Discovering Evolutionary Theme Patterns from Text
An Exploration of Temporal Text Mining

Qiaozhu Mei ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana Champaign

August 21-24, 2005 / KDD'05

Michael Stockerl

Outline

- Introduction
- Problem Formulation
- Evolution Graph Discovery
  - Theme Extraction
  - Evolutionary Transition Discovery
- Theme Life Cycles
- Experiments
- Summary

Temporal Text Mining (TTM) is concerned with discovering temporal patterns in text information collected over time.

Discovering Evolutionary Theme Patterns from Text

An Exploration of Temporal Text Mining

almost every document has a meaningful time stamp, therefore we could find. . .

- Temporal patterns
- An underlying temporal and evolutionary structure consisting of suptopics/themes
- The start, progression of the event and the impact on other events

Task: Find these evolutionary theme patterns (ETP) automatically

Why are we interested in ETP?

- Organization of the stream according to the underlying thematic structure
- Navigation through all these documents
- Summarization of the event/topic, including
  - Subtopics
  - Threads
- Life cycles

How will we find the ETP?

1. Discovering interesting global and outstanding local themes in a given time range
2. Discovering theme evolutionary relations and building an evolution graph of themes
3. Modeling theme strength over time and analyzing the life cycles of themes

Applications

- Mining user logs
- Mining costumer reviews
- Email analysis
- Finding trends in social media
- Recommendation system
- Etc.

Outline

- Introduction
- Problem Formulation
- Evolution Graph Discovery
  - Theme Extraction
  - Evolutionary Transition Discovery
- Theme Life Cycles
- Experiments
- Summary

Definition 1: Theme

- probabilistic distribution of words that characterizes a topic
- a theme is represented by a unigram language model Θ in the following
- high probability words are mostly what the theme about

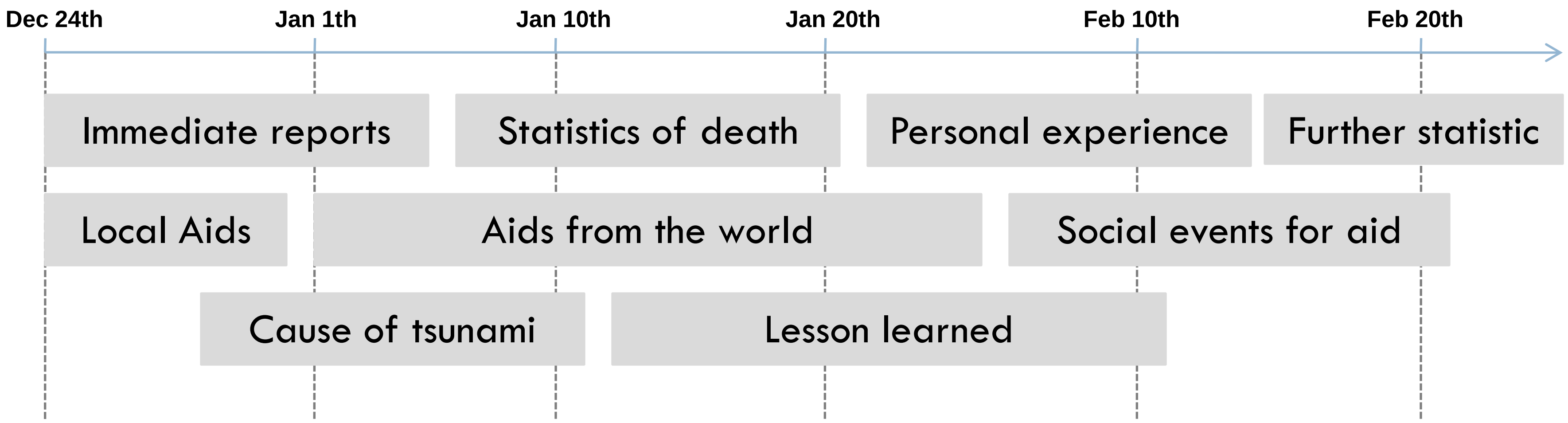
Immediate reports | Statistics of death | Personal experience | Further statistic

Local Aids | Aids from the world | Social events for aid

Cause of tsunami | Lesson learned

Definition 2: Theme span

- A theme Θ that spans a given interval *I*
- Represented by ‹Θ, s(γ), t(γ)›
- useful to correlate themes with time
- we will use themes and theme spans as synonyms
- a theme span is a transcollection theme, if s = 1 and t = T



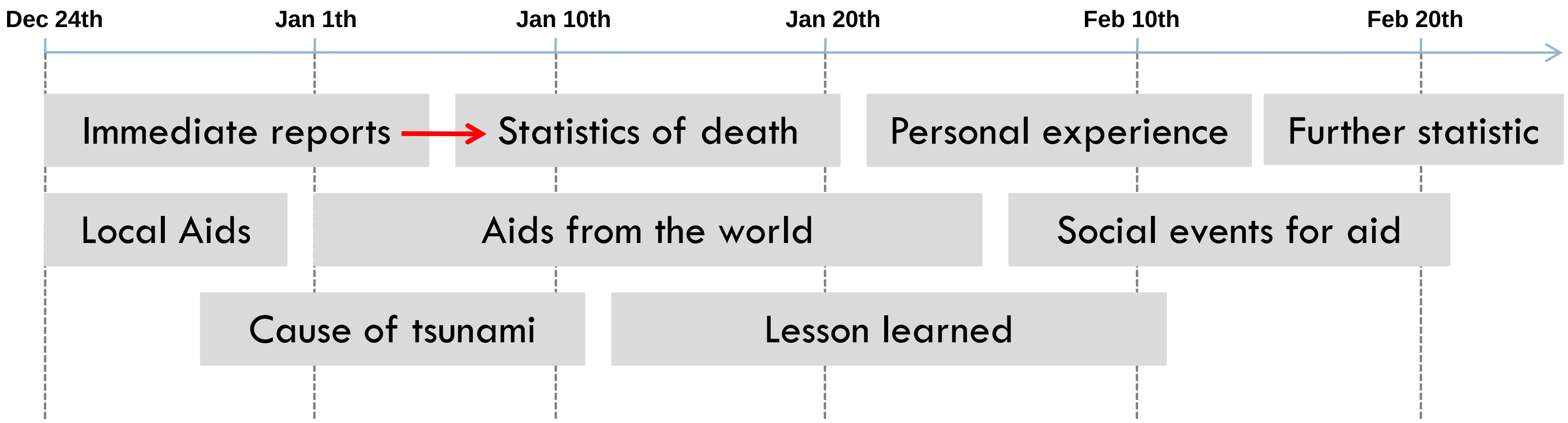
Definition 3: Evolutionary Transition

Given: $\gamma_1 = \langle \Theta_1, s(\gamma_1), t(\gamma_1) \rangle$ and $\gamma_2 = \langle \Theta_2, s(\gamma_2), t(\gamma_2) \rangle$

There is an evolutionary transition from γ_1, γ_2 (denoted: $\gamma_1 \prec \gamma_2$), if

- The similarity between γ_1 and γ_2 is above a threshold
- $t(\gamma_1) \leq s(\gamma_2)$

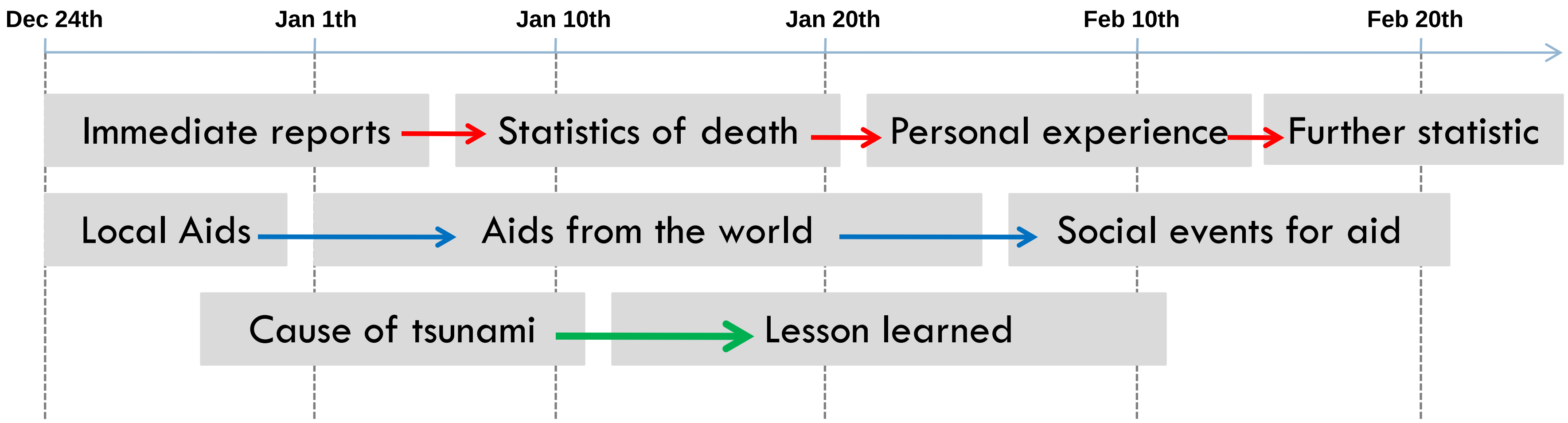
We can describe relations between themes now.



Definition 4: Theme Evolution Graph

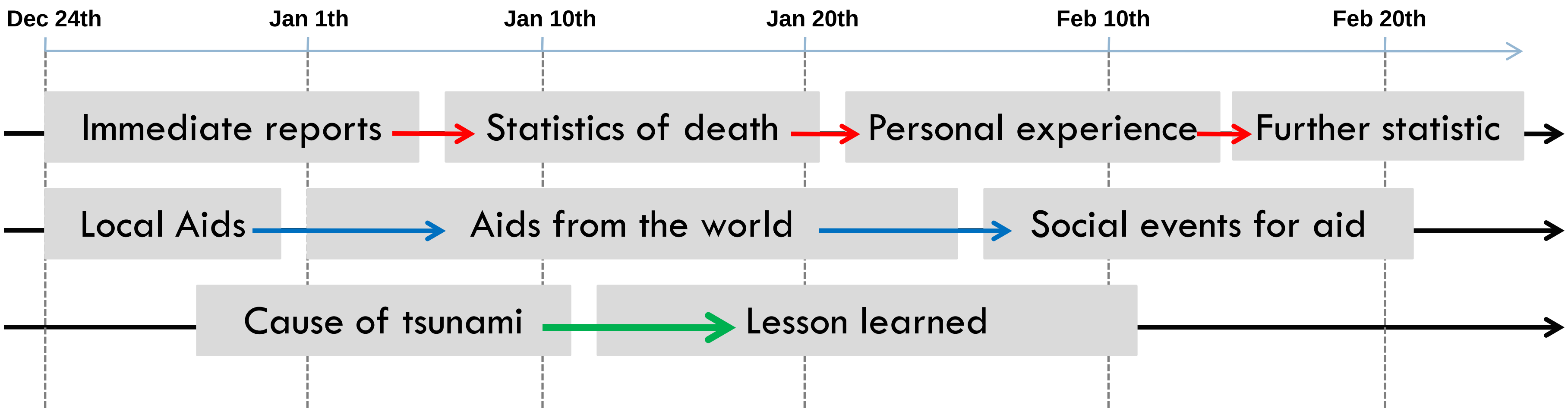
Weighted directed graph G = (N,E), where

- Each vertex $v \in N$ is a theme span
- Each edge $e \in E$ is an evolutionary transition
- The weight on the edge represents the evolutionary distance



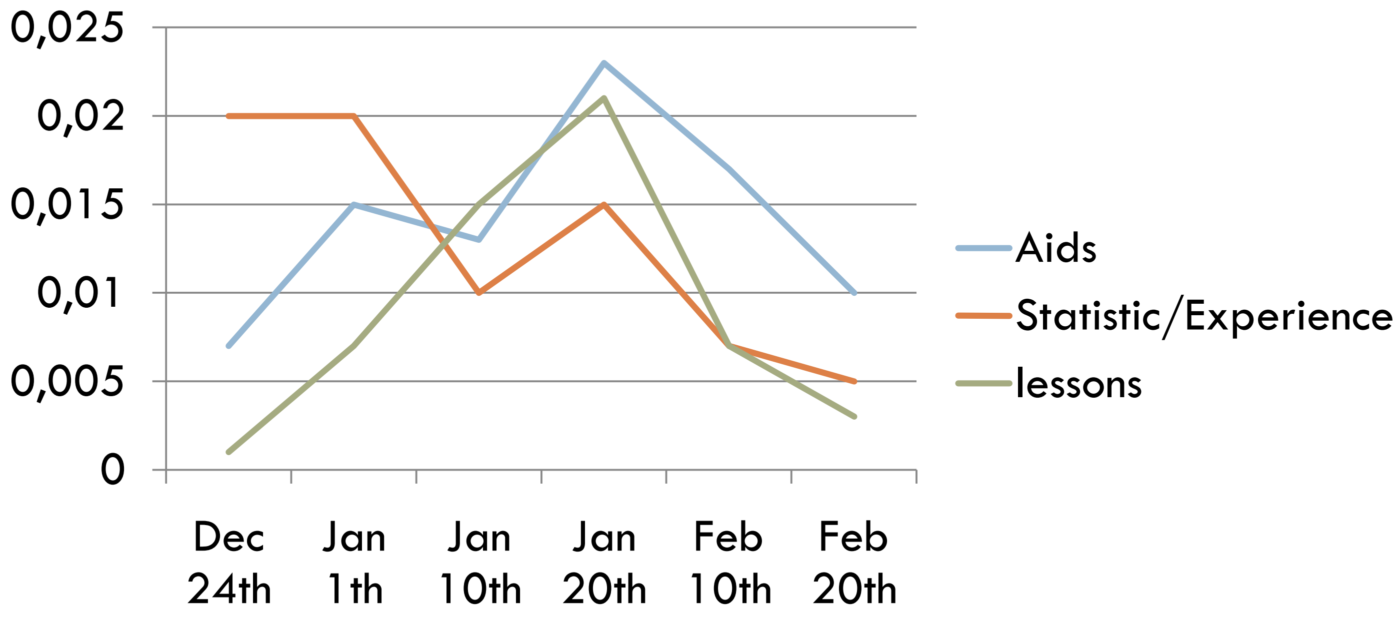
Definition 5: Theme Evolution Thread

- each path through the graph is a theme evolution thread
- characterize how related themes evolve over time



Definition 6: Theme Life Cycle of a theme

- strength distribution of the theme over the entire time line
- strength is measured by the number of words generated by the topic in a time interval
- two strength types:
  - relative strength: normalized with the total number of words in the period
  - absolute strength: normalized by the number of time points

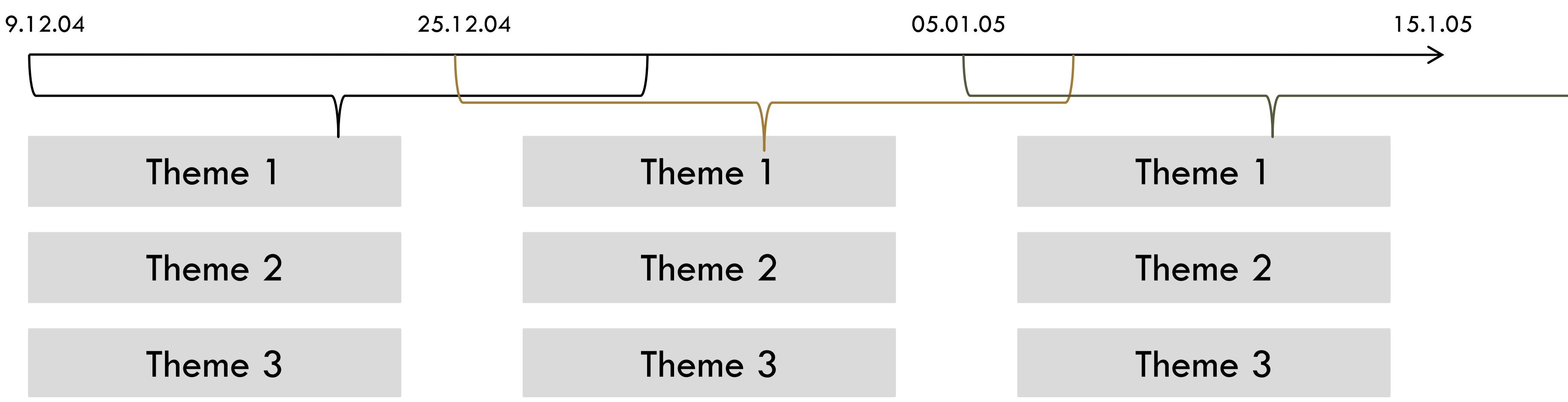


Outline

- Introduction
- Problem Formulation
- Evolution Graph Discovery
  - Theme Extraction
  - Evolutionary Transition Discovery
- Theme Life Cycles
- Experiments
- Summary

Roughly process

1. Partition the documents into n (possibly overlapping) subcollections with fixed or variable time interval
2. Extract the most outstanding themes from each subcollections using a probabilistic mixture model
3. Find the evolutionary transitions based on the similarity of the themes

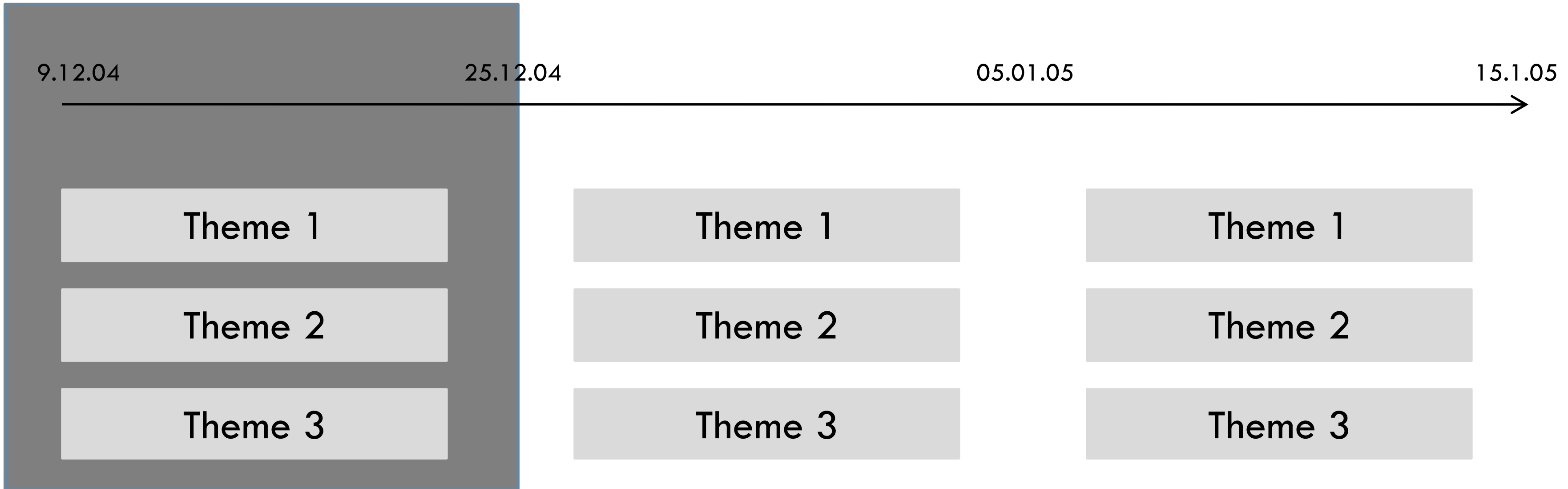


Outline

- Introduction
- Problem Formulation
- Evolution Graph Discovery
  - Theme Extraction
  - Evolutionary Transition Discovery
- Theme Life Cycles
- Experiments
- Summary

Theme Extraction

- Extracting themes from each subcollection, using a simple probabilistic mixture model
- The model could be estimated using the Expectation Maximization algorithm
- To extract the trans-collection themes, apply the model on the whole collection



The mixture model

- Words are regarded as data drawn from the mixture model
- Words in the same document share the same mixing weight $\pi_{d,j}$
- We expect *k* themes in every collection
- Each is characterized by a unigram language model
  - e.g. word distribution
- A background model should swallow the non-discriminative words

A document *d* is regarded as a sample of the following mixture model

$$p(w:d) = \lambda_B p(w \mid \theta_B) + (1-\lambda_B)\sum_{j=1}^{k}[\pi_{d,j}\, p(w \mid \theta_j)]$$

To make it easier to find the maximum, we could use the log-likelihood

$$\log p(C:\Lambda) = \sum_{d \in C_i}\sum_{w \in V}[c(w,d) * \log(\lambda_B p(w \mid \theta_B) + (1-\lambda_B)\sum_{j=1}^{k}(\pi_{d,j}\, p(w \mid \theta_j)))]$$

Task of the EM algorithm

Estimate the missing parameters with the following update formulas:

$$p(z_{d,w} = j) = \frac{\pi_{d,j}^{(n)} p^{(n)}(w \mid \theta_j)}{\sum \pi_{d,j'}^{(n)} p^{(n)}(w \mid \theta_k)} \qquad p(z_{d,w} = B) = \frac{\lambda_B p(w \mid \theta_B)}{\lambda_B p(w \mid \theta_B) + (1-\lambda_B)\sum_{j=1}^{k}[\pi_{d,j} p(w \mid \theta_j)]}$$

$$\pi_{d,j}^{(n+1)} = \frac{\sum_{w \in V} c(w,d)(1 - p(z_{d,w} = B))\, p(z_{d,w} = j)}{\sum_{l=1}^{k} \sum_{w \in V} c(w,d)(1 - p(z_{d,w} = B))\, p(z_{d,w} = l)}$$

$$p^{(n+1)}(w \mid \theta_j) = \frac{\sum_{d \in C} c(w,d)(1 - p(z_{d,w} = B))\, p(z_{d,w} = j)}{\sum_{w' \in V} \sum_{d \in C} c(w',d)(1 - p(z_{d,w'} = B))\, p(z_{d,w'} = j)}$$

Outline

- Introduction
- Problem Formulation
- Evolution Graph Discovery
  - Theme Extraction
  - Evolutionary Transition Discovery
- Theme Life Cycles
- Experiments
- Summary

Kullback-Leibler divergence

- Measure of the difference between two probability distributions P and Q, whereas…
  - P represents a true distribution (data, observations or precisely calculated theoretical distribution)
  - Q represents a theory, model, description or approximation of P

→ Measures the information gain from a prior to a posterior distribution

- Formula: $D(P \| Q) = \sum_{i=1}^{|V|} \ln\left(\frac{P(i)}{Q(i)}\right) P(i)$
- Non-symmetric
- D(P || Q) = 0, if and only if P = Q
- Only defined, when P and Q both sum to 1
- If Q(i) = 0 → P(i) = 0, for all i

Evolutionary Transition Discovery

Let $\gamma_1 = \langle \theta_1, s(\gamma_1), t(\gamma_1) \rangle$ and $\gamma_2 = \langle \theta_2, s(\gamma_2), t(\gamma_2) \rangle$ be two theme spans, where $t(\gamma_1) \le s(\gamma_2)$

- If the language models θ_2 and θ_1 are close to each other, γ_1 and γ_2 have a small evolution distance
- KL –Divergence $D(\theta_2 \| \theta_1)$ can model the new information from θ_2 compared to θ_1
- If $D(\theta_2 \| \theta_1)$ is below a threshold, there exists a evolutionary transition (denoted as $\gamma_1 \prec \gamma_2$)

Summary of theme evolutionary graph

- right now: microcosmic view of the ETPs
  - major themes of every time interval
  - evolutionary structure of the themes
- in the following: macroscopic view of the ETPs
  - global evolutionary patterns of the transcollection themes
  - analyze the life cycle of every theme

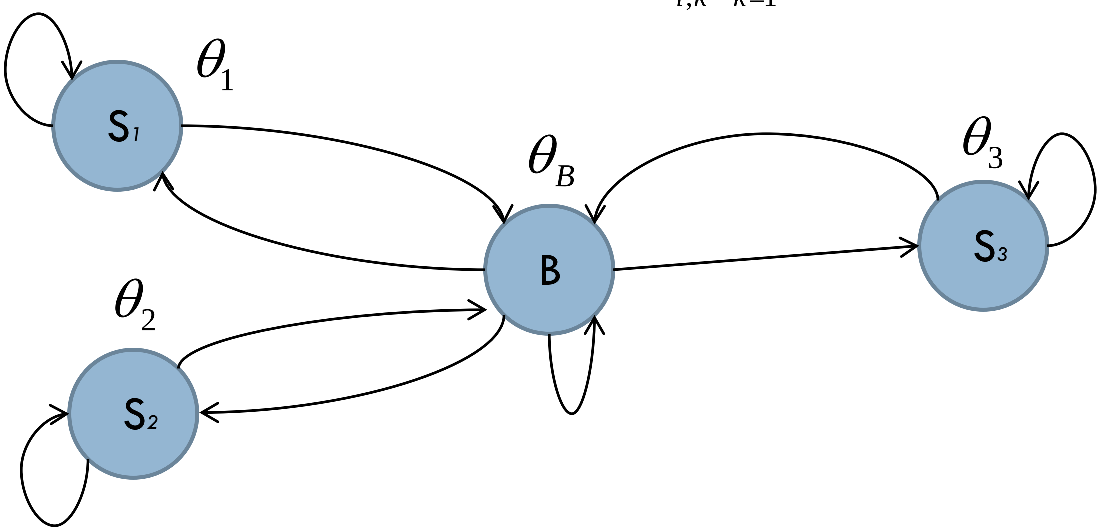
Outline

- Introduction
- Problem Formulation
- Evolution Graph Discovery
  - Theme Extraction
  - Evolutionary Transition Discovery
- Theme Life Cycles
- Experiments
- Summary

Hidden Markov Models (HHM)

An HMM could be characterized by …

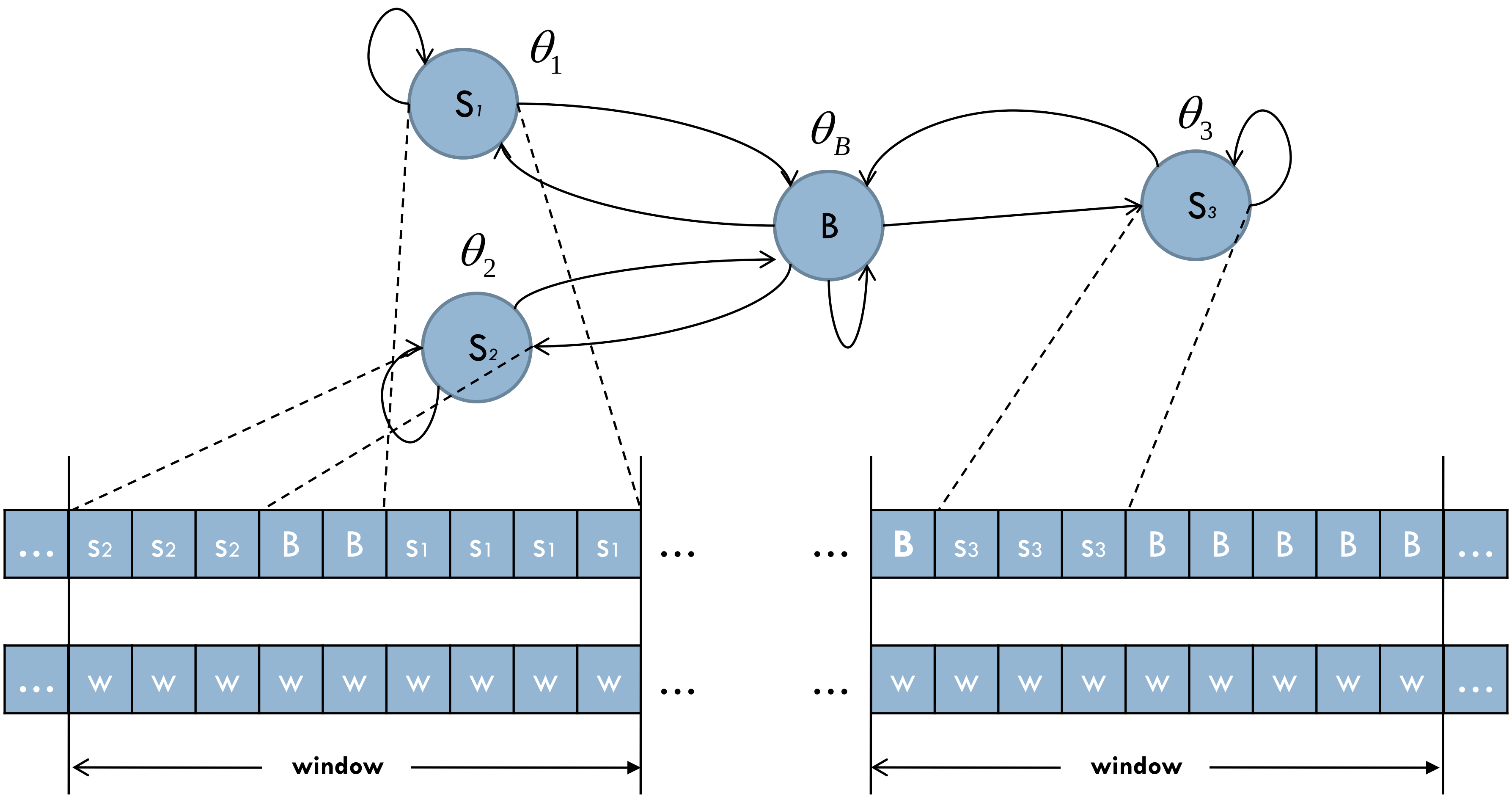
- A set of hidden states O = {s_1,…,s_n}
- A set of observable output symbols O = {o_1,…,o_m}
- A initial state probability distribution $\{\pi\}_{j=1}^{n}$
- A state transition probability distribution $\{a_{i,j}\}_{j=1}^{n}$ for each state s_i
- A output probability distribution $\{b_{i,k}\}_{k=1}^{m}$ for each state s_i



Model the theme shifts

1. Construct an HMM to model how themes shift
   - Extract k trans-collection themes from the text data
   - Construct a fully connected HMM with k+1 states
2. Estimate the unknown parameters of the HMM using the whole collection as observed data
3. Decode the collection and label each word with the hidden theme model from which it is generated
4. Analyze when the themes start , when they terminate and how they develope over time

Decoding the model



Absolute strength and relative strength

$$AStrength\ (i,t) = \frac{1}{W} \sum_{t' \in [t-\frac{W}{2}, t+\frac{W}{2}]} \sum_{j=1}^{|d_{t'}|} \delta(d_{t'j}, i)$$

$$NStrength\ (i,t) = \frac{AStrength\ (i,t)}{\sum_{j=1}^{k} AStrength\ (j,t)}$$

$$= \frac{\sum_{t' \in [t-\frac{W}{2}, t+\frac{W}{2}]} \sum_{j=1}^{|d_{t'}|} \delta(d_{t'j}, i)}{\sum_{t' \in [t-\frac{W}{2}, t+\frac{W}{2}]} |d_{t'}|}$$

Where $\delta(d_{t'j}, i) = 1$, if word $d_{t'j}$ is labeled as theme i

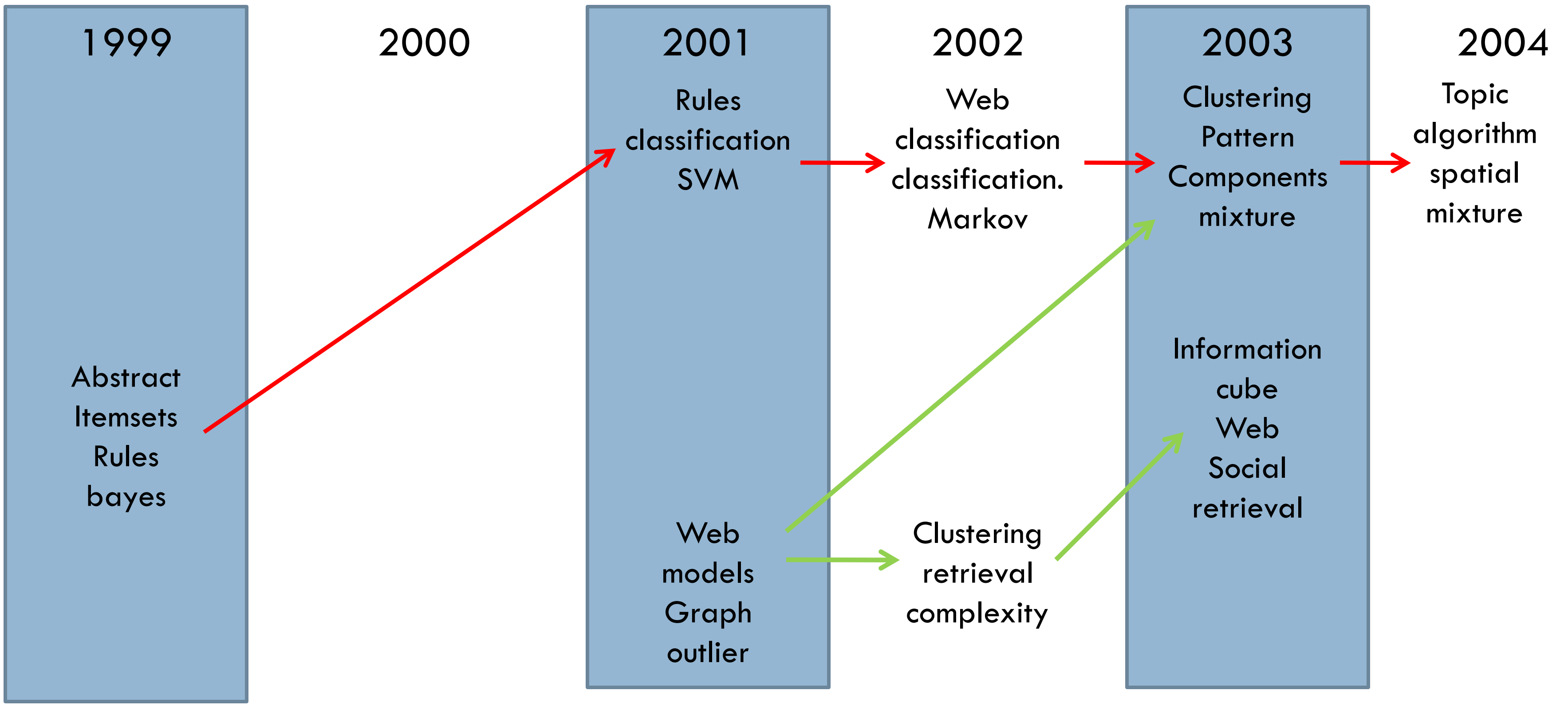
Outline

- Introduction
- Problem Formulation
- Evolution Graph Discovery
  - Theme Extraction
  - Evolutionary Transition Discovery
- Theme Life Cycles
- Experiments
- Summary

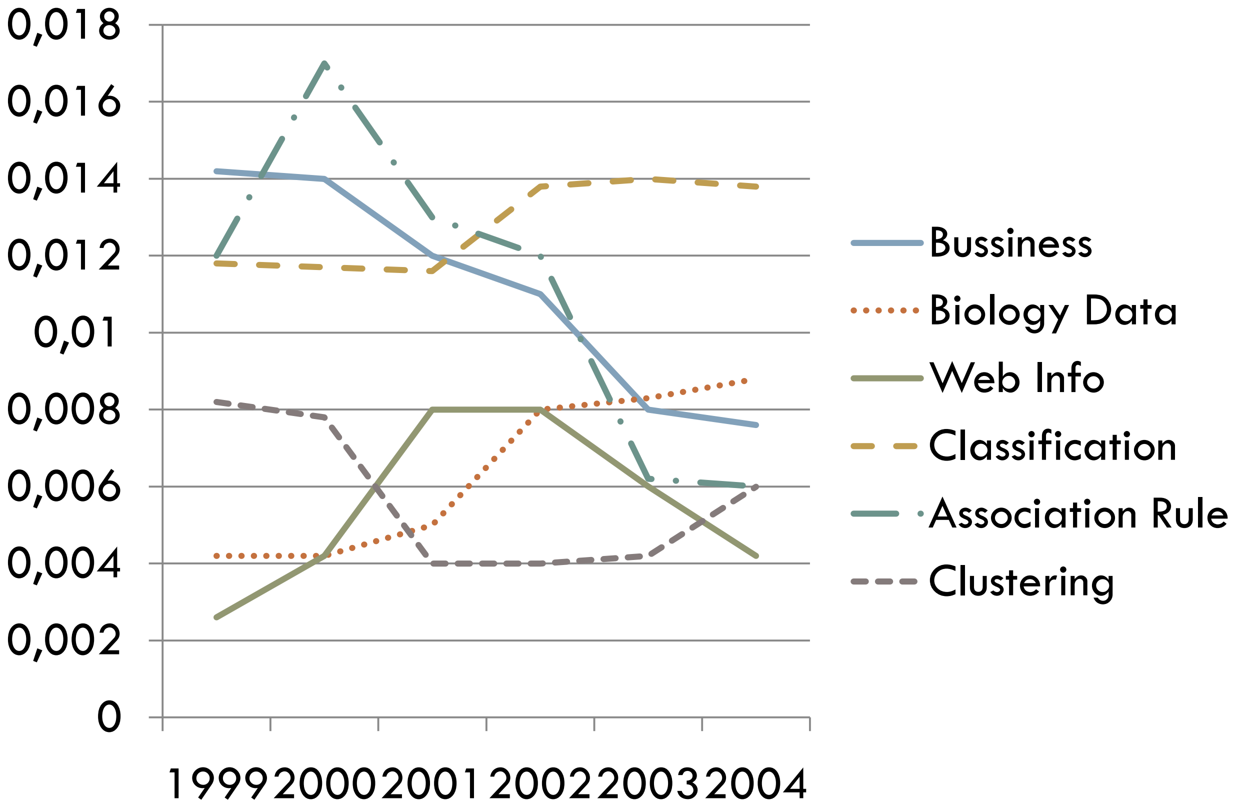
Data sets

- 7468 news articles about the Asian Tsunami from 19.12.2004 to 8.2.2005
- 469 abstracts in KDD conference proceedings from 1999 to 2004

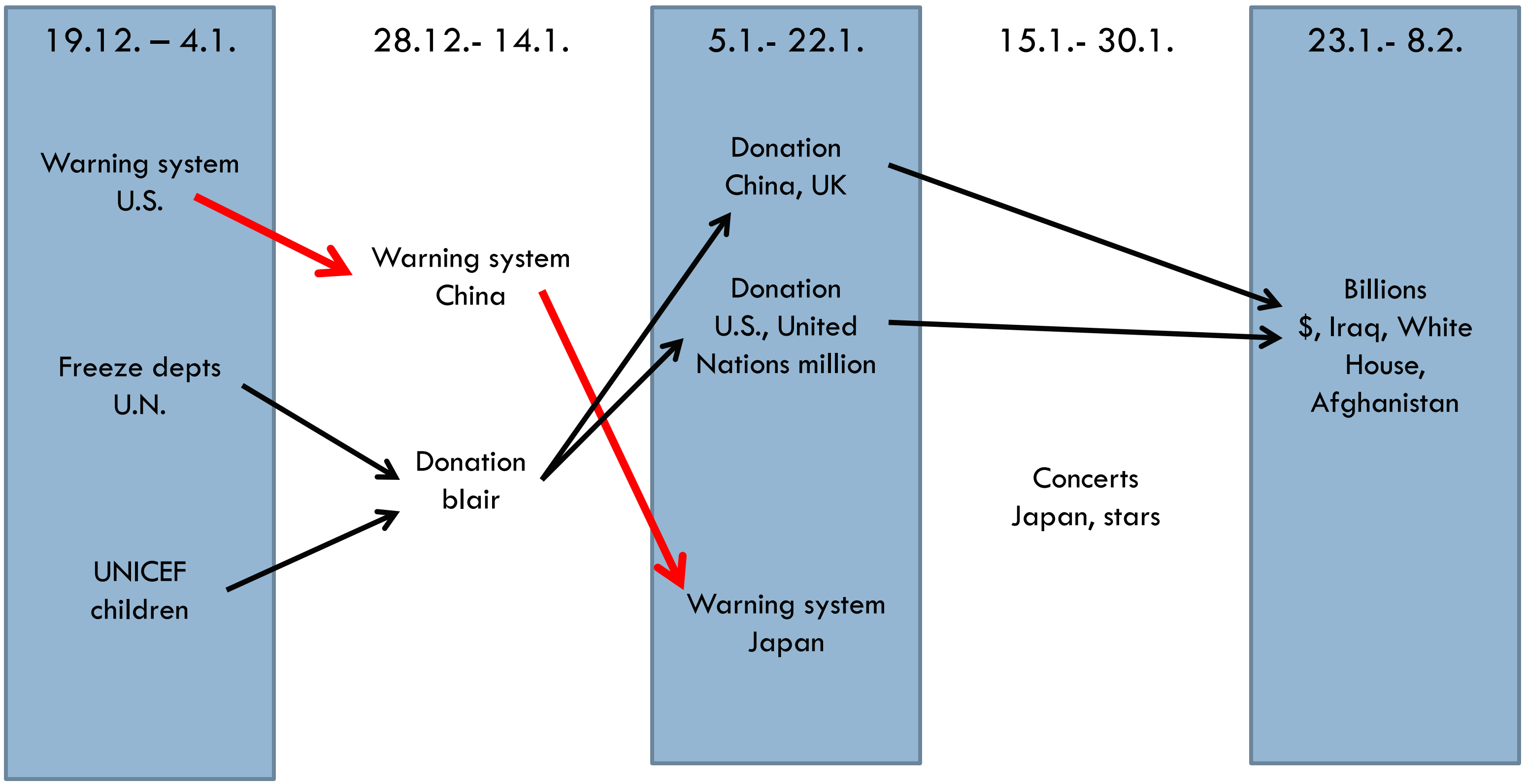
Theme evolutionary graph (KDD example)



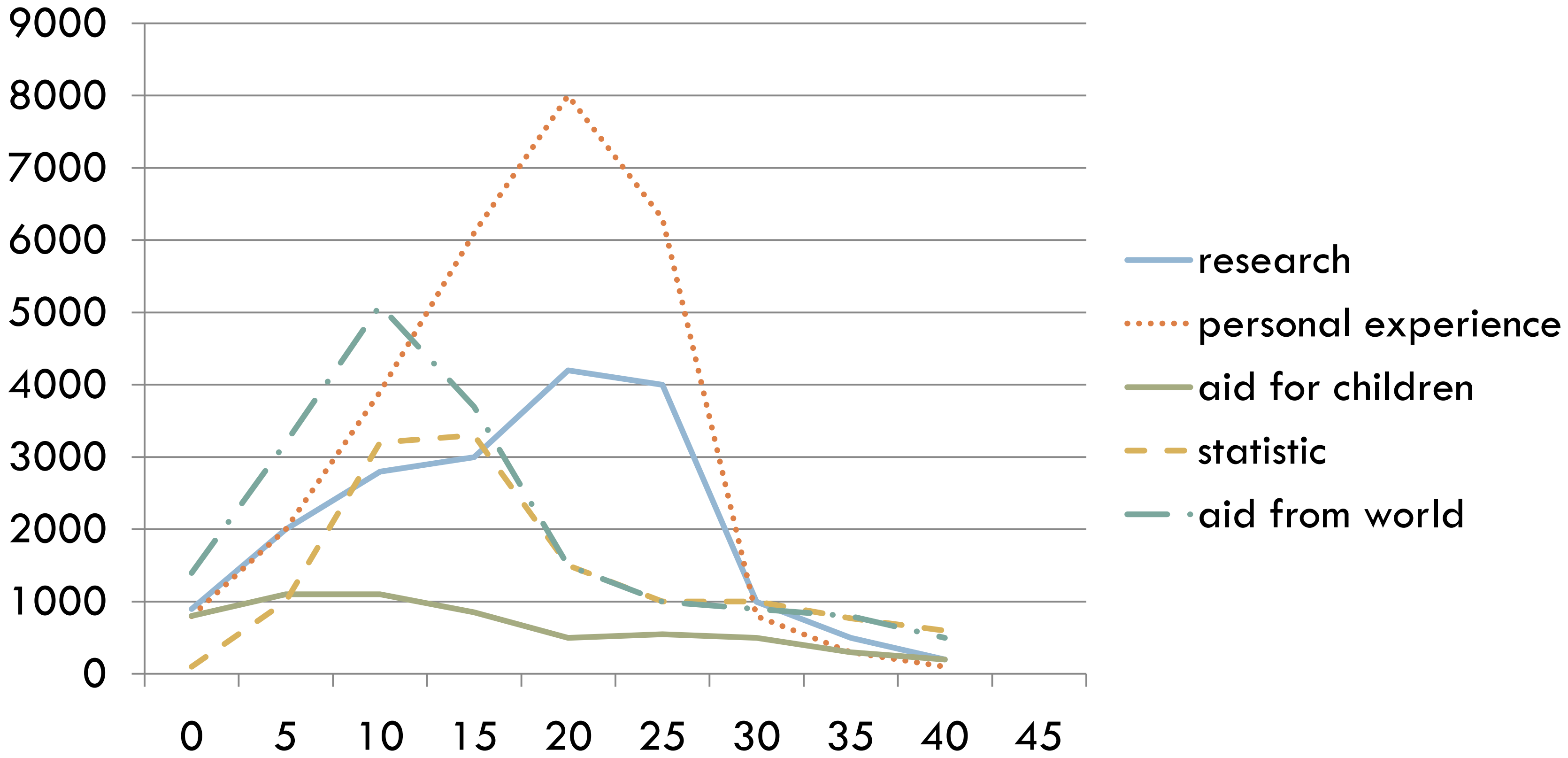
Life cycle of the KDD example



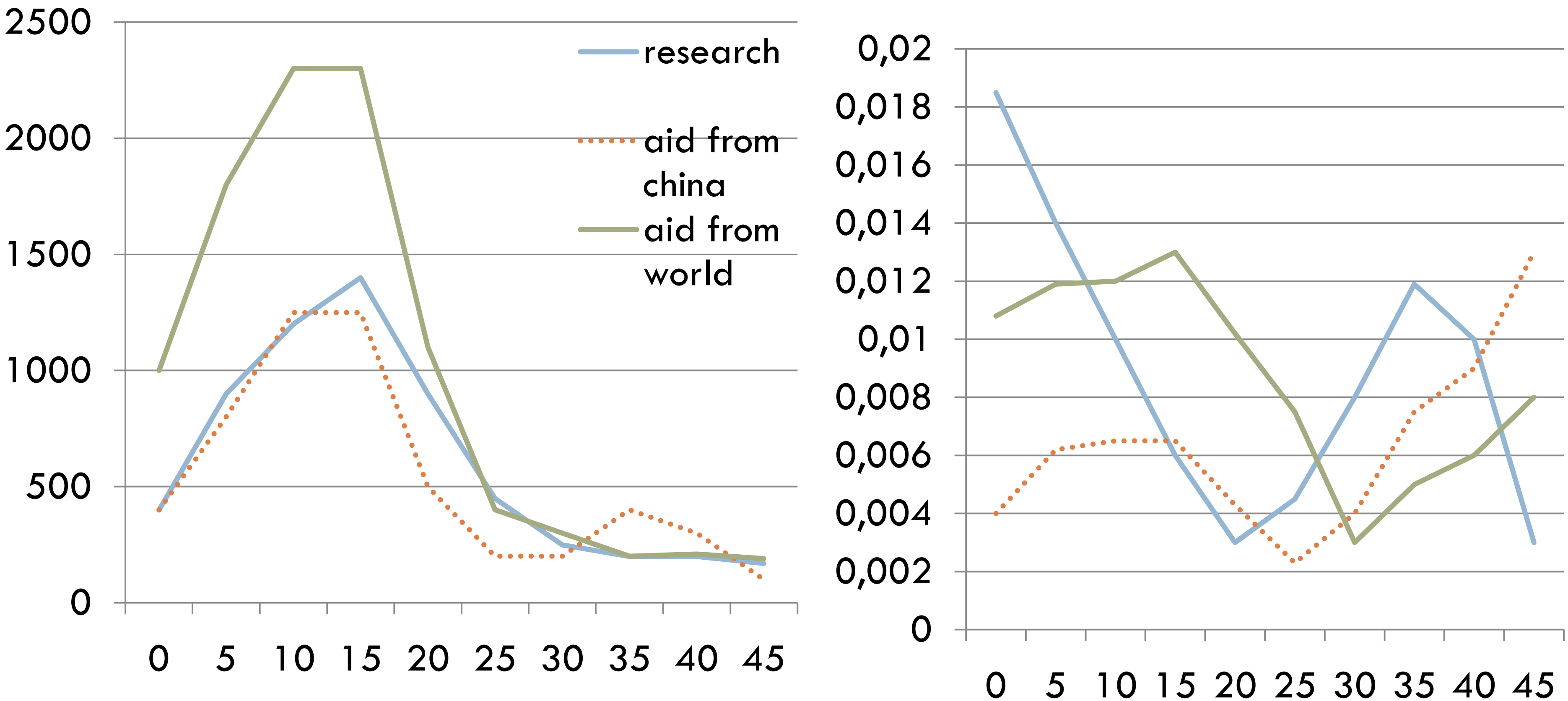
Theme evolutionary graph (Tsunami example)



Life cycle of the Tsunami example (CNN)



Life cycle of the Tsunami example (Xinhua)



Outline

- Introduction
- Problem Formulation
- Evolution Graph Discovery
  - Theme Extraction
  - Evolutionary Transition Discovery
- Theme Life Cycles
- Experiments
- Summary

Summary

- Given a text stream C, the most important task of ETP discovery problem is to extract a theme evolutionary graph from C automatically.
- graph could be used as summary of the themes and their evolutionary relationship
- can organize the data in a meaningful way

Pro & Contra

- Advantages:
  - unsupervised task
  - summary of a complete topic
  - navigation through the data stream
  - robust (no stemming and stopword removal)
- Disadvantages:
  - unsupervised task
  - expensive calculation
  - extracted words are not always meaningful
  - EM - algorithm only finds local maximums