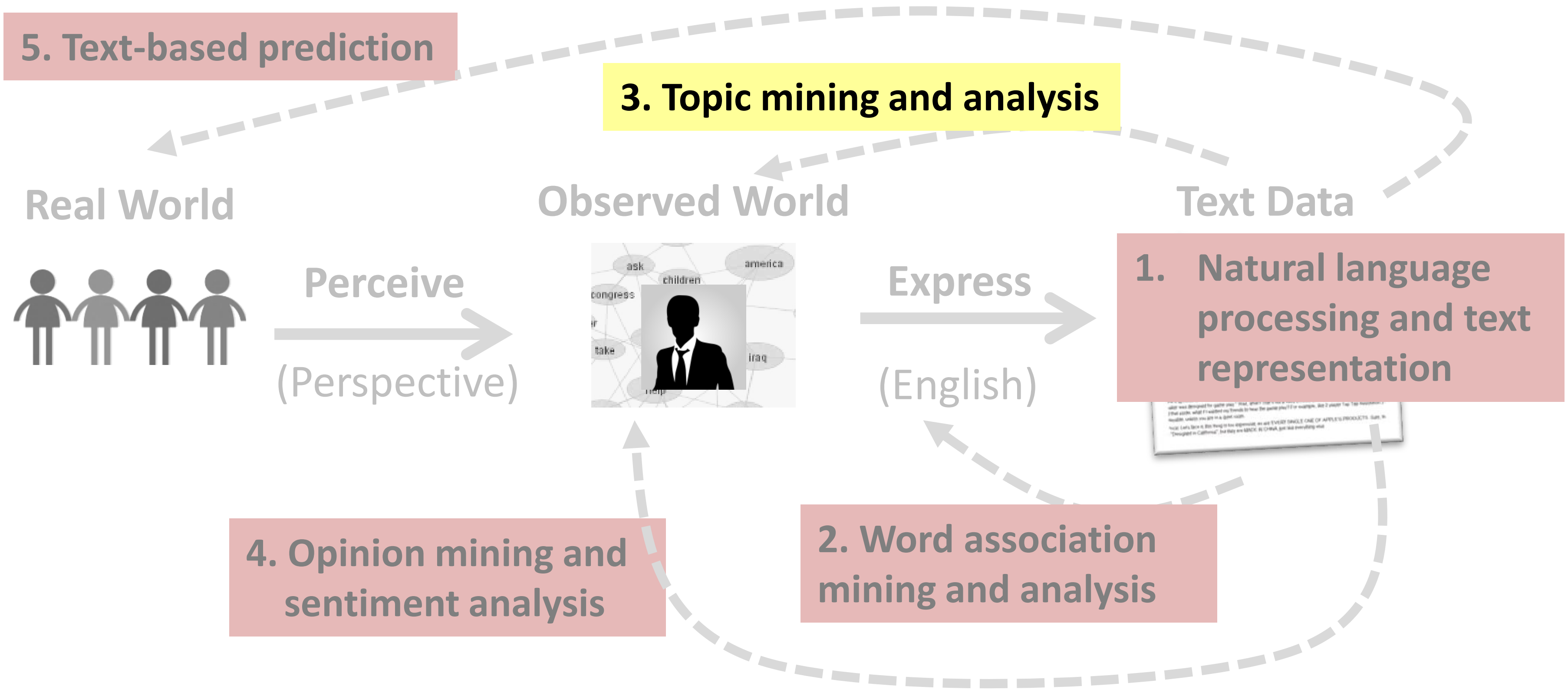
Topic Mining and Analysis: Mining One Topic

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

Probabilistic Topic Models: Mining One Topic



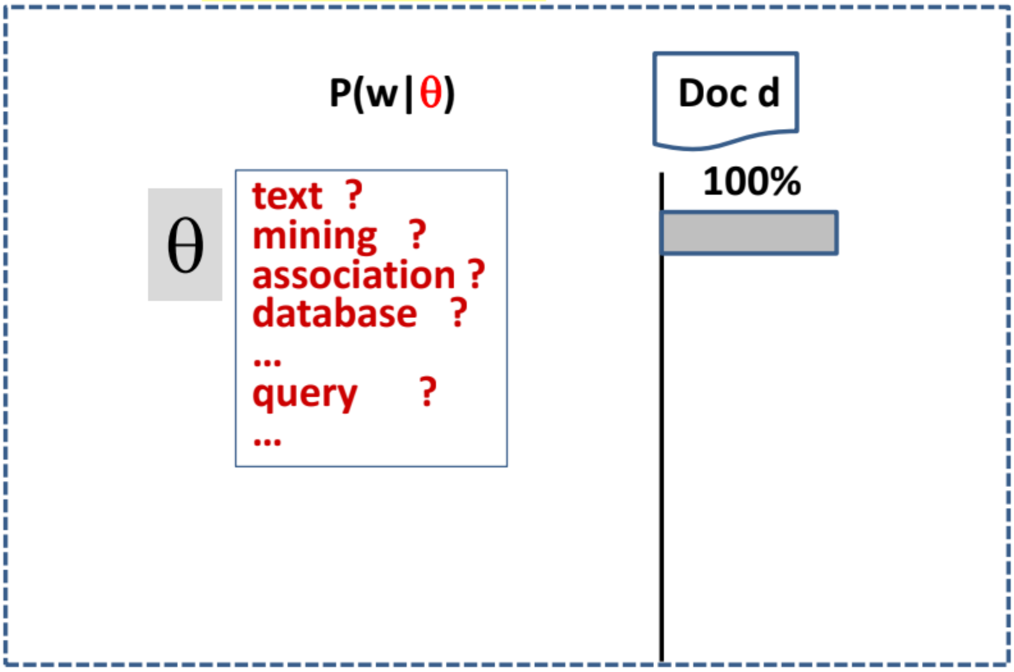
Simplest Case of Topic Model: Mining One Topic

INPUT: C={d}, V

K=1

Text Data

OUTPUT: { θ}



Language Model Setup

- **Data**: Document d= $x_1\ x_2 \ldots x_{|d|}$, $x_i \in V=\{w_1, \ldots, w_M\}$ is a word

- **Model**: Unigram LM θ(=topic) : $\{\theta_i=p(w_i|\theta)\}$, i=1, …, M; $\theta_1+\ldots+\theta_M=1$ θ_i 表示 i 在 d 中的概率.

- **Likelihood** function: $p(d|\theta) = p(x_1|\theta) \times \ldots \times p(x_{|d|}|\theta)$

$$\boxed{= p(w_1|\theta)^{c(w_1,d)} \times \ldots \times p(w_M|\theta)^{c(w_M,d)}}$$

w_M 在 d 中出现的次数.

有几个出现就
有几个相乘

$$= \prod_{i=1}^{M} p(w_i|\theta)^{c(w_i,d)} = \prod_{i=1}^{M} \theta_i^{c(w_i,d)}$$

$x^n = x \cdot \cdot x \cdots \cdot x$

$x^0 = 1$

将参数作为挖掘后的知识

目标, 找到一组 θ_i. 使 $p(d|\theta)$ 最大.

- ML **estimate:** $(\hat{\theta}_1, \ldots, \hat{\theta}_M) = \arg\max_{\theta_1,\ldots,\theta_M} p(d|\theta) = \arg\max_{\theta_1,\ldots,\theta_M} \prod_{i=1}^{M} \theta_i^{c(w_i,d)}$

Computation of Maximum Likelihood Estimate

Maximize p(d|θ) $(\hat{\theta}_1,\ldots,\hat{\theta}_M) = \arg\max_{\theta_1,\ldots,\theta_M} p(d\,|\,\theta) = \arg\max_{\theta_1,\ldots,\theta_M} \prod_{i=1}^{M} \theta_i^{c(w_i,d)}$

Max. Log-Likelihood $(\hat{\theta}_1,\ldots,\hat{\theta}_M) = \arg\max_{\theta_1,\ldots,\theta_M} \log[p(d\,|\,\theta)] = \arg\max_{\theta_1,\ldots,\theta_M} \boxed{\sum_{i=1}^{M} c(w_i,d)\log\theta_i}$

取log更易计算.

Subject to constraint: $\sum_{i=1}^{M}\theta_i = 1$

Use Lagrange multiplier approach

拉格朗日乘数法

Lagrange function: $f(q\,|\,d) = \sum_{i=1}^{M} c(w_i,d)\log q_i + \lambda(\sum_{i=1}^{M} q_i - 1)$

Normalized Counts

求偏导

$$\frac{\partial f(q\,|\,d)}{\partial q_i} = \frac{c(w_i,d)}{q_i} + \lambda = 0 \rightarrow \boxed{q_i = -\frac{c(w_i,d)}{\lambda}} \quad \because \sum_{i=1}^{M} q_i = 1$$

代入

$$\therefore \sum_{i=1}^{M} -\frac{c(w_i,d)}{\lambda} = 1 \rightarrow \boxed{\lambda = -\sum_{i=1}^{N} c(w_i,d)} \rightarrow \hat{q}_i = p(w_i\,|\,\hat{q}) = \frac{c(w_i,d)}{\sum_{i=1}^{M} c(w_i,d)} = \frac{c(w_i,d)}{|d|}$$

What Does the Topic Look Like?

$p(w|\theta)$

d

Text mining paper

the 0.031
a 0.018
…
text 0.04
mining 0.035
association 0.03
clustering 0.005
computer 0.0009
…
food 0.000001
…

Can we get rid of these common words?