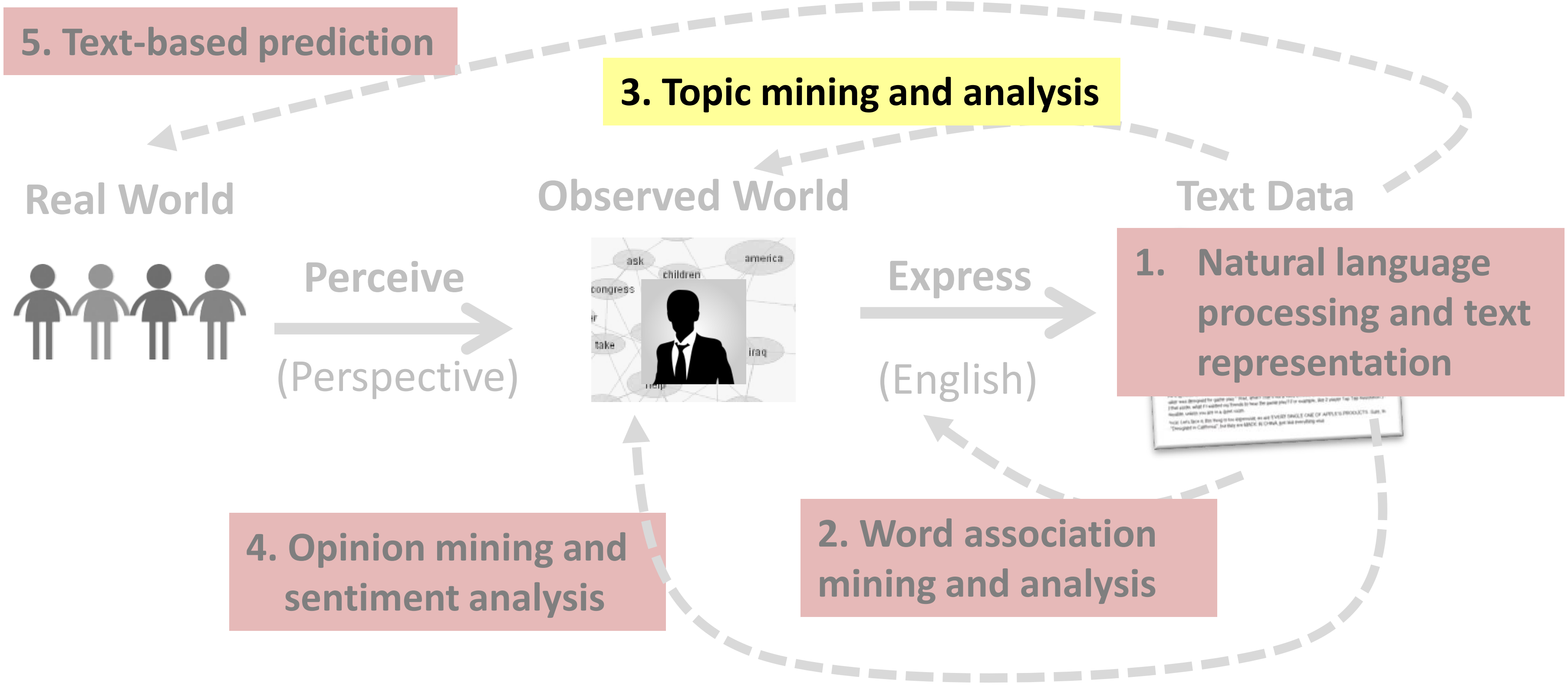
Text Clustering: Generative Probabilistic Models

Part 2

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

Text Clustering: Generative Probabilistic Models (Part 2)



5. Text-based prediction

3. Topic mining and analysis

Real World

Perceive

(Perspective)

Observed World

Express

(English)

Text Data

1. Natural language processing and text representation

4. Opinion mining and sentiment analysis

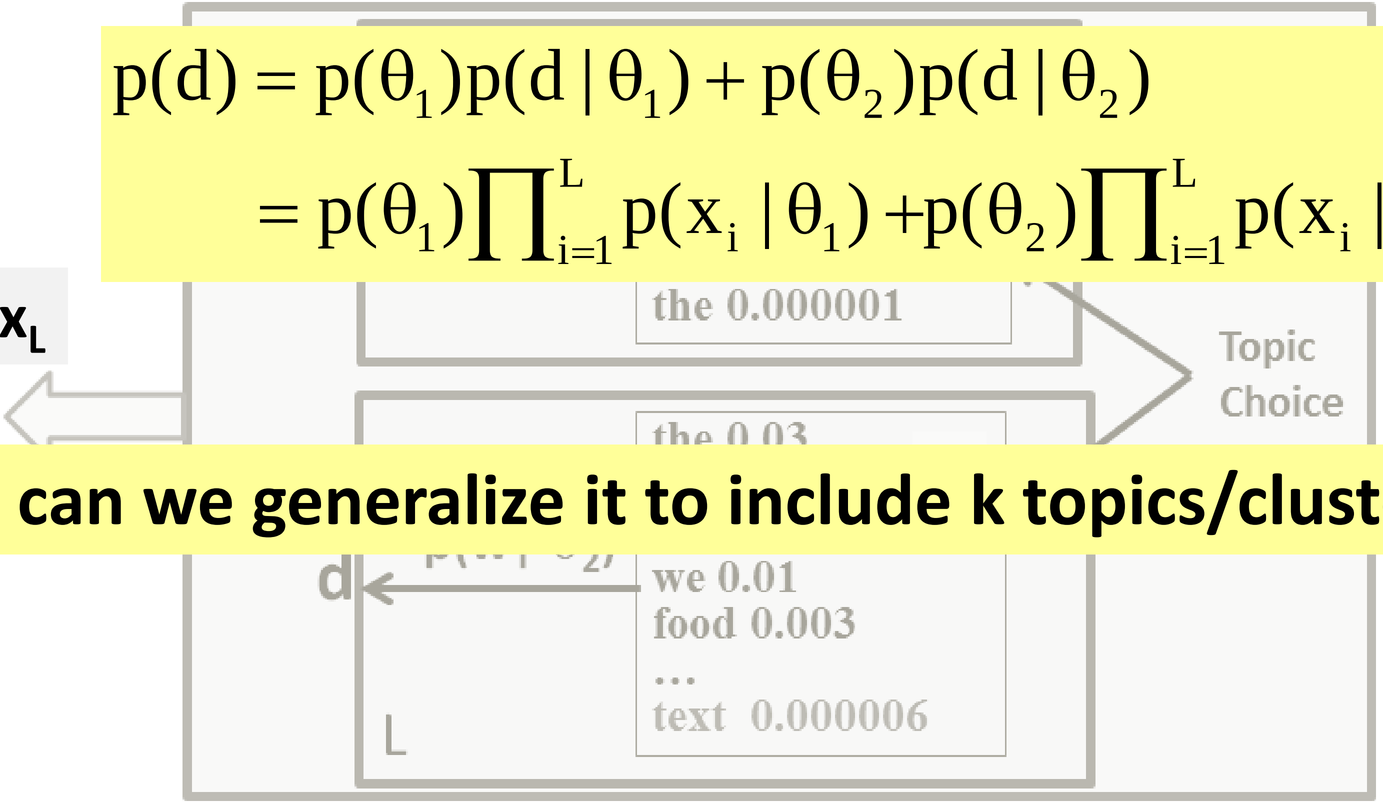
2. Word association mining and analysis

Likelihood Function: p(d)=?

$$p(d) = p(\theta_1)p(d\,|\,\theta_1) + p(\theta_2)p(d\,|\,\theta_2)$$

$$= p(\theta_1)\prod_{i=1}^{L} p(x_i\,|\,\theta_1) + p(\theta_2)\prod_{i=1}^{L} p(x_i\,|\,\theta_2)$$

$d=x_1\ x_2 \ldots x_L$



How can we generalize it to include k topics/clusters?

Mixture Model for Document Clustering

- Data: a collection of documents C={d_1, …, d_N} 在k个分布中选择生成文档的概率
- Model: mixture of k unigram LMs: $\Lambda=(\{\theta_i\}; \{p(\theta_i)\})$, $i\in[1,k]$
  - To generate a document, first **choose a** θ_i according to $p(\theta_i)$, and then generate **all** words in the document using $p(w|\theta_i)$

先选择θi，
再则用θi生成

- Likelihood:

$$p(d\mid\Lambda)=\sum_{i=1}^{k}[p(\theta_i)\prod_{j=1}^{|d|}p(x_j|\theta_i)]$$

$$=\sum_{i=1}^{k}[p(\theta_i)\prod_{w\in V}p(w|\theta_i)^{c(w,d)}]$$

- Maximum Likelihood estimate

用最大似然估计

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

Cluster Allocation After Parameter Estimation

- **Parameters** of the mixture model: $\Lambda=(\{\theta_i\}; \{p(\theta_i)\}), i\in[1,k]$
  - Each θ_i represents the **content of cluster i** : $p(w| \theta_i)$
  - $p(\theta_i)$ indicates the **size of cluster i** （该类用于生成文档的可能性，可以转换成聚类i的大小）
  - Note that unlike in PLSA, $p(\theta_i)$ doesn't depend on d!
- Which cluster should document d belong to? c_d=? $c_d \in [1,k]$
  - **Likelihood only**: Assign d to the cluster corresponding to the topic θ_i that most likely has been used to generate d

$$c_d = \arg\max_i p(d\,|\,\theta_i)$$

  - **Likelihood + prior $p(\theta_i)$ (Bayesian):** favor large clusters 考虑了聚类的大小..

$$c_d = \arg\max_i p(d\,|\,\theta_i)p(\theta_i)$$

大聚类生成d的可能性更大.