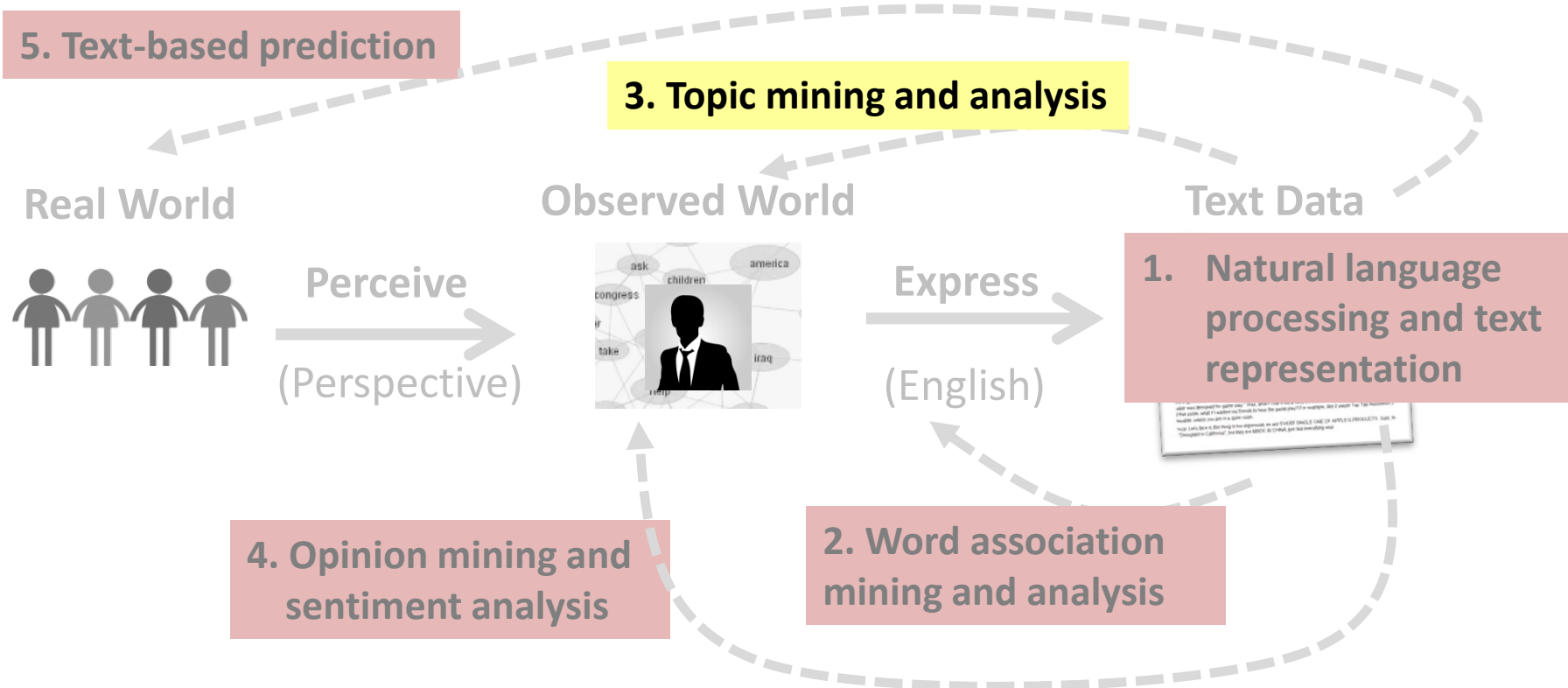




Topic Mining and Analysis: Mining One Topic

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

... we are not going to give up on our vision for a new, more natural and expressive way of talking to things. I am not considering taking back. Here's why.

Speaker quality is ABSOLUTELY HORRENDOUS. The speaker is simply not functional, unless you are in perfect recording studio. When you turn it up all the way because you can't hear it it becomes fuzzy, sounding like a blow dryer. What is this thing, the size of a hair? And if there is ANY background noise you cannot hear it at all with the volume up or even you're wondering, no, the speaker did not fail (recharged the iPad, and the second one started to make really bad). I can't share videos, share music, share anything, unless my friends and I are in a wine cellar.

I heard reviews for the speaker's embarrassing quality, including "It's a small device". My cheap Samsung cell phone's speaker is 4x better than the iPod Touch, and the phone is smaller. The other reason I've heard is "The speaker was designed for game play." What, what? That's not a valid excuse, because the iPod is a MUSIC device, of that aside, what if I wanted my friends to hear the game play? For example, the 2 player "Tap Tap Revolution 3" playable, unless you are in a quiet room.

Price. Let's face it, this thing is too expensive, as are EVERY SINGLE ONE OF APPLE'S PRODUCTS. Sure, the "Designed in California", but they are MADE IN CHINA, just like everything else.

More from iDAYS



OUTPUT: $\{\theta\}$

$P(w|\theta)$

θ

text ?
mining ?
association ?
database ?

...
query ?
...

Doc d

100%

Language Model Setup

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

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

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

$$\begin{aligned} &= p(w_1 | \theta)^{c(w_1, d)} \times \dots \times p(w_M | \theta)^{c(w_M, d)} \\ &= \prod_{i=1}^M p(w_i | \theta)^{c(w_i, d)} = \prod_{i=1}^M \theta_i^{c(w_i, d)} \end{aligned}$$

将参数作为数据的知识

目标: 找到一组 θ_i 使 $p(d | \theta)$ 最大.

*w_M 在 doc 中出现次数.
 有可能出现多次
 有可能不出现
 $x^n = x \cdot x \cdot \dots \cdot x$
 $x^0 = 1$*

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

Computation of Maximum Likelihood Estimate

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

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

使用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

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

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

What Does the Topic Look Like?

d

Text mining
paper

$p(w | \theta)$

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?