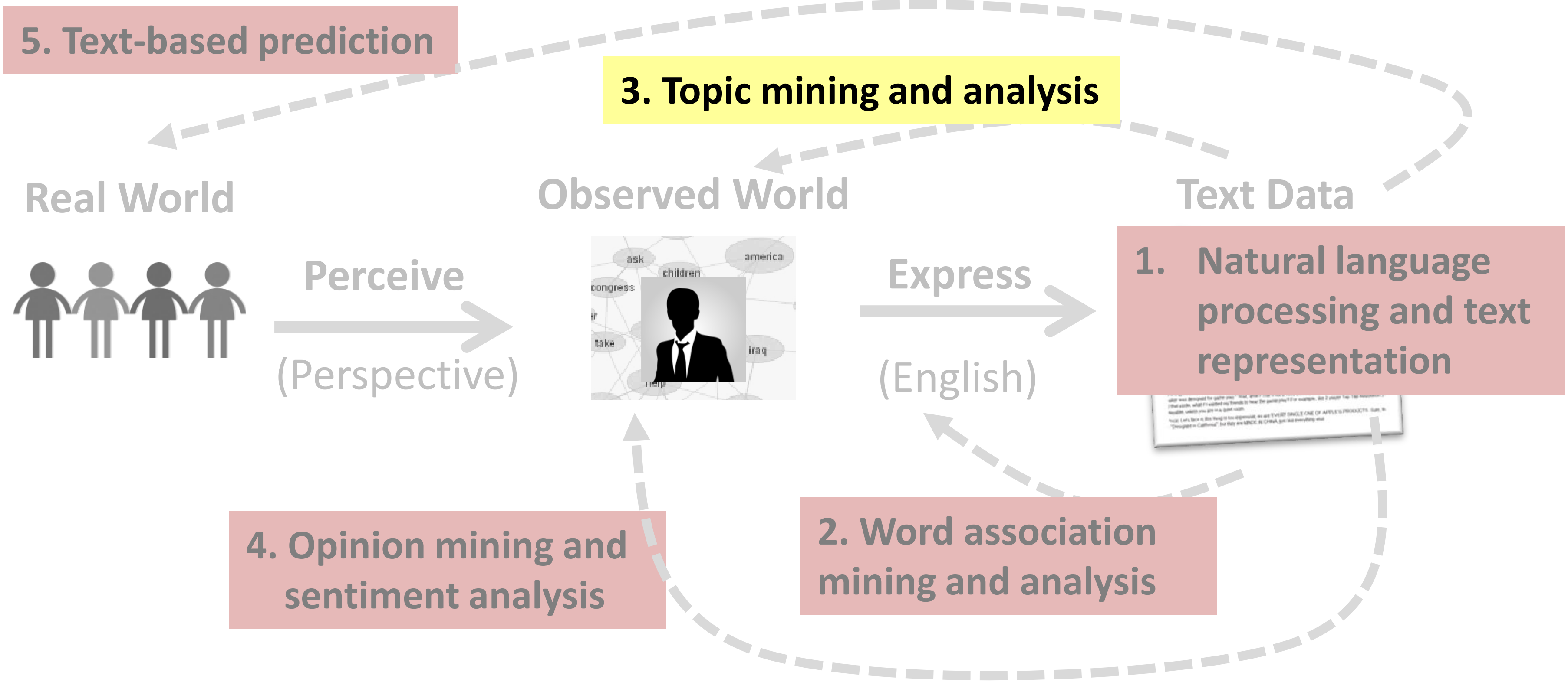
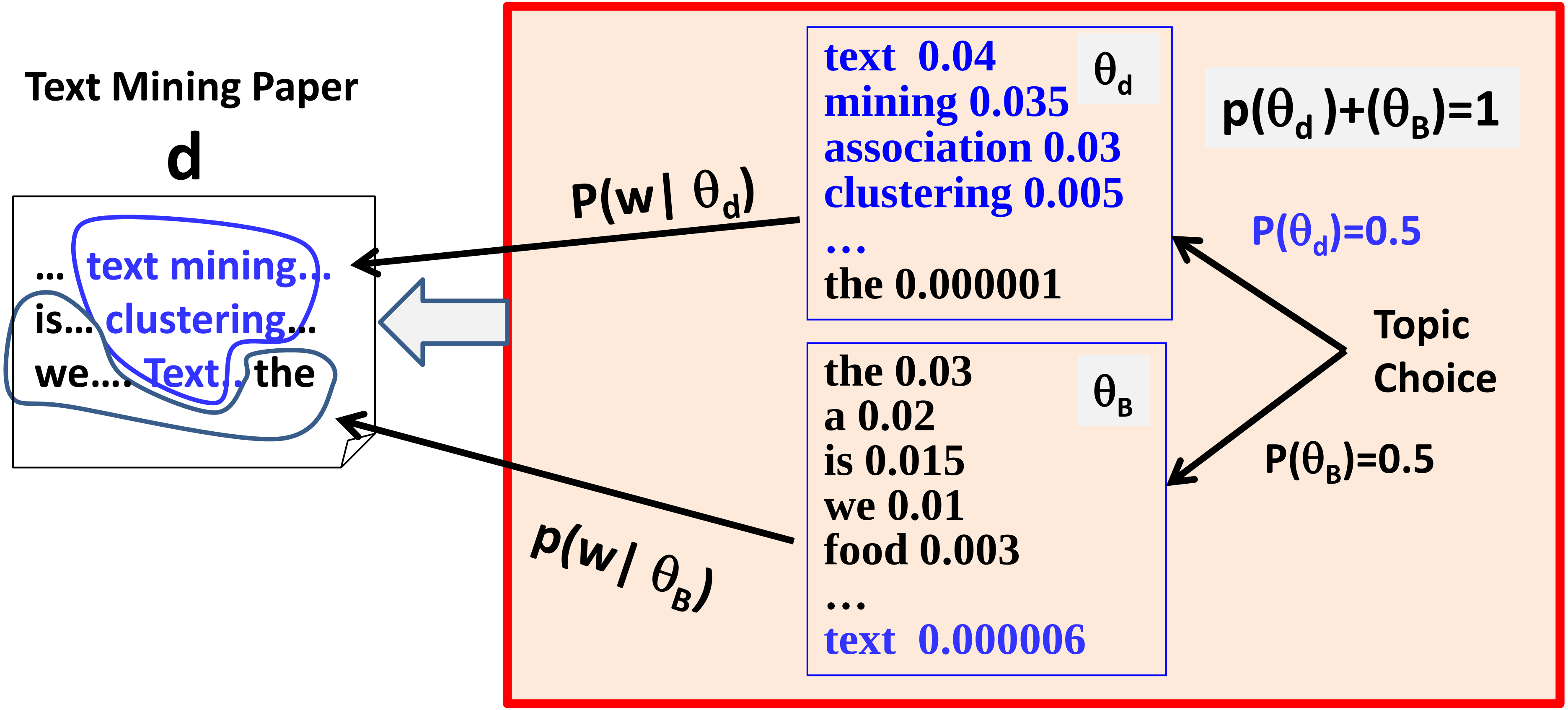
Probabilistic Topic Models: Mixture Model Estimation

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

Probabilistic Topic Models: Mixture Model Estimation



Back to Factoring out Background Words



Estimation of One Topic: **$P(w|\theta_d)$**

Adjust θ_d to maximize $p(d|\Lambda)$
(all other parameters are known)

Would the ML estimate <u>demote</u> background words in θ_d ?

d

... text mining...
is... clustering...
we.... Text.. the

θ_d

text ?
mining ?
association ?
clustering ?
…
the ?

$p(\theta_d)+(\theta_B)=1$

$P(\theta_d)=0.5$

Topic Choice

$P(\theta_B)=0.5$

θ_B

the 0.03
a 0.02
is 0.015
we 0.01
food 0.003
…
text 0.000006

Behavior of a Mixture Model

d = text the

Likelihood:

$$P(\text{“text”}) = p(\theta_d)p(\text{“text”}|\theta_d) + p(\theta_B)p(\text{“text”}|\theta_B)$$
$$= 0.5 * p(\text{“text”}|\theta_d) + 0.5*0.1$$

$$P(\text{“the”}) = 0.5 * p(\text{“the”}|\theta_d) + 0.5*0.9$$

$$p(d|\Lambda) = p(\text{“text”}|\Lambda)\, p(\text{“the”}|\Lambda)$$
$$= [0.5 * p(\text{“text”}|\theta_d) + 0.5*0.1] \times$$
$$[0.5 * p(\text{“the”}|\theta_d) + 0.5*0.9]$$

text ?	θ_d
the ?	

$P(\theta_d)=0.5$

$P(\theta_B)=0.5$

the 0.9	θ_B
text 0.1	

How can we set $p(\text{“text”}|\theta_d)$ & $p(\text{“text”}|\theta_d)$ to maximize it?

Note that $p(\text{“text”}|\theta_d) + p(\text{“the”}|\theta_d) = 1$

“Collaboration” and “Competition” of θ_d and θ_B

$p(d|\Lambda) = p(\text{“text”}|\Lambda)\, p(\text{“the”}|\Lambda)$

$= [0.5 * p(\text{“text”}|\theta_d) + 0.5*0.1] \times$

$[0.5 * p(\text{“the”}|\theta_d) + 0.5*0.9]$

d = text the

text ?	θ_d
the ?	

$P(\theta_d)=0.5$

$P(\theta_B)=0.5$

the 0.9	θ_B
text 0.1	

Note that $p(\text{“text”}|\theta_d) + p(\text{“the”}|\theta_d) = 1$

If $x + y = constant$, then xy reaches maximum when $x = y$.

$0.5 * p(\text{“text”}|\theta_d) + 0.5*0.1 = 0.5 * p(\text{“the”}|\theta_d) + 0.5*0.9$

→ $p(\text{“text”}|\theta_d)=0.9 \quad >> \quad p(\text{“the”}|\theta_d) = 0.1$!

Behavior 1: if $p(w1|\theta_B) > p(w2|\theta_B)$, then $p(w1|\theta_d) < p(w2|\theta_d)$

Response to Data Frequency

d = text the

$$p(d|\Lambda) = [0.5*p(\text{"text"}|\theta_d) + 0.5*0.1] \times [0.5*p(\text{"the"}|\theta_d) + 0.5*0.9]$$

$\rightarrow p(\text{"text"}|\theta_d)=0.9 \quad >> \quad p(\text{"the"}|\theta_d) = 0.1$!

d' = text the the the the ...the

$$p(d'|\Lambda) = [0.5*p(\text{"text"}|\theta_d) + 0.5*0.1] \times [0.5*p(\text{"the"}|\theta_d) + 0.5*0.9] \times [0.5*p(\text{"the"}|\theta_d) + 0.5*0.9] \times [0.5*p(\text{"the"}|\theta_d) + 0.5*0.9] \cdots \times [0.5*p(\text{"the"}|\theta_d) + 0.5*0.9]$$

What if we increase $p(\theta_B)$?

What's the optimal solution now? $p(\text{"the"}|\theta_d) > 0.1$? or $p(\text{"the"}|\theta_d) < 0.1$?

Behavior 2: high frequency words get higher $p(w|\theta_d)$

Summary

- General behavior of a mixture model:
  - Every component model attempts to assign high probabilities to highly frequent words in the data (to "collaboratively maximize likelihood")
  - Different component models tend to "bet" high probabilities on different words (to avoid "competition" or "waste of probability")
  - The probability of choosing each component "regulates" the collaboration/competition between the component models
- Fixing one component to a background word distribution (i.e., background language model):
  - Helps "get rid of background words" in other component
  - Is an example of imposing a prior on the model parameters (prior = one model must be exactly the same as the background LM)