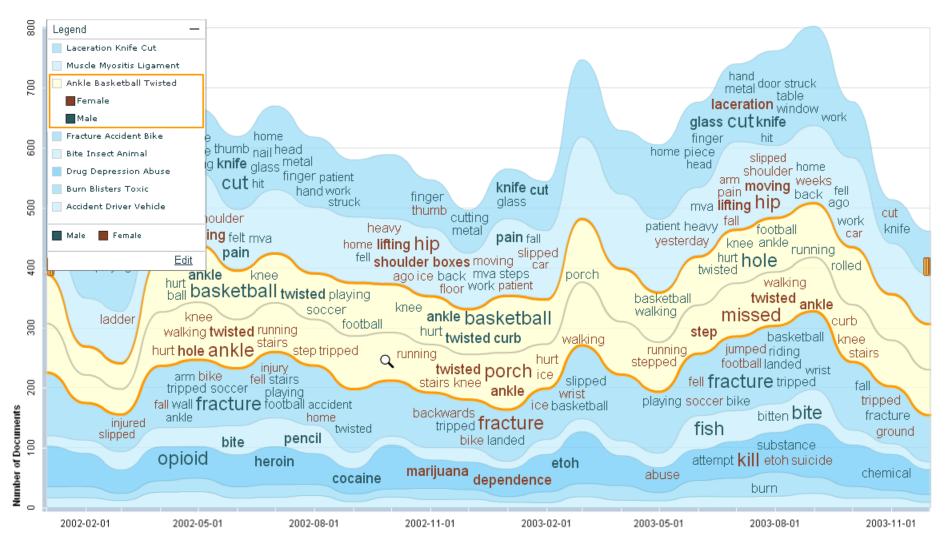


TIARA helped knowledgeable users more

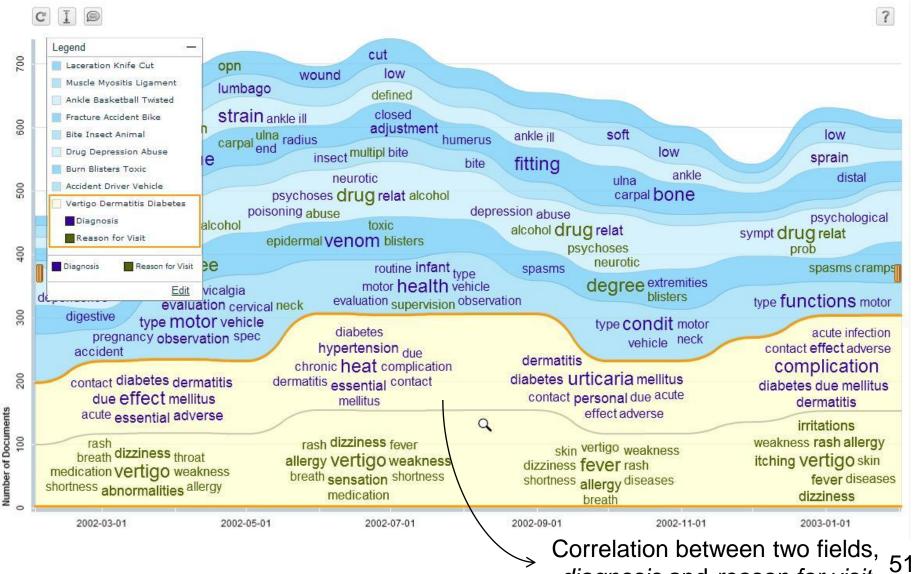
## **Application Example: Healthcare**

- Visualize unstructured data (text) to facilitate analysis
  - Cause of injury
  - Reason for visit
  - Diagnosis
- Handle multiple fields of text data and show their correlation
- Leverage structured data to help better illustrate text information
  - Gender + Cause of injury

## **Correlation between Structured and Text Fields**



## **Correlation between Two Text Fields**



diagnosis and reason for visit

# Outline

Example tasks in text analytics

# Visually analyzing textual information

- Dynamic Word Cloud
- Topic-based Visual Text Summarization
- TextFlow: Towards Better Understanding of Evolving Topics in Text
- Text Visualization Perspectives

## TextFlow: Towards Better Understanding of Evolving Topics in Text

Cui et al. Infovis 11

#### Problems

- Understanding topic evolution in large text collections is important
  - Keep abreast of hot, new, and intertwining topics
  - Gain insight into the latent topics

#### Challenges

- Model topic merging/splitting patterns
- Visually convey the topic merging/splitting patterns in an intuitive way
- Facilitate analytical reasoning

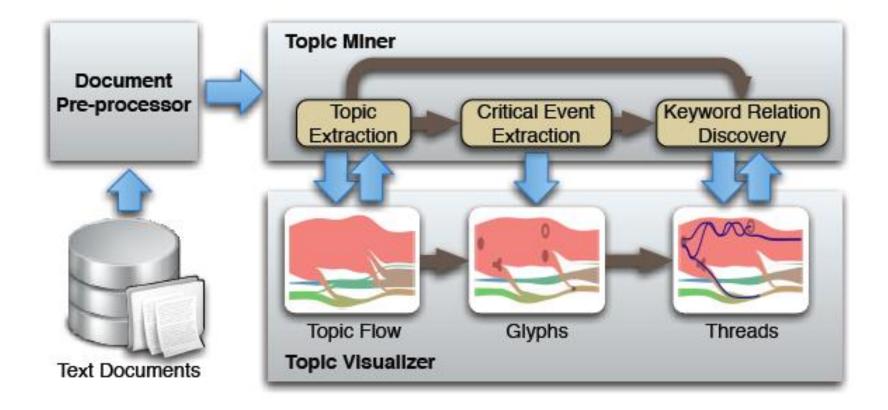
#### Solutions

- Leverage *Hierarchical Dirichlet Processes* to model topic merging/splitting
- Augment the familiar visual metaphor, the river flow, to convey the complex analytic results
- Interact with the topic from global structure to local salient features

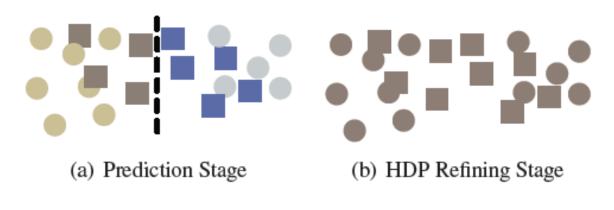
## **Related Work**

- Little work has focused on studying topic merging and splitting patterns
- It has barely been touched by using visual analysis techniques to interactively analyze complex topic evolution from multiple perspectives

## **TextFlow Overview**



## **Topic Data and Relationship Extraction**



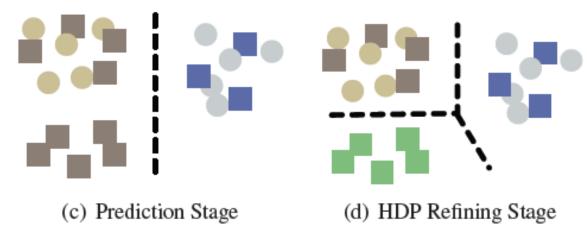


Fig. 3. An example of splitting/merging of clusters: circles representing samples at time t - 1; rectangles encoding samples at time t.

## **Merging/Splitting Relationships**

• Merging input from time *t-1* to *t*:

$$P_t^{in}(s \to r) \stackrel{\Delta}{=} \frac{\sum_{\tau=t-T_{win}+1}^{t} \sum_{i,j} I(z_{ji}^{t,old} = s \& z_{ji}^{t,new} = r)}{\sum_{\tau=t-T_{win}+1}^{t} \sum_{i=1}^{n^t} I(z_{ji}^{t,new} = r)}$$

Splitting output from time t-1 to t:

$$P_{t-1}^{out}(s \to r) \stackrel{\Delta}{=} \frac{\sum_{\tau=t-T_{win}+1}^{t} \sum_{j,i} I(z_{ji}^{\tau,old} = s \& z_{ji}^{\tau,new} = r)}{\sum_{\tau=t-T_{win}+1}^{t} \sum_{j,i} I(z_{ji}^{\tau,old} = s)}.$$

## **Critical Event Extraction**

### Critical events

- Birth, death, merge, and split

### Score of a merging event

Domain-dependent activeness metric

$$R(r,t) = |\mathcal{N}_r| \cdot H_t(r) = |\mathcal{N}_r| \cdot \kappa_B \sum_{s \in \mathcal{N}_r} -P_t^{in}(s \to r) \ln P_t^{in}(s \to r)$$
  
Neighborhood Entropy score set

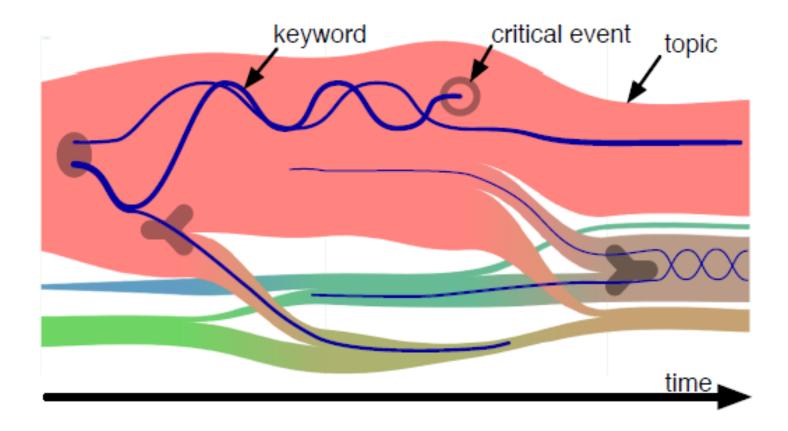
Score of a splitting event

$$R(s,t) = |\mathcal{N}_s| \cdot H_t(r) = |\mathcal{N}_s| \cdot \kappa_B \sum_{r \in \mathcal{N}_s} -P_{t-1}^{out}(s \to r) \ln P_{t-1}^{out}(s \to r)$$

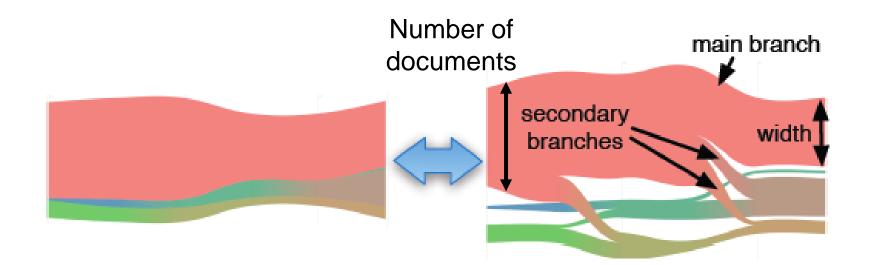
## **Keyword Correlation Discovery**

 Extract "noun phrases," "verb phrases", and "named entities" in each document, and count co-occurrences among them

## **Visualization Design**



## **Topic Evolution as Flow**

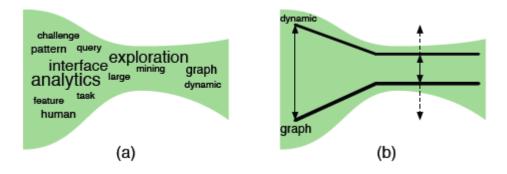


## **Critical Event as Glyph**

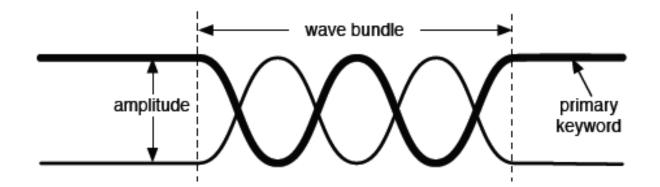


# **Keyword Correlation as Thread**

### Alternatives



Keyword thread



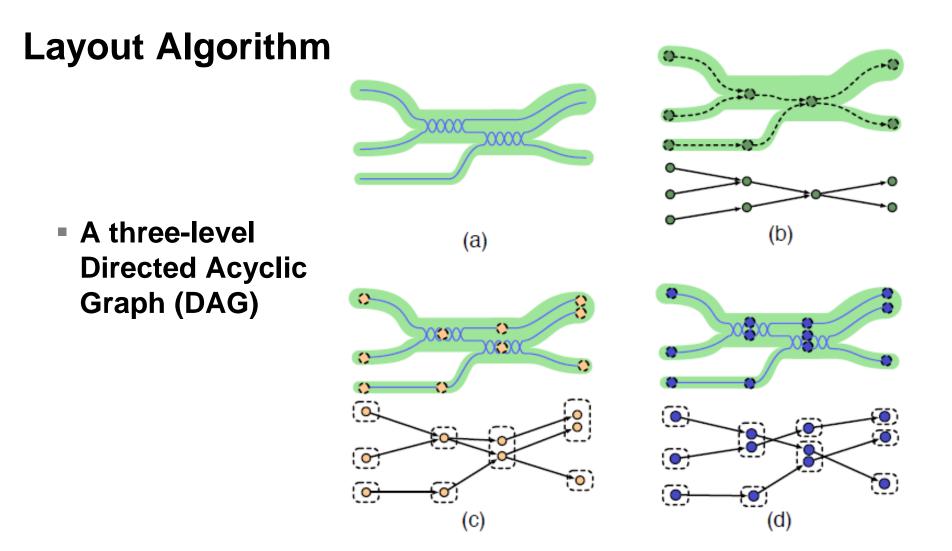


Fig. 9. The three-level model for topic flow graph layout (dotted rectangles indicating boundary constraints): (a) the original structure; (b) first level: topic flows; (c) second level: bundles; (d) third level: threads.

## **Interactive Exploration**

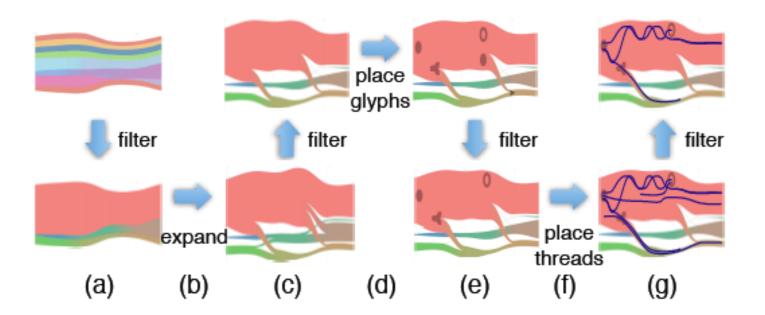
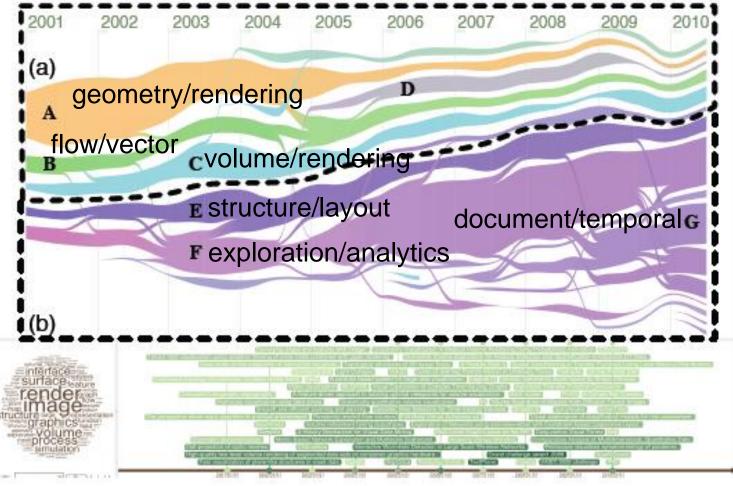


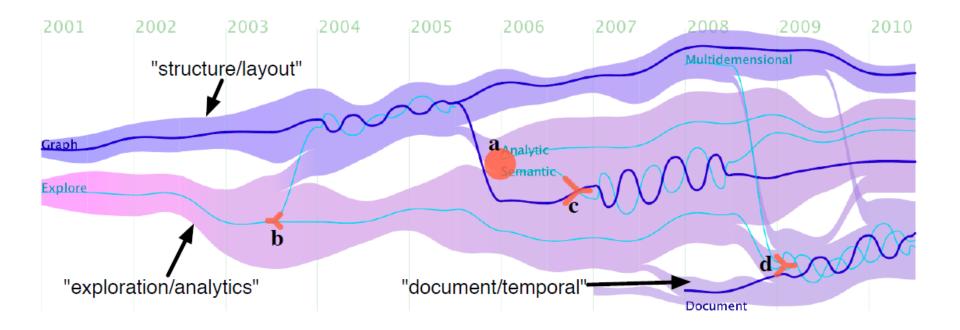
Fig. 11. The exploration pipeline: (a) select a subset of interest based on users' interest or system recommendation; (b) expand the topics to see their splitting and merging patterns; (c) remove trivial or irrelevant branches; (d) and (e) extract major critical events based on current flow patterns; (f) and (g) explore and adjust threads around the selected critical events based on users' interest or system recommendation.

## **Application Example – VisWeek Pulications**

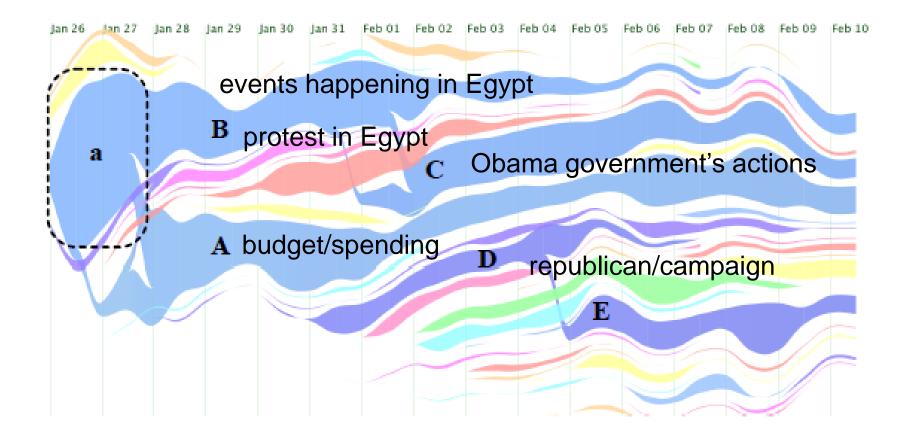


933 research papers

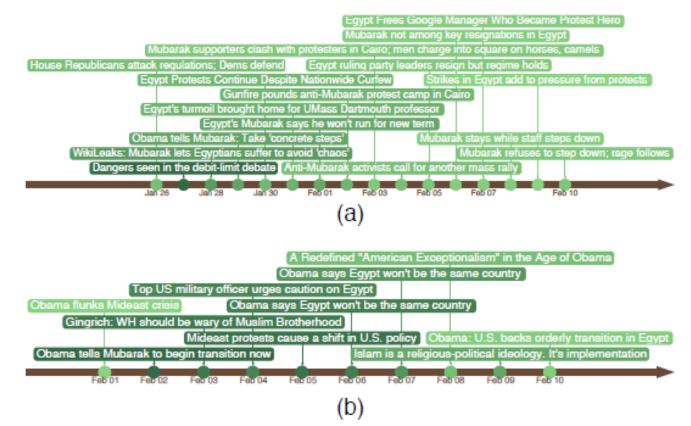
### **Application Example - VisWeek Pulications**



## **Application Example – Bing News**

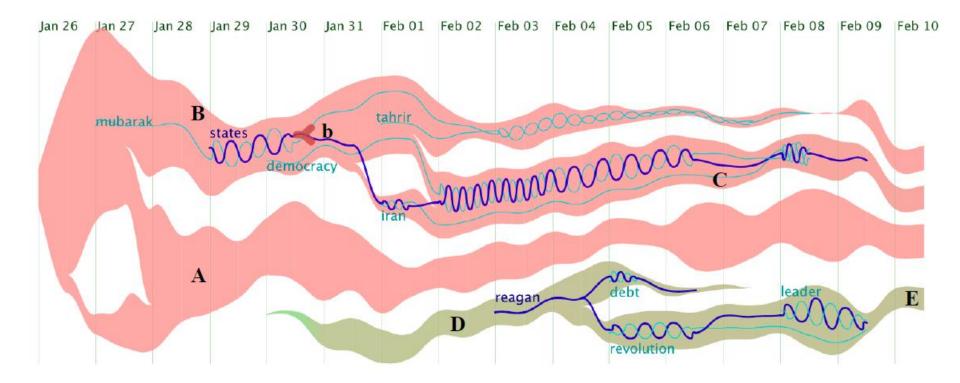


## **Application Example – Bing News**

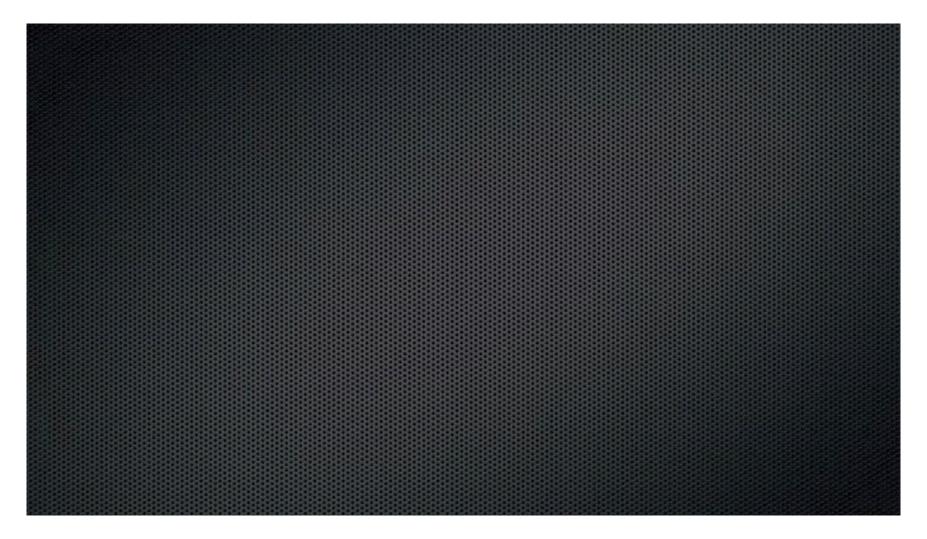


Comparing timelines extracted from different topics: (a) timeline extracted from topic B, which mainly describes protest events happening in Egypt; (b) timeline extracted from topic C, which focuses on the Obama's government's actions on events in Egypt.

## **Application Example – Bing News**



## Video



# Outline

- Example tasks in text analytics
- Visually analyzing textual information
  - Dynamic Word Cloud
  - Topic-based Visual Text Summarization
  - TextFlow: Towards Better Understanding of Evolving Topics in Text

# Text Visualization Perspectives

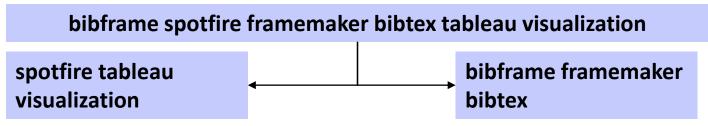
# **Future Text Visualization Topics**

Interactive, incremental text analytics

#### **Topic editing**

TIARA LDA latent text semantic **David** models **edt** keywords summarization topic

#### **Topic split**



- Multi-level visual text summarization (keywords + sentences)
- Multi-faceted text analytics (e.g., summarization + sentimental analysis)
- Multimedia document summarization (text + image + video)
- Interactive, visual social media analysis

## **Acknowledgements**

Nan Cao, Yingcai Wu, Weiwei Cui, Prof. Huamin Qu (HKUST) Dr. Michelle X Zhou (IBM Almaden Research Center)



## Thanks a lot for your attention!

