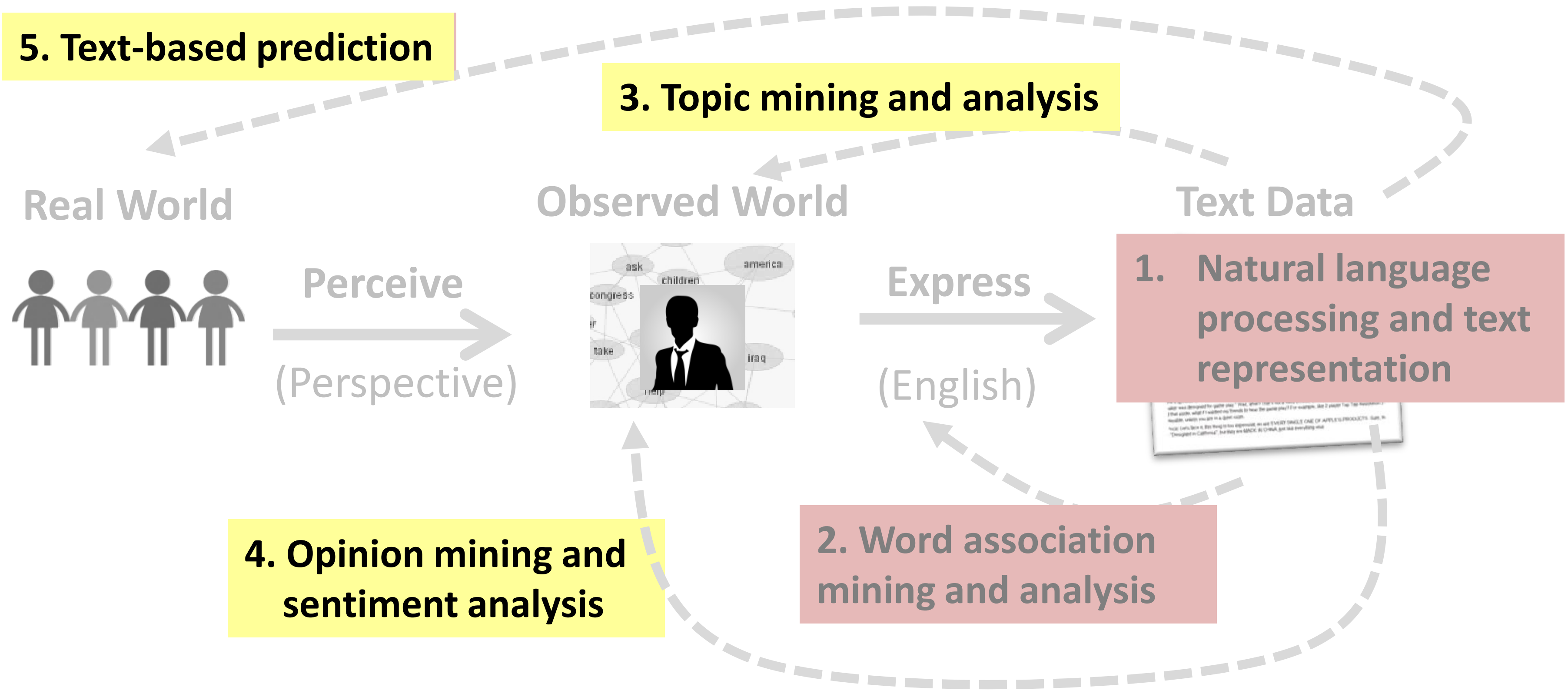
# Contextual Text Mining: Mining Topics with Social Network Context

ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

# Contextual Text Mining:
# Mining Topics with Social Network as Context



# Topic Analysis with Network Context

- The **context** of a text article can form a **network**, e.g.,
  - Authors of research articles may form **collaboration networks**
  - Authors of social media content form **social networks**
  - Locations associated with text can be connected to form a **geographic network**
- **Benefit** of **joint analysis** of text and its network context
  - Network imposes **constraints** on topics in text (**authors connected in a network tend to write about similar topics**)
  - Text helps **characterize** the content associated with each subnetwork (e.g., difference in opinions expressed in two subnetworks?)

# Network Supervised Topic Modeling: General Idea [Mei et al. 08]

- Probabilistic topic modeling as optimization: maximize likelihood

$$\Lambda^* = \arg\max_{\Lambda} p(\text{TextData} \mid \Lambda)$$

- Main idea: network imposes constraints on model parameters $\Lambda$
  - The text at two adjacent nodes of the network tends to cover similar topics
  - Topic distributions are smoothed over adjacent nodes
  - Add network-induced regularizers to the likelihood objective function

**Any generative model**

**Any network**

$$\Lambda^* = \arg\max_{\Lambda} f(p(\text{TextData} \mid \Lambda), r(\Lambda, \text{Network}))$$

**Any way to combine**

**Any regularizer**

# Instantiation: NetPLSA [Mei et al. 08]

Network-induced prior: Neighbors have similar topic distribution

**Modified objective function**

**Text collection**

**PLSA log-likelihood**

$$O(C,G) = (1-\lambda)\cdot\boxed{\left(\sum_{d}\sum_{w} c(w,d)\log\sum_{j=1}^{k} p(\theta_j \mid d)p(w \mid \theta_j)\right)}$$

**Network graph**

$$+\boxed{\lambda}\cdot\left(-\frac{1}{2}\sum_{\langle u,v\rangle\in E} \underline{w(u,v)}\,\underline{\sum_{j=1}^{k}\left(p(\theta_j \mid u)-p(\theta_j \mid v)\right)^2}\right)$$

**Influence of network constraint**

**Weight of edge (u,v)**

**Quantify the difference in the topic coverage at node u and v**

# Mining 4 Topical Communities: Results of PLSA

**Can’t uncover the 4 communities (IR, DM, ML, Web)**

| Topic 1 | | Topic 2 | | Topic 3 | | Topic 4 | |
|---|---|---|---|---|---|---|---|
| term | 0.02 | peer | 0.02 | visual | 0.02 | interface | 0.02 |
| question | 0.02 | patterns | 0.01 | analog | 0.02 | towards | 0.02 |
| protein | 0.01 | mining | 0.01 | neurons | 0.02 | browsing | 0.02 |
| training | 0.01 | clusters | 0.01 | vlsi | 0.01 | xml | 0.01 |
| weighting | 0.01 | stream | 0.01 | motion | 0.01 | generation | 0.01 |
| multiple | 0.01 | frequent | 0.01 | chip | 0.01 | design | 0.01 |
| recognition | 0.01 | e | 0.01 | natural | 0.01 | engine | 0.01 |
| relations | 0.01 | page | 0.01 | cortex | 0.01 | service | 0.01 |
| library | 0.01 | gene | 0.01 | spike | 0.01 | social | 0.01 |

# Mining 4 Topical Communities: Results of NetPLSA

**Uncovers the 4 communities well**

| Information Retrieval | | Data Mining | | Machine Learning | | Web | |
|---|---|---|---|---|---|---|---|
| retrieval | 0.13 | mining | 0.11 | neural | 0.06 | web | 0.05 |
| information | 0.05 | data | 0.06 | learning | 0.02 | services | 0.03 |
| document | 0.03 | discovery | 0.03 | networks | 0.02 | semantic | 0.03 |
| query | 0.03 | databases | 0.02 | recognition | 0.02 | services | 0.03 |
| text | 0.03 | rules | 0.02 | analog | 0.01 | peer | 0.02 |
| search | 0.03 | association | 0.02 | vlsi | 0.01 | ontologies | 0.02 |
| evaluation | 0.02 | patterns | 0.02 | neurons | 0.01 | rdf | 0.02 |
| user | 0.02 | frequent | 0.01 | gaussian | 0.01 | management | 0.01 |
| relevance | 0.02 | streams | 0.01 | network | 0.01 | ontology | 0.01 |

# Text Information Network

- In general, we can view text data that naturally "lives" in a rich information network with all other related data
- Text data can be associated with
  - Nodes of the network
  - Edges of the network
  - Paths of the network
  - Subnetworks
  - …
- Analysis of text should be using the entire network!

# Suggested Reading

- **[Mei et al. 08]** Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. 2008. Topic modeling with network regularization. In *Proceedings of the 17th international conference on World Wide Web* (WWW 2008). ACM, New York, NY, USA, 101-110. DOI=10.1145/1367497.1367512