Probabilistic Topic Model

Jie Tang

Department of Computer Science and Technology
Tsinghua University

Why Topic—Recall Our Input

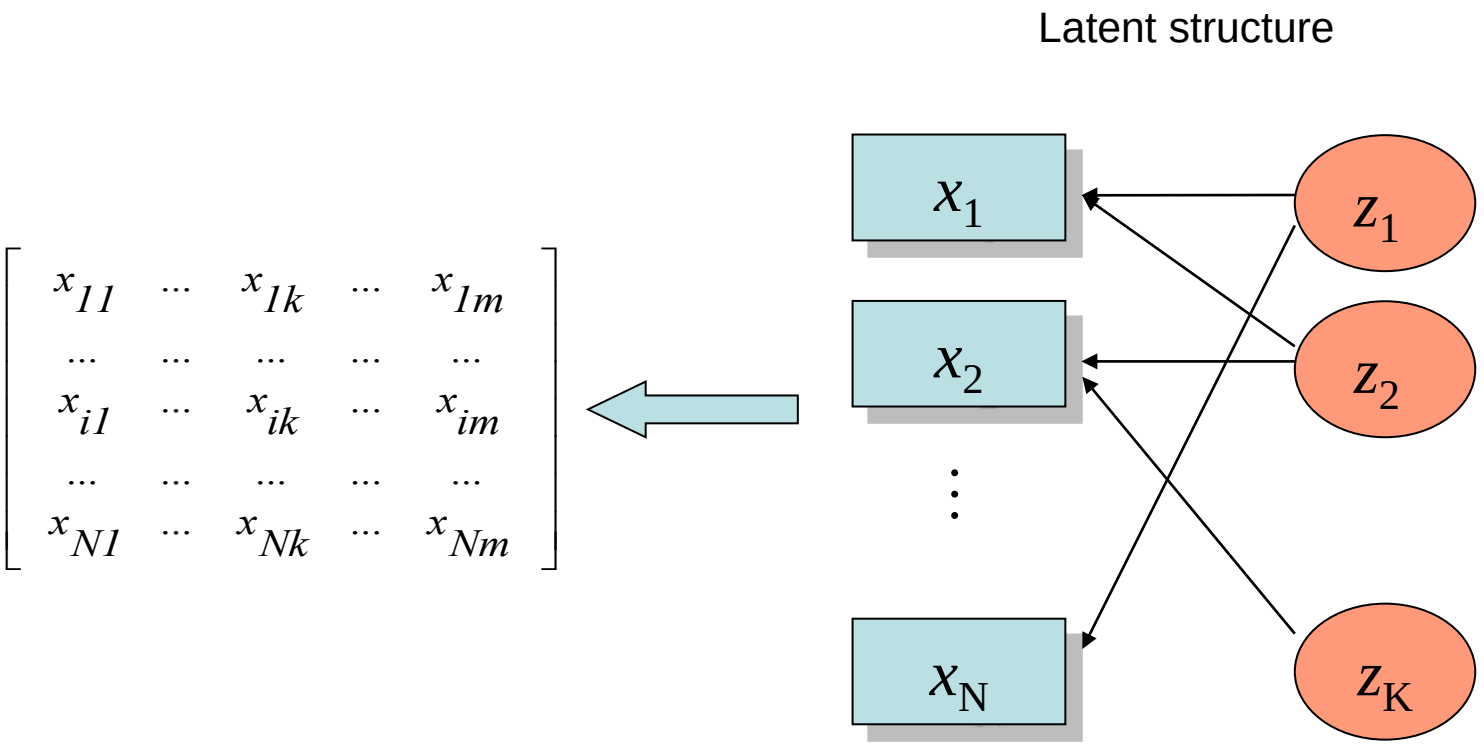
- An example with p features

$$\boldsymbol{x}_i=[x_{i1}, x_{i2},\ldots, x_{im}]$$

- Data (set) matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1k} & \cdots & x_{1m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{ik} & \cdots & x_{im} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{N1} & \cdots & x_{Nk} & \cdots & x_{Nm} \end{bmatrix}$$

Extend Our Input



Why Topic? (cont.)

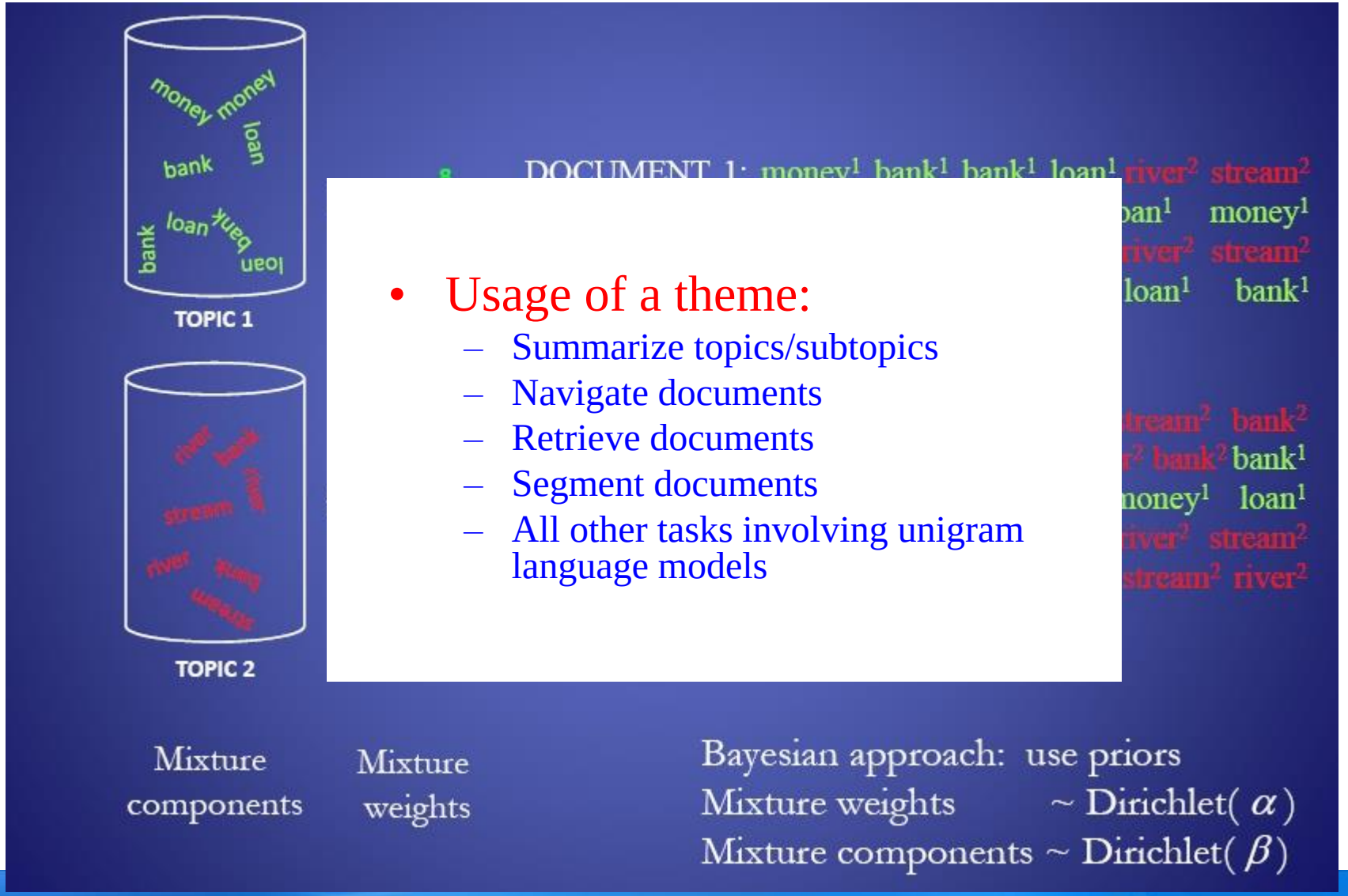
- IR – text to real number vector (Baeza-Yates and Ribeiro-Neto, 1999), tfidf (Salton and McGill, 1983)
  - Tfidf – shortcoming: (1) Lengthy and (2) Cannot model inter- and intra- document statistical structure
- LSI – dimension reduction (Deerwester et al., 1990)
  - Advantages: achieve significant compression in large collections and capture synonymy and polysemy.
- Generative probabilistic model – to study the ability of LSI (Papadimitriou et al., 1998)
  - Why LSI, we can model the data directly using maximum likelihood or Bayesian methods.

What is Topic?

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

What is Topic? (cont.)

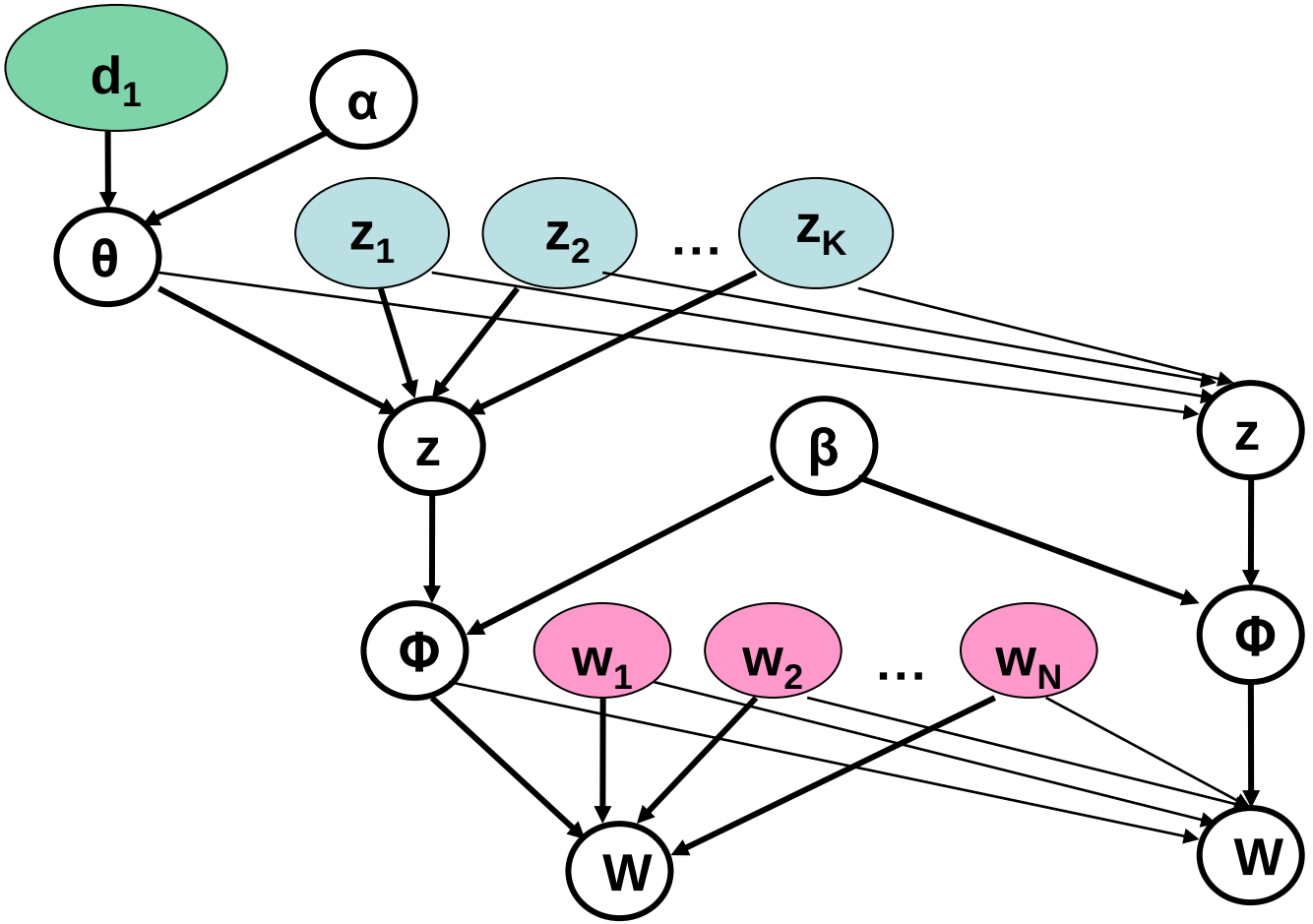


- Usage of a theme:
  - Summarize topics/subtopics
  - Navigate documents
  - Retrieve documents
  - Segment documents
  - All other tasks involving unigram language models

Topic Model Family Member

- Directed graphical model (Bayesian network model)
- Undirected graphical model (Markov random field, restricted boltzmann machines)
- We will only focus on directed graphical model in this talk.

Bayesian Network for modeling document generation



Two state-of-the-art topic models:

- PLSI
- LDA

Notations

- M: #documents in the whole corpus
- K: #topics
- V: #vocabulary

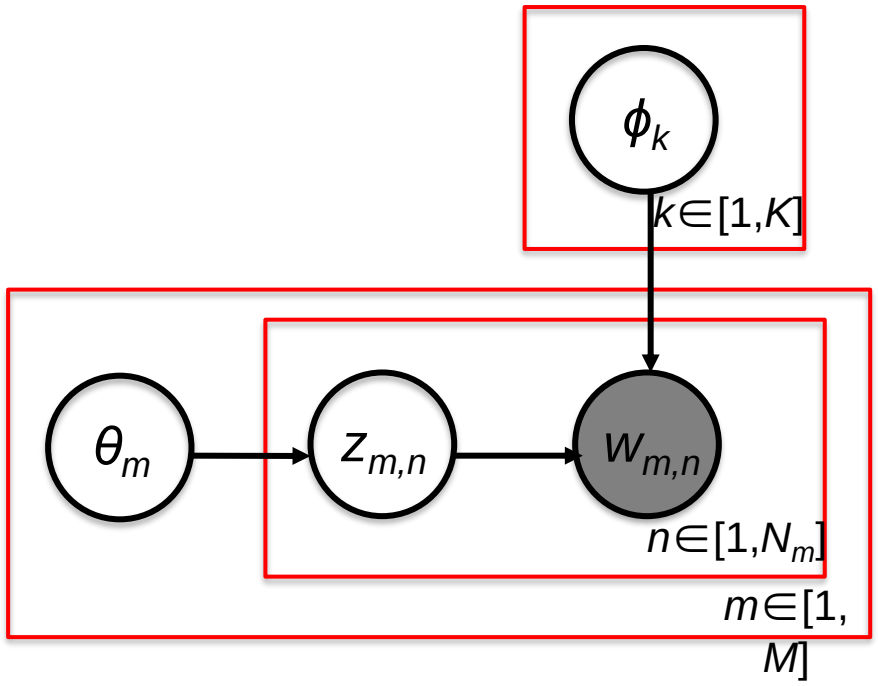
- **w, z** : bold font means words or topics in the whole corpus
- w : a word, z : a topic, d : a document

- m: the index of document, m = 1 to M
- N_m: #words in m^{th} document
- n: the index of word in a document, n=1 to N_m
  - $w_{m,n}$: the n^{th} word of m^{th} document
  - $z_{m,n}$: the n^{th} word of m^{th} document is denoted by topic $z_{m,n}$

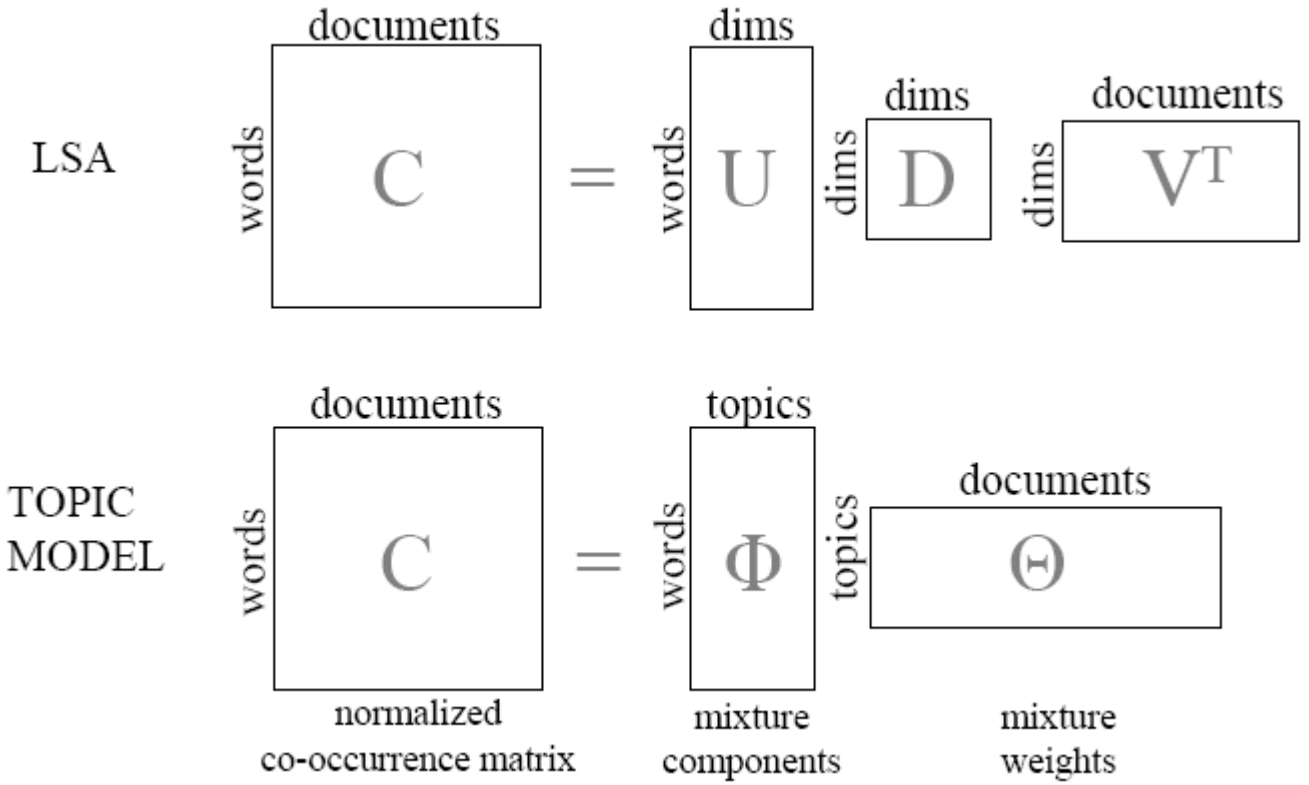
- k: the index of topic, k=1 to K
- t: the index of word in the vocabulary, t=1 to V
- $n_{.,k,t}$: the number of times word w_t is generated by topic z_k
- $n_{m,k,.}$: the number of times topic z_k occurs in document d_m

Probabilistic Latent Semantic Indexing (PLSI)

PLSI Model

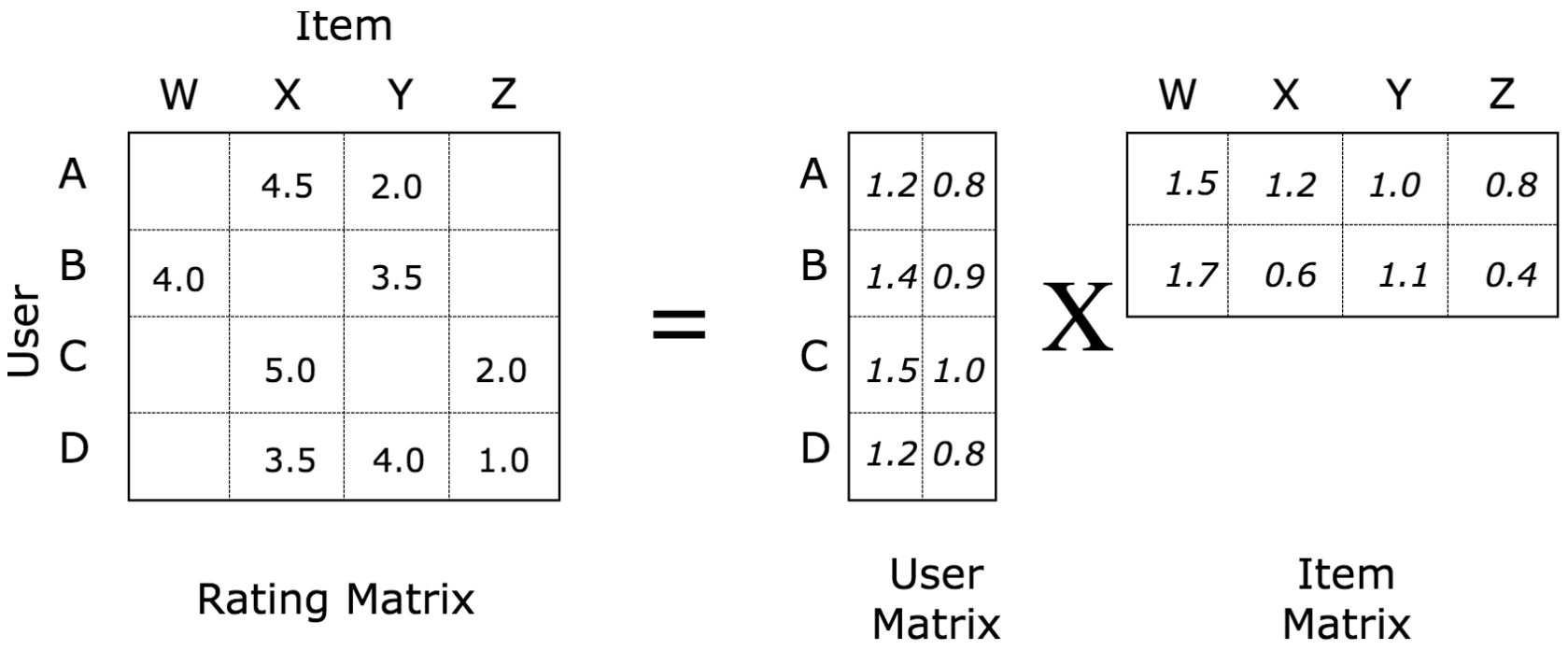


Matrix Interpretation



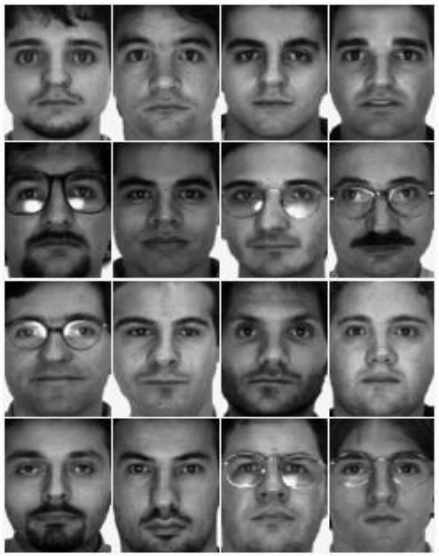
- In the topic model, the word-document co-occurrence matrix is split into two parts: a topic matrix Φ and a document matrix Θ. Note that the diagonal matrix D in LSA can be absorbed in the matrix U or V, making the similarity between the two representations even clearer.

Example 1: products

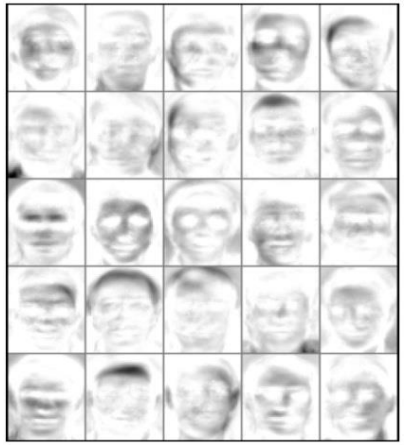


- Benefits:
  - Discovering “latent profile” in input matrix
  - Recommending new products to customers
  - More

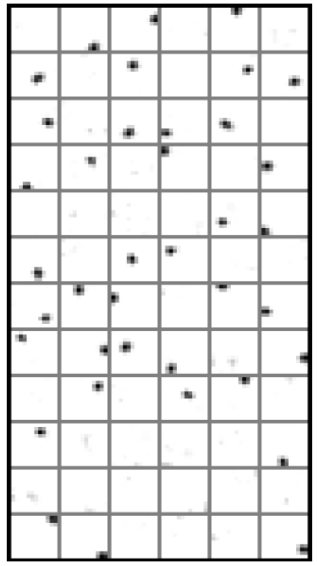
Example 2: images



$\approx$



$\times$



$\mathbf{X}$

input images

$\mathbf{U}$

face patterns

$\mathbf{V}^{\mathrm{T}}$

images represented with patterns

Example 3: documents

Encyclopedia entry: 'Constitution of the United States'

president (148)
congress (124)
power (120)
united (104)
constitution (81)
amendment (71)
government (57)
law (49)

$\approx$

court government council culture supreme constitutional rights justice	president served governor secretary senate congress presidential elected
flowers leaves plant perennial flower plants growing annual	disease behaviour glands contact symptoms skin pain infection

$\times$



$\mathbf{X}$

input documents

$\mathbf{U}$

latent topics

$\mathbf{V}^{\mathrm{T}}$

docs represented with topics

Example 4: videos

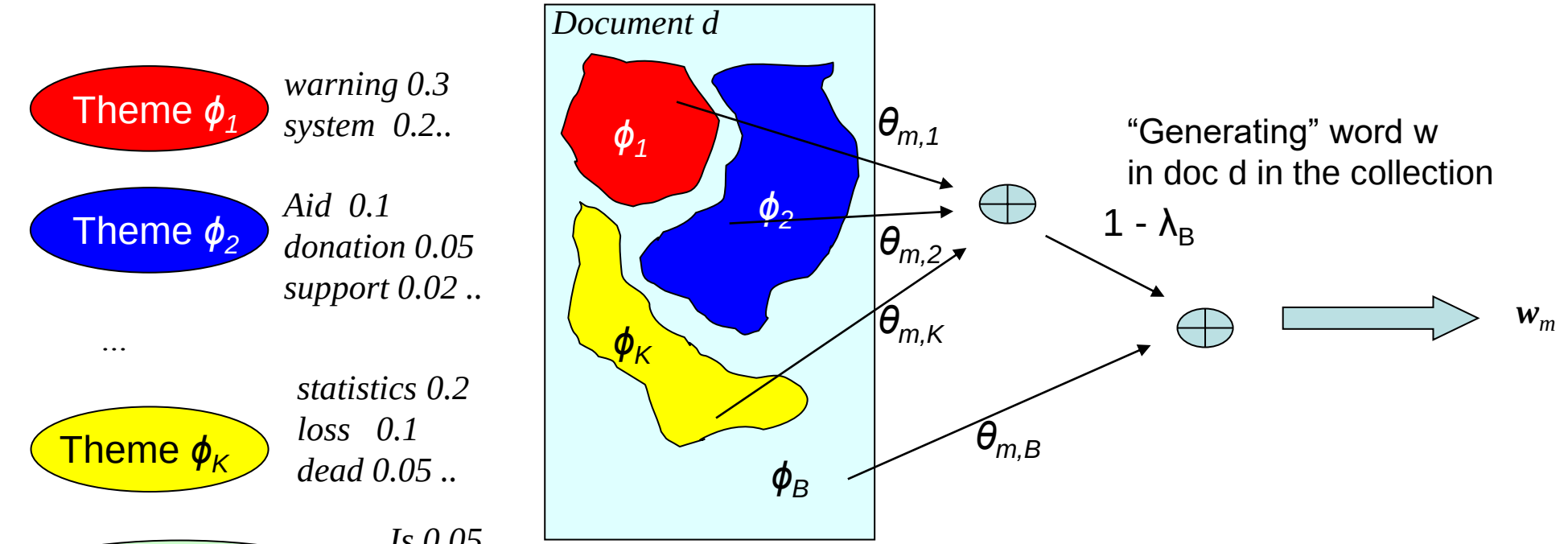
Video D = Low-rank appx. A + Sparse error E



PLSI as a Mixture Model

$$p(w_{m,n}) = \lambda_B p(w_{m,n} \mid \phi_B) + (1-\lambda_B)\sum_{k=1}^{K} \theta_{m,k} p(w_{m,n} \mid \phi_k)$$

$$\log p(\mathbf{w}_m) = \sum_{n=1}^{N_m} \log[\lambda_B p(w_{m,n} \mid \phi_B) + (1-\lambda_B)\sum_{k=1}^{K} \theta_{m,k} p(w_{m,n} \mid \phi_k)]$$



Parameters:

λ_B=noise-level (manually set)

θ's and ϕ's are estimated with Maximum Likelihood

Formal Description

Maximum-likelihood

- Definition
  - We have a density function $P(x|\Theta)$ that is govened by the set of parameters Θ, e.g., P might be a set of Gaussians and Θ could be the means and covariances
  - We also have a data set $X=\{x_1,\ldots,x_N\}$, supposedly drawn from this distribution P, and *a*ssume these data vectors are i.i.d. with P.
  - Then the log-likehihood function is:

$$L(\Theta \mid X) = \log p(X \mid \Theta) = \log \prod_i p(x_i \mid \Theta) = \sum_i \log p(x_i \mid \Theta)$$

  - The log-likelihood is thought of as a function of the parameters Θ where the data X is fixed. Our goal is to find the Θ that maximizes L. That is

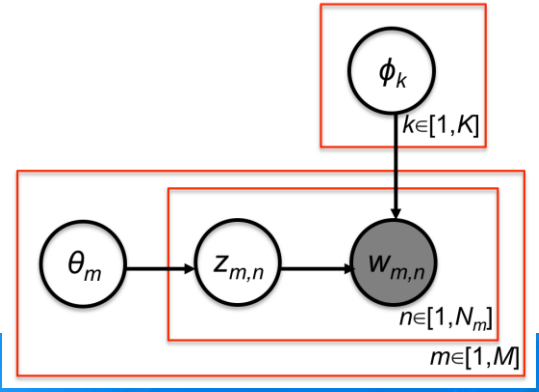
$$\Theta^* = \arg\max_{\Theta} L(\Theta \mid X)$$

PLSI Model

- Following the likelihood principle, we determines ϕ and θ by maximizing of the log-likelihood function

$$\begin{aligned}
\mathcal{L}(\Phi,\Theta) &= \sum_{m=1}^{M}\sum_{n=1}^{N} n(d_m,w_{m,n})\log P(d_m,w_{m,n}) \\
&= \sum_{m=1}^{M} n(d_m)\left(\log P(d_m)+\sum_{n=1}^{N_m}\frac{n(d_m,w_{m,n})}{n(d_m)}\log\sum_{k=1}^{K}P(w_{m,n}|z_k)P(z_k|d_m)\right) \\
&= \sum_{m=1}^{M} n(d_m)\left(\log P(d_m)+\sum_{n=1}^{N_m}\frac{n(d_m,w_{m,n})}{n(d_m)}\log\sum_{k=1}^{K}\phi_{k,w_{m,n}}\theta_{m,k}\right)
\end{aligned}$$

where $\Phi=[\phi_1,\cdots,\phi_K], z_k\in\mathcal{Z}$
and $\Theta=[\theta_m,\cdots,\theta_M], d_m\in\mathcal{D}$



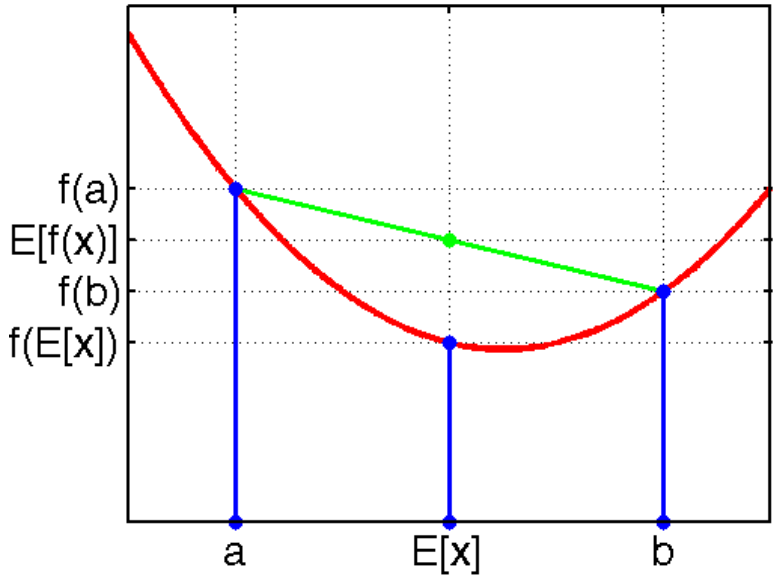
Recall Jensen Inequality

Andrew Ng. Part IX: The EM algorithm. CS229 Lecture Notes.

Jensen's Inequality

- Recall that *f* is a convex function if *f* "(x)≥0, and *f* is strictly convex function if *f* "(x)>0
- Let *f* be a convex function, and let *X* be a random variable, then:

$$E[f(X)] \geq f(EX)$$



- Moreover, if *f* is strictly convex, then $E[f(X)]=f(EX)$ holds true if and only if $X=E[X]$ with probability 1 (i.e., if *X* is a constant)

Basic EM Algorithm

- However, Optimizing the likelihood function is analytically intractable but when the likelihood function can be simplified by assuming the existence of and values for additional but missing (or hidden) parameters:

$$L(\Theta \mid X) = \sum_i \log p(x_i \mid \Theta) = \sum_i \log \sum_z p(x_i, z \mid \Theta)$$

- Maximizing L(Θ) explicitly might be difficult, and the strategy is to instead repeatedly construct a lower-bound on L(E-step), and then optimize that lower bound (M-step).
  - For each i, let Q_i be some distribution over z ($\sum_z Q_i(z)=1$, $Q_i(z)\geq 0$), then

$$\sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \Theta) = \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}; \Theta)}{Q_i(z^{(i)})} \geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \Theta)}{Q_i(z^{(i)})}$$

  - The above derivation used Jensen's inequality. Specifically, f(x) = logx is a concave function, since f"(x)=$-1/x^2<0$

$$E[f(X)] \geq f(EX)$$

Basic EM Algorithm (cont')

- In the summation is just an expectation of the quantity $[p(x^{(i)},z^{(i)}|\Theta)/Q_i(z^{(i)})]$ with respect to $z^{(i)}$ drawn according to the distribution given by Q_i. By Jensen's inequality, we have

$$f(E_{z^{(i)}\sim Q_i}[\frac{p(x^{(i)},z^{(i)};\Theta)}{Q_i(z^{(i)})}]) \geq E_{z^{(i)}\sim Q_i}[f(\frac{p(x^{(i)},z^{(i)};\Theta)}{Q_i(z^{(i)})})]$$

  - Now, for any set of distributions Q_i, we gives a lower-bound on $L(\Theta)$.
- There are many possible choices for the Q_i. Which should we choose?
  - Have current guess Θ, try to make the lower-bound tight at the value of Θ.
  - To make the bound tight for a particular value of Θ, we need Jensen's inequality to hold with equality. we require that X is a constant, i.e.,

$$\frac{p(x^{(i)},z^{(i)};\Theta)}{Q_i(z^{(i)})} = c$$

  - Since we know that $\sum_z Q_i(z)=1$, this further tells us that

$$Q_i(z^{(i)}) = \frac{p(x^{(i)},z^{(i)};\Theta)}{\sum_z p(x^{(i)},z;\Theta)} = \frac{p(x^{(i)},z^{(i)};\Theta)}{p(x^{(i)};\Theta)} = p(z^{(i)} \mid x^{(i)};\Theta)$$

  - Thus, we simply set the Q_i's to be the posterior distribution of the $z^{(i)}$'s given $x^{(i)}$ and the setting of the parameters Θ

Basic EM Algorithm (cont')

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) = p(z^{(i)} \mid x^{(i)}; \Theta)$$

(M-step) Set

$$\Theta := \arg\max_{\Theta} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \Theta)}{Q_i(z^{(i)})}$$

}

Proof of Convergence

- Suppose $\Theta^{(t)}$ and $\Theta^{(t+1)}$ are the parameters from two successive iterations of EM. We will now prove that $L(\Theta^{(t)}) \leq L(\Theta^{(t+1)})$, which shows EM always monotonically improves the log-likelihood.

$$L(\Theta^{(t+1)}) \geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \Theta^{(t+1)})}{Q_i(z^{(i)})} \qquad (1)$$

$$\geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \Theta^{(t)})}{Q_i(z^{(i)})} \qquad (2)$$

$$= L(\Theta^{(t)})$$

  - Inequality (1) comes from the fact that

$$L(\Theta) \geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \Theta)}{Q_i(z^{(i)})}$$

    - Holds for any values of Q_i and Θ, and in particular holds for $Q_i = Q_i^{(t)}$, $\Theta = \Theta^{(t+1)}$.
  - Inequality (2) used the fact of M-step that $\Theta^{(t+1)}$ is chosen to maximize the lower bound.

Another interpretation

Our goal then is to find some q(z) trying to make the new function *L* to be the same as the original function *f*

$$f = p(X \mid \Theta) = \int_z \frac{p(X, z, \Theta)}{q(z)} q(z)dz \geq L = \prod_{i=1}^{N} \left(\frac{p(X, z, \Theta)}{q(z)}\right)$$

$$s.t. \qquad \int_z q(z) = 1$$

$$\Theta = \{P(w \mid z), p(z \mid d), p(z)\}$$

As q(z) is a valid probability distribution over h. Thus we have a logarithm form of the bound (Neal and Hinton, 1993)

$$\log\left(\prod_{i=1}^{N} \left(\frac{p(X, z, \Theta)}{q(z)}\right)^{q(z)}\right) = \int_z q(z) \log p(X, z, \Theta) - q(z) \log q(z)dz$$

Adding Lagrange multiplier for the constant q(z), we have

$$\log L = \lambda(1 - \int_z q(z)dz) + \int_z q(z) \log p(X, z, \Theta) - q(z) \log q(z)dz$$

$$\frac{dL}{dq(z)} = -\lambda - 1 + \log p(X, z, \Theta) - \log q(z) = 0$$

$$q(z) = \frac{p(X, z, \Theta)}{\int_z p(X, z, \Theta)dz} = p(z \mid X, \Theta)$$

When dL/dq(z)=0 the lower bound L equates to the original problem f

By KL-Divergence

$$L = \int_z q(z) \log p(X, z, \Theta) - q(z) \log q(z) dz$$

$$= \int_z q(z) \log \frac{p(X, z, \Theta)}{q(z)} dz$$

$$= E_{q(z)} \left[\log \frac{p(X, z, \Theta)}{q(z)} \right]$$

$$= -E_{q(z)} \left[\log \frac{q(z)}{p(z \mid X, \Theta)} \right] + \log p(X, \Theta)$$

$$= -D(q(z) \,||\, p(z \mid X, \Theta)) + \log p(X, \Theta)$$

Recall what is *X* in our problem.

Finding q to get a good bound is the E-step of the algorithm.
To get the next guess for θ, we maximize the bound over θ. this is the M-step. This step is problem-dependent The relevant term of G is

Using EM to derive PLSI

E-Step

- Given parameters, calculate the posterior probability of hidden variable.

$$
\begin{aligned}
q(z_k) &= P(z_k|d_m, w_t; \Theta, \Phi) \\
&= \frac{P(z_k, d_m, w_t)}{\sum_{l=1}^{K} P(z_l, d_m, w_t)} \\
&= \frac{P(w_t|d_m, z_k)P(z_k|d_m)P(d_m)}{\sum_{l=1}^{K} P(w_t|d_m, z_l)P(z_l|d_m)P(d_m)} \\
&= \frac{P(w_t|d_m, z_k)P(z_k|d_m)}{\sum_{l=1}^{K} P(w_t|d_m, z_l)P(z_l|d_m)} \\
&= \frac{\phi_{k,t}\theta_{m,k}}{\sum_{l=1}^{K} \phi_{l,t}\theta_{m,l}}
\end{aligned}
$$

Lower Bound

- Introduce posterior probability of hidden variable into log-likelihood and get lower bound of log-likelihood, and then maximize the lower bound to calculate new parameters.
- First, we assume $p(d_i) \propto n(d_i)$, thus it does not influence the log-likelihood:

$$\mathcal{L} = \sum_{m=1}^{M} \sum_{n=1}^{N_m} n(d_m, w_{m,n}) \log \sum_{k=1}^{K} \phi_{k,w_{m,n}} \theta_{m,k}$$

- Second, we get the lower bound as:

$$\begin{aligned} \mathcal{L} &= \sum_{m=1}^{M} \sum_{n=1}^{N_m} n(d_m, w_{m,n}) \log \sum_{k=1}^{K} P(z_k|d_m, w_{m,n}) \frac{\phi_{k,w_{m,n}} \theta_{m,k}}{P(z_k|d_m, w_{m,n})} \\ &\geq \sum_{m=1}^{M} \sum_{n=1}^{N_m} n(d_m, w_{m,n}) \sum_{k=1}^{K} P(z_k|d_m, w_{m,n}) \log \frac{\phi_{k,w_{m,n}} \theta_{m,k}}{P(z_k|d_m, w_{m,n})} \end{aligned}$$

M-Step

We introduce lagrange multiplier λ with the constraint $\sum_{t=1}^{V} \phi_{k,t} = 1$, and derive the partial derivative of $\phi_{k,t}$ from the lower bound of $\mathcal{L}$:

$$\frac{\partial \mathcal{L}}{\partial \phi_{k,t}} = \frac{\sum_{m=1}^{M} n(d_m, w_t) P(z_k|d_m, w_t)}{\phi_{k,t}} + \lambda = 0$$

With constraint $\sum_{t=1}^{V} \phi_{k,t} = 1$, we can get:

$$\lambda = -\sum_{m=1}^{M} \sum_{n=1}^{N_m} n(d_m, w_{m,n}) P(z_k|d_m, w_{m,n})$$

Thus:

$$\phi_{k,t} = \frac{\sum_{m=1}^{M} n(d_m, w_t) P(z_k|d_m, w_t)}{\sum_{m=1}^{M} \sum_{n=1}^{N_m} n(d_m, w_{m,n}) P(z_k|d_m, w_{m,n})}$$

Then $\theta_{m,k}$ is calculated in the same way.

The final update Equations

- E-step:

$$q(z) = P(z \mid d, w) = \frac{\theta_{k,t}\phi_{m,k}}{\sum_{l=1}^{K}\theta_{l,t}\phi_{m,l}}$$

- M-step:

$$\phi_{k,t} = P(w_t \mid z_k) = \frac{\sum_{m=1}^{M} n(d_m, w_t) P(z_k \mid d_m, w_t)}{\sum_{m=1}^{M}\sum_{n=1}^{N_m} n(d_m, w_{m,n}) P(z_k \mid d_m, w_{m,n})}$$

$$\theta_{m,k} = P(z_k \mid d_m) = \frac{\sum_{n=1}^{N_m} n(d_m, w_{m,n}) P(z_k \mid d_m, w_t)}{\sum_{m=1}^{M}\sum_{n=1}^{N_m} n(d_m, w_{m,n}) P(z_k \mid d_m, w_{m,n})}$$

$$P(z_k) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N_m} n(d_m, w_{m,n}) P(z_k \mid d_m, w_t)}{\sum_{m=1}^{M}\sum_{n=1}^{N_m} n(d_m, w_{m,n})}$$

Coding Design

- Variables:
  - **double**[][] *p_dz_n* // *p(d|z)*, |*D*|*|*Z*|
  - **double**[][] *p_wz_n* // *p(w|z)*, |*W*|*|*Z*|
  - **double**[] *p_z_n* // *p(z)*, |*Z*|
- Running Processing:
  1. Read dataset from file

     ArrayList<DocWordPair> doc; // all the docs

     DocWordPair – (word_id, word_frequency_in_doc)
  2. Parameter Initialization

     Assign each elements of *p_dz_n, p_wz_n and p_z_n* with a random double value, satisfying $\sum_d p_dz_n = 1$, $\sum_d p_wz_n = 1$, and $\sum_d p_z_n = 1$
  3. Estimation (Iterative processing)
     1. Update *p_dz_n, p_wz_n and p_z_n*
     2. Calculate Log-likelihood function to see where (|Log-likelihood – old_Log-likelihood| < threshold)
  4. Output *p_dz_n, p_wz_n and p_z_n*

Coding Design

- **Update *p_dz_n***

```
For each doc d{
   For each word w included in d   {
       denominator = 0;
       nominator = new double[Z];
       For each topic z   {
          nominator[z] = p_dz_n[d][z]* p_wz_n[w][z]* p_z_n[z]
           denominator  +=nominator[z];
       } // end for each topic z
       For each topic z    {
          P_z_condition_d_w = nominator[j]/denominator;
          nominator_p_dz_n[d][z] += tf_wd*P_z_condition_d_w;
           denominator_p_dz_n[z] += tf_wd*P_z_condition_d_w;
       } // end for each topic z
    }// end for each word w included in d
}// end for each doc d

For each doc d {
      For each topic z                {
         p_dz_n_new[d][z] = nominator_p_dz_n[d][z]/ denominator_p_dz_n[z];
      } // end for each topic z
}// end for each doc d
```

Coding Design

- **Update *p_wz_n***

```
For each doc d{
   For each word w included in d   {
        denominator = 0;
        nominator = new double[Z];
        For each topic z   {
          nominator[z] = p_dz_n[d][z]* p_wz_n[w][z]* p_z_n[z]
           denominator  +=nominator[z];
        } // end for each topic z
        For each topic z    {
          P_z_condition_d_w = nominator[z]/denominator;
          nominator_p_wz_n[w][z] += tf_wd*P_z_condition_d_w;
           denominator_p_wz_n[z] += tf_wd*P_z_condition_d_w;
        } // end for each topic z
    }// end for each word w included in d
}// end for each doc d

For each w {
     For each topic z                {
        p_wz_n_new[w][z] = nominator_p_wz_n[w][z]/ denominator_p_wz_n[z];
     } // end for each topic z
}// end for each doc d
```

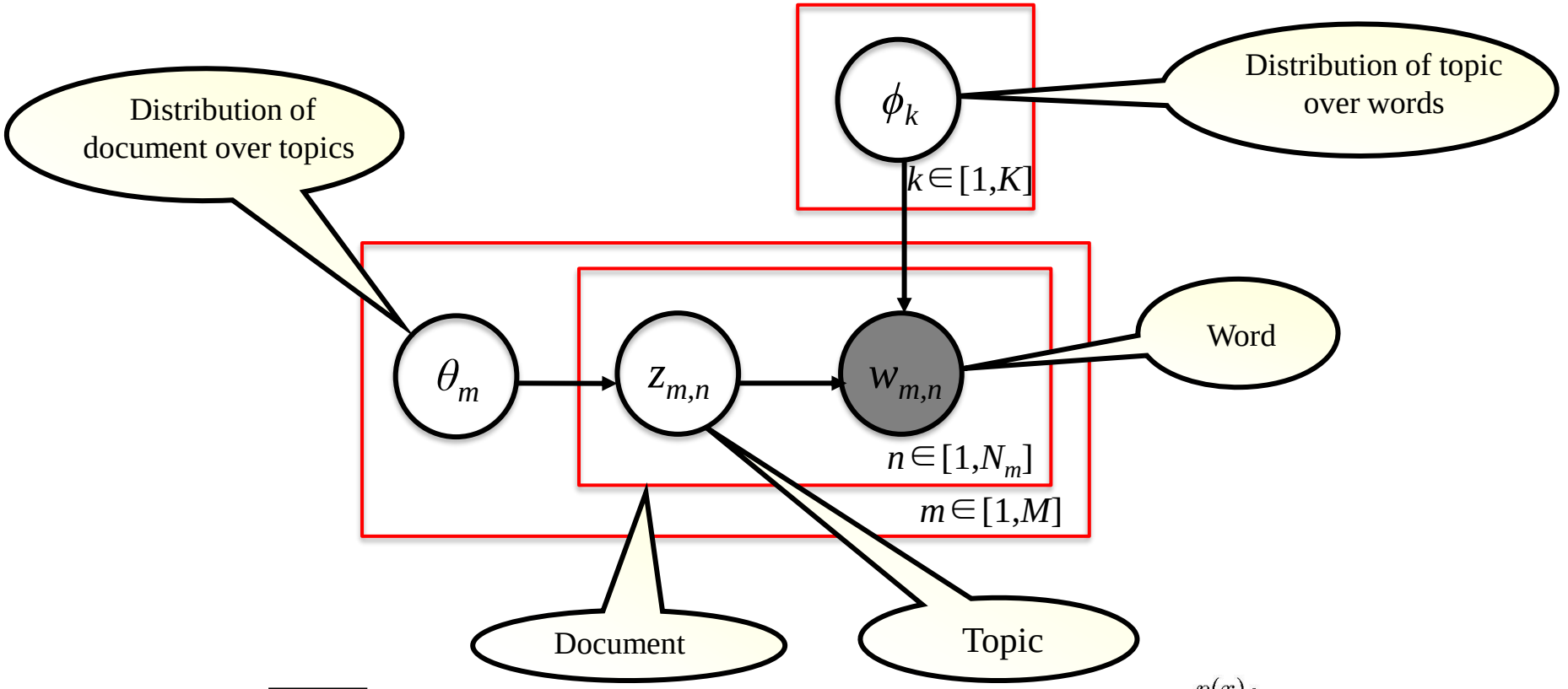
Coding Design

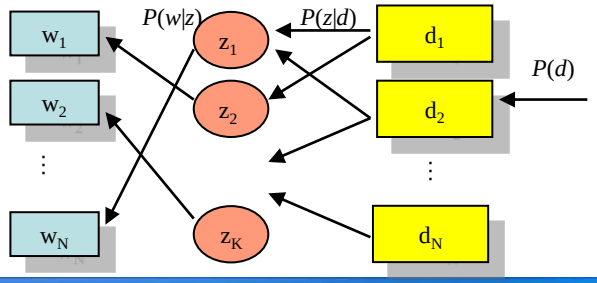
- **Update *p_z_n***

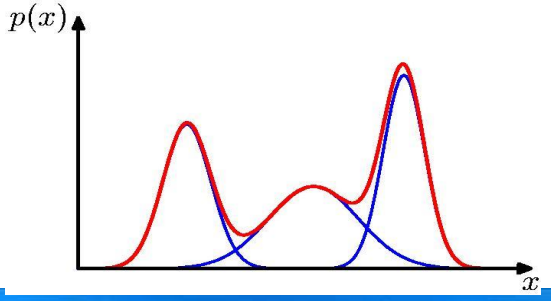
```
For each doc d{
   For each word w included in d   {
       denominator = 0;
       nominator = new double[Z];
       For each topic z   {
           nominator[z] = p_dz_n[d][z]* p_wz_n[w][z]* p_z_n[z]
            denominator  +=nominator[z];
       } // end for each topic z
       For each topic z    {
           P_z_condition_d_w = nominator[z]/denominator;
           nominator_p_z_n[z] += tf_wd*P_z_condition_d_w;
       } // end for each topic z
       denominator_p_z_n[z] += tf_wd;
    }// end for each word w included in d
}// end for each doc d

For each topic z{
        p_dz_n_new[d][z] = nominator_p_z_n[z]/ denominator_p_z_n;
} // end for each topic z
```

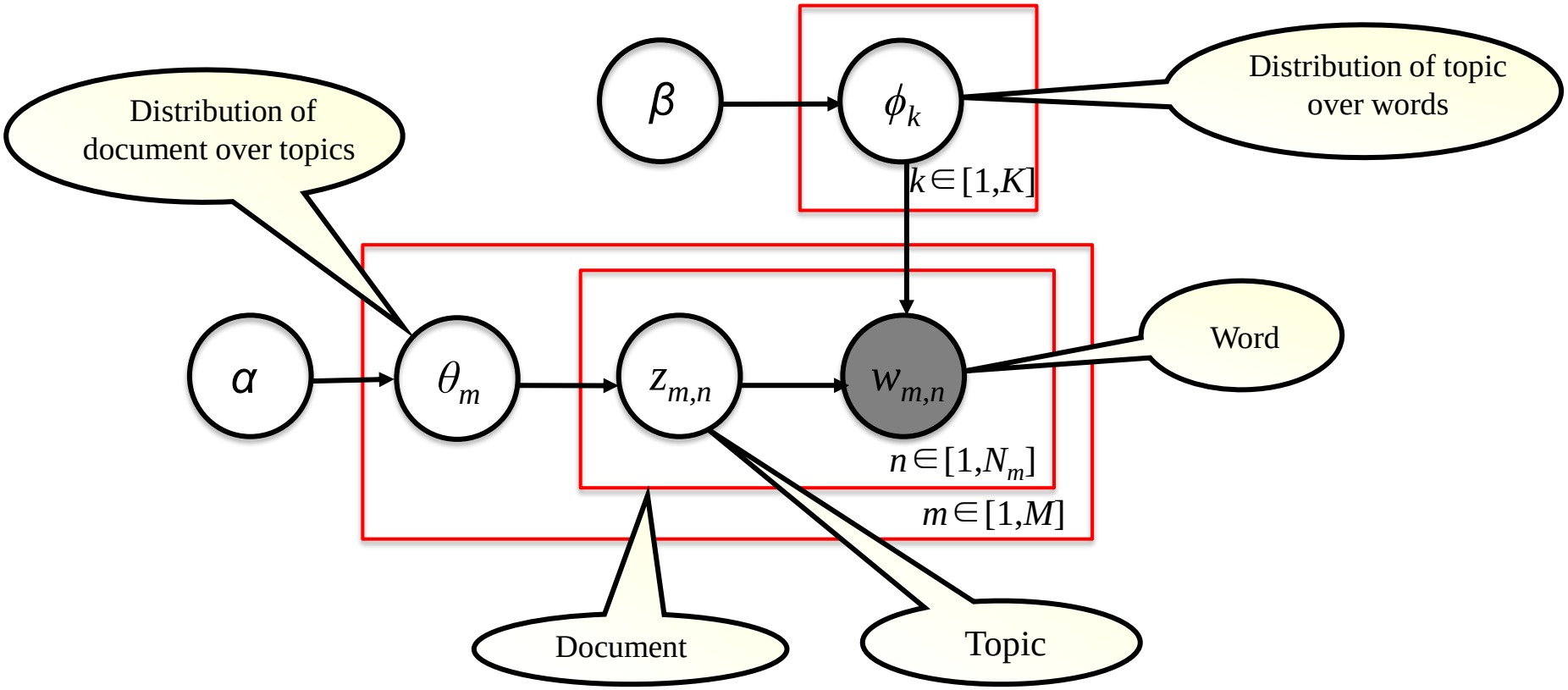
PLSI(SIGIR'99)







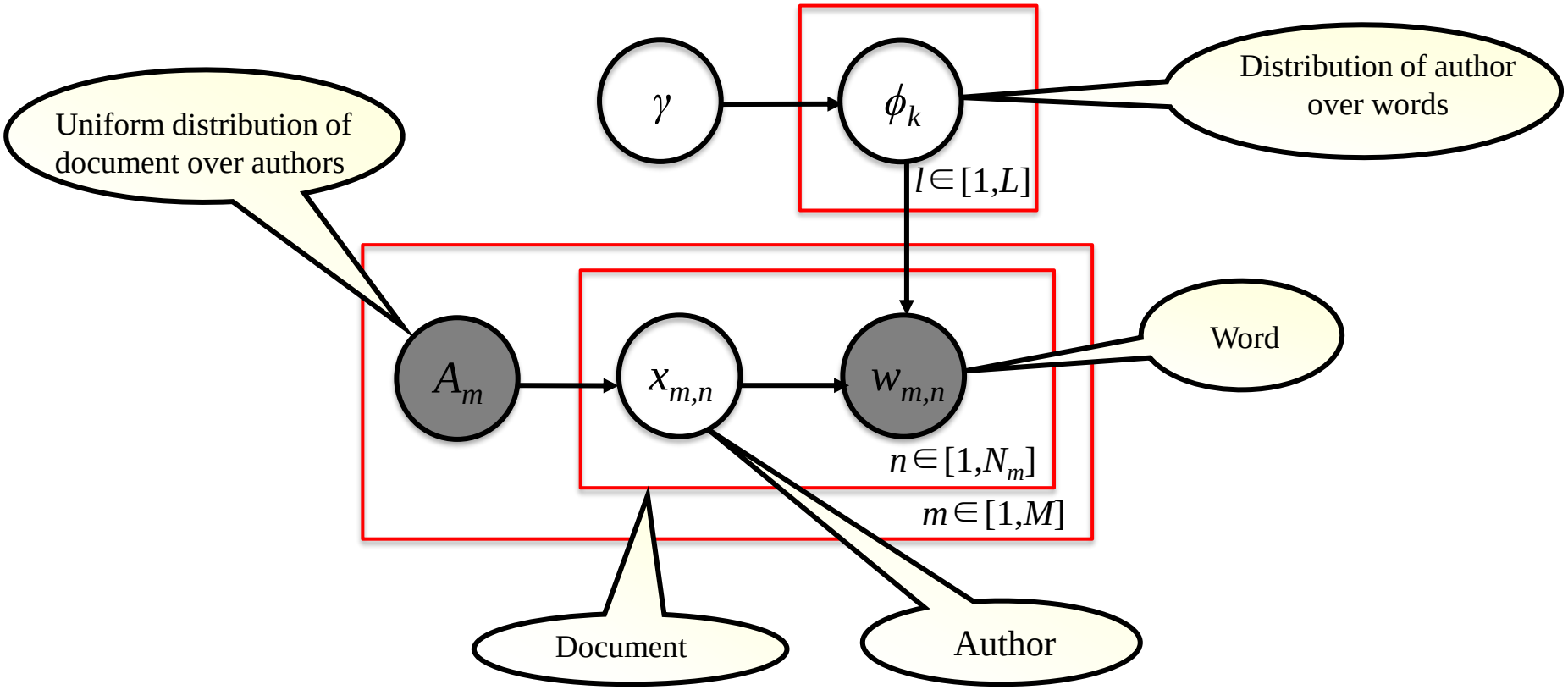
LDA (JMLR'03)



α, β: Prior distributions

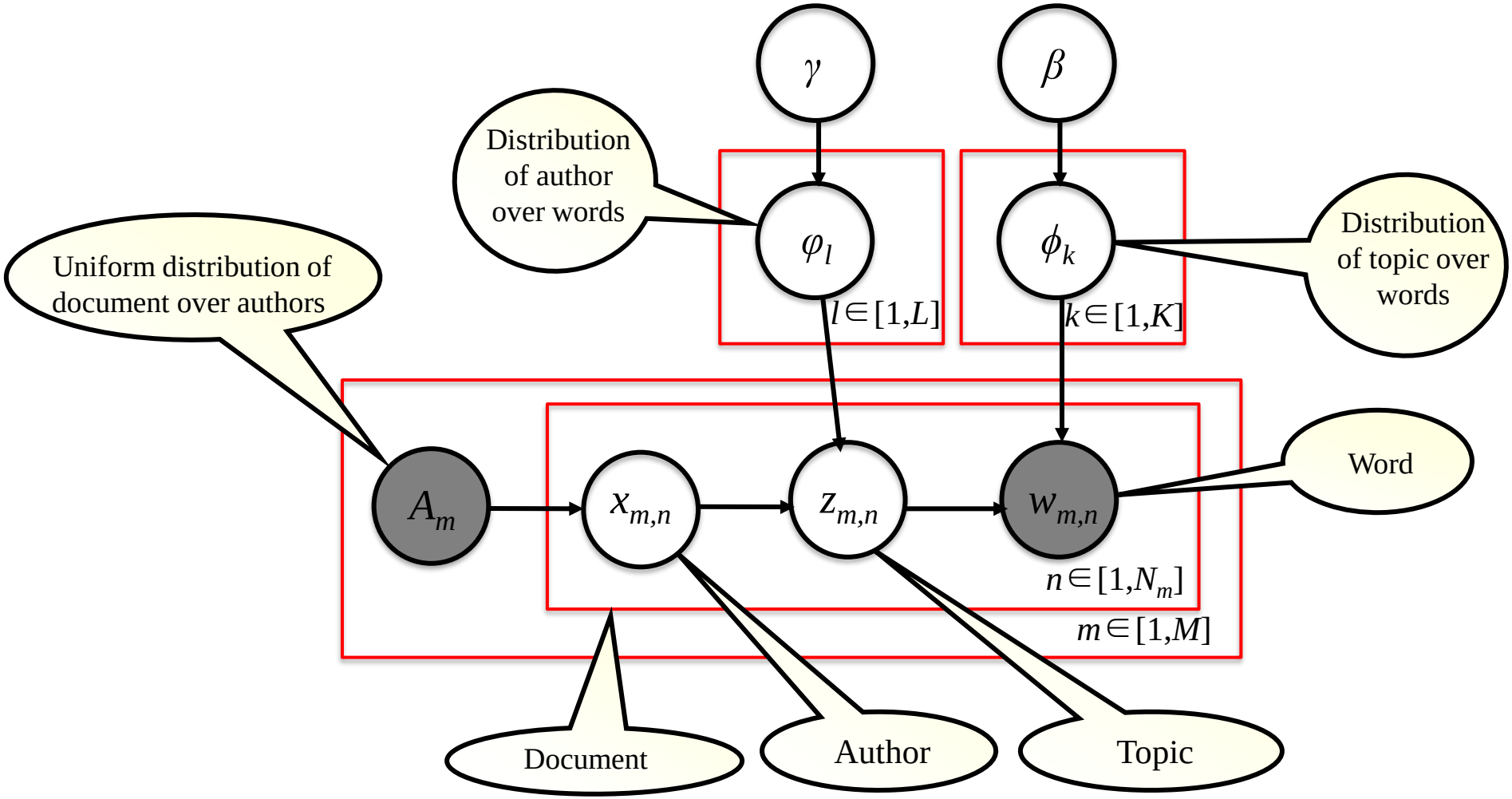
Modeling Authors with Words

(AAAI'09 workshop)



γ, β: Prior distributions

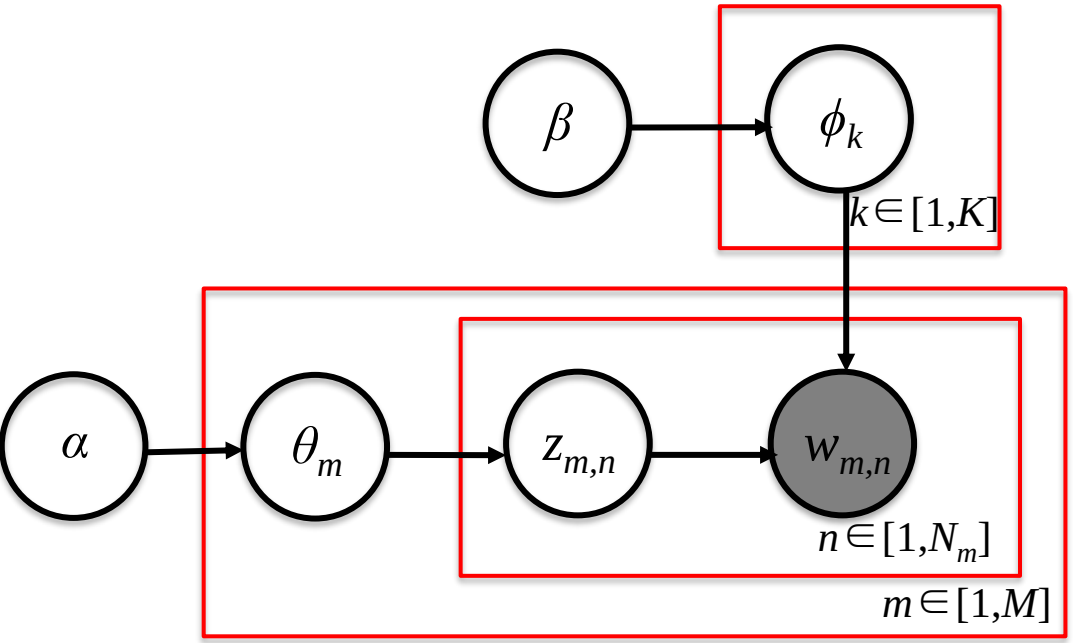
Author-Topic Model (UAI'04, KDD'05)



γ, β: Prior distributions

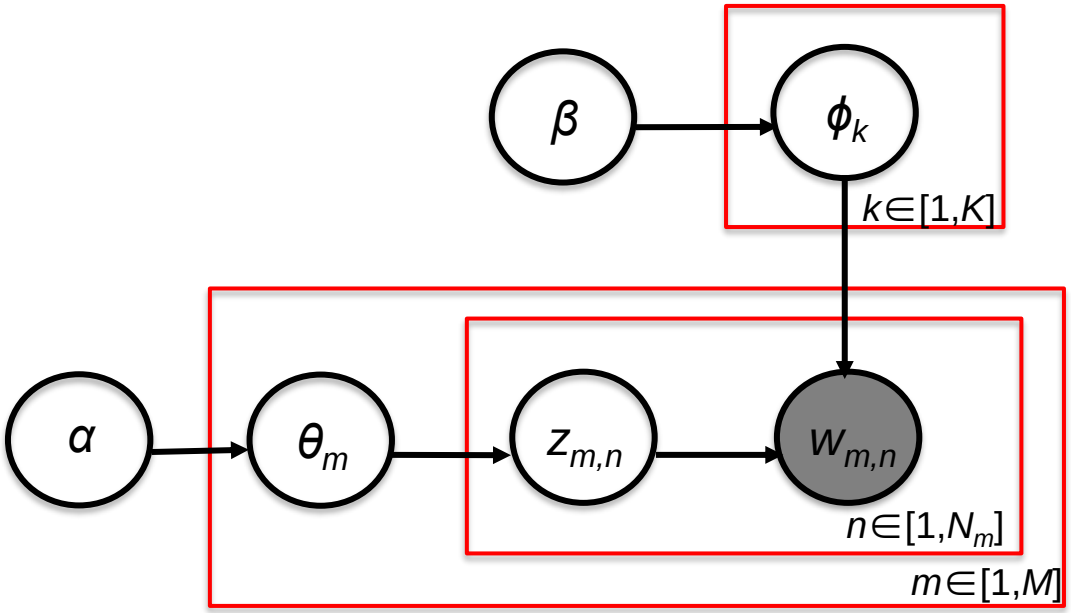
Latent Dirichlet Allocation (LDA)

LDA Model



- For each topic z_k, with $k \in [1, K]$:
  - generate a topic distribution over the vocabulary $\vec{\phi_k} \sim \text{Dir}(\vec{\beta})$;
- For each doc d_m, with $m \in [1, M]$:
  - Generate a document distribution over topics $\vec{\theta_m} \sim \text{Dir}(\vec{\alpha})$;
  - Generate document length $N_m \sim \text{Poiss}(\xi)$.
  - For each word $w_{m,n}$ with $n \in [1, N_m]$:
    - Generate a topic $z_{m,n} \sim \text{Mult}(\vec{\theta}_m)$;
    - Generate a word $w_{m,n} \sim \text{Mult}(\vec{\phi}_{z_{m,n}})$.

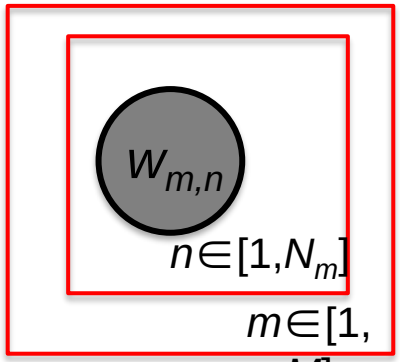
Three-level Representation



Similar model can be referred to hierarchical models (Gelman et al., 1995), or more precisely as conditionally independent hierarchical models (Kass and Steffey, 1989).

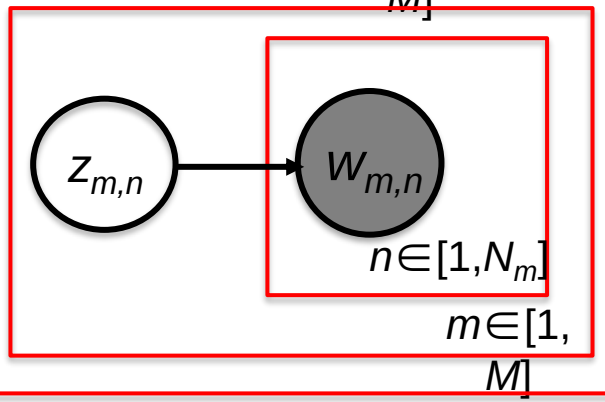
- **α ,β,** Φ are corpus-level parameters
  - Sampled once when generating a corpus
- $\boldsymbol{\theta}_m$ are document-level variables
  - Sampled once per document
- $z_{m,n}$,$w_{m,n}$ are word-level variables
  - Sampled once for each word in each document

Relationships with Other Latent Models



Unigram

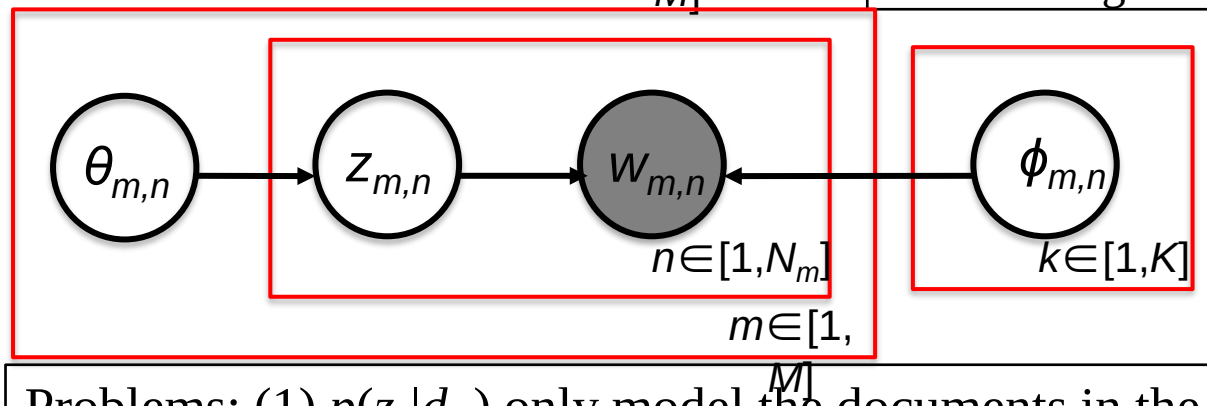
$$p(\mathbf{w}_m) = \prod_{n=1}^{N_m} p(w_{m,n})$$



Mixture of unigrams

$$p(\mathbf{w}_m) = \sum_{k=1}^{K} p(z_k) \prod_{n=1}^{N_m} p(w_{m,n} \mid z_k)$$

Problems: Each document exhibits exactly one topic. This assumption is often too limiting to effectively model a large collection of documents.

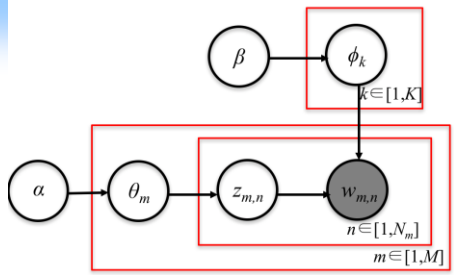


PLSI/Aspect model

$$p(\mathbf{w}_m) = \prod_{n=1}^{N_m} \sum_{k=1}^{K} p(w_{m,n} \mid z_k) p(z_k \mid d_m)$$

Problems: (1) $p(z_k|d_m)$ only model the documents in the training data set and cannot for the unseen document; and (2) the parameter $KV+KM$ grows linearly in M and thus overfitting.

Joint Probability



- The probability of the n-th word in m-th document being word t is:

$$P(w_{m,n} = t|\vec{\theta}_m, \Phi) = \sum_{k=1}^{K} P(w_{m,n} = t|\vec{\phi}_k) P(z_{m,n} = k|\vec{\theta}_m)$$

where $\Phi = \{\vec{\phi}_k\}_{k=1}^{K}$.

- We write the joint probability of words, hidden variables, and parameters in one document as:

$$P(\vec{w}_m, \vec{z}_m, \vec{\theta}_m, \Phi|\vec{\alpha}, \vec{\beta}) = P(\vec{\theta}_m|\vec{\alpha}) P(\Phi|\vec{\beta}) \prod_{n=1}^{N_m} P(w_{m,n}|\vec{\phi}_{z_{m,n}}) P(z_{m,n}|\vec{\theta}_m)$$

- Then the joint probability in the whole corpus as:

$$P(\mathcal{W}, \mathcal{Z}, \Theta, \Phi|\alpha, \beta) = \prod_{k=1}^{K} P(\vec{\phi}_k|\vec{\beta}) \prod_{m=1}^{M} \left(P(\vec{\theta}_m|\vec{\alpha}) \prod_{n=1}^{N_m} P(w_{m,n}|\vec{\phi}_{z_{m,n}}) P(z_{m,n}|\vec{\theta}_m) \right) \tag{13}$$

Gibbs Sampling Derivation for LDA

Gibbs Sampling

- The key idea of the Gibbs sampling is to sequentially update each variable of the distribution according the conditional probability given all the other variables.
- Take a distribution of the three variable (X, Y, Z) for example, the Gibbs sampler makes the Markov chain transit from (x_i, y_i, z_i) to (x_{i+1}, y_{i+1},z_{i+1}) by the following steps:
  - Draw x_{i+1} from p($X|Y=y_i$, $Z=z_i$)
  - Draw y_{i+1} from p($Y|X=x_{i+1}$, $Z=z_i$)
  - Draw z_{i+1} from p($Z|X=x_{i+1}$, $Y=y_{i+1}$)

Recall Dirichlet Distribution

Thomas P. Minka. Estimating a Dirichlet Distribution. 2003, Technique Report.

Multinomial Distribution

- Multinomial distribution is a multivariate generalization of the binomial distribution. There are n independent trials, where each trial results in exactly one of some fixed finite number k possible outcomes, with probabilities $p_1, \cdots, p_k$ (so that $p_i > 0$ for $i = 1, \cdots, k$ and $\sum_{i=1}^{k} p_i = 1$):

$$P(X_1 = x_1, \cdots, X_k = x_k; n, p_1, \cdots, p_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

where X_i means the number of event i, and $\sum_{i=1}^{k} x_i = n$

Dirichlet Distribution

- Dirichlet distribution is a multivariate generalization of beta distribution. The probability density function of the Dirichlet distribution, with shape parameters $\alpha_1, \cdots, \alpha_K > 0$, is:

$$f(x_1, \cdots, x_{K-1}; \alpha_1, \cdots, \alpha_K) = \frac{1}{\Delta(\vec{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

where $\frac{1}{\Delta(\vec{\alpha})} = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}$, and $\sum_{i=1}^{K} x_i = 1$ and $x_1, \cdots, x_K > 0$.

Normalization factor

To Understand Dirichlet Distribution

- Dirichlet distribution is the distribution of distribution.
- Example
  - Conduct a trial of tossing a coin 10000 times.
  - If we get the result (2000,2000,2000,1500,1500,1000), which are the occurring times of each number from 1 to 6, then the value of $x_1,\dots, x_K$ are (0.2,0.2,0.2,0.15,0.1).
  - Suppose we conduct 10000 times of the above trial, we can get the occurring probability of the distribution (0.2,0.2,0.2,0.15,0.1).

To Understand Dirichlet Distribution (Cont.)

- Dirichlet distribution are very often used as prior distribution in Bayesian inference.
- When we have no prior information to favor one component over any other, the simplest Dirichlet prior is selected the symmetric Dirichlet distribution where all parameters are equal.
- In this case, the single value α is called the concentration parameter.
  - With a value much less than 1, the mass will be highly concentrated in a few components, and all the rest will have almost no mass.
  - With a value much greater than 1, the mass will be dispersed almost equally among all the components.

Multinomial-Dirichlet Conjugate

- Assume prior distribution as $P(x;\alpha_1,\cdots,\alpha_k)=\frac{1}{\Delta(\vec{\alpha})}\prod_{i=1}^{K}x_i^{\alpha_i-1}$ and likelihood distribution as $P(\vec{y}|\vec{x})=\frac{n!}{n_1!\cdots n_k!}\prod_{i=1}^{K}x_i^{n_i}$;
- We can know the posterior distribution as:

$$P(\vec{x}|\vec{y})\propto P(\vec{y}|\vec{x})P(\vec{x})\propto\prod_{i=1}^{K}x_i^{\alpha_i+n_i-1}$$

- Since we need $\int_0^1\prod_{i=1}^{K}x_i^{\alpha_i+n_i-1}dx_i=1$ and we know the definition $\Delta(\vec{x})=\int_0^1\prod_{i=1}^{K}x_i^{n_i}dx_i$
- Thus $\int_0^1\prod_{i=1}^{K}(x_i+\alpha_i-1)^{n_i}dx_i=\Delta(\vec{\alpha}+\vec{n}-1)$, where $\vec{n}=\{n_t\}_{t=1}^{V}$
- Therefore, $P(\vec{x}|\vec{y})\propto\frac{1}{\Delta(\vec{\alpha}+\vec{n}-1)}\prod_{i=1}^{k}x_i^{\alpha_i+n_i-1}$

Expectation of Dirichlet

- If $\vec{p} \sim \text{Dir}(\vec{p}|\vec{\alpha})$, then the expectation of any element p_i is:

$$
\begin{aligned}
E(p_i) &= \int_0^1 p_i \text{Dir}(\vec{p}|\vec{\alpha}) dp_i \\
&= \frac{\Gamma(\sum_{k=1}^K)\alpha_k}{\prod_{k=1}^K \Gamma(\alpha_k)} \int_0^1 \prod_{k=1}^K p_i^{\beta_i - 1} \\
&= \frac{\Gamma(\sum_{k=1}^K)\alpha_k}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\prod_{k=1}^K \Gamma(\beta_k)}{\Gamma(\sum_{k=1}^K \beta_k)} \\
&= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\Gamma(\alpha_i)} \frac{\Gamma(\alpha_i + 1)}{\Gamma(\sum_{k=1}^K \alpha_k + 1)} \\
&= \frac{\alpha_i}{\sum_{k=1}^K \alpha_k}, \text{ Use } \Gamma(x+1) = x\Gamma(x)
\end{aligned}
$$

$$
\beta_k = \begin{cases} \alpha_k + 1 & \text{if } k = i \\ \alpha_k & \text{otherwise.} \end{cases}
$$

Collapsed Gibbs Sampling Derivation for LDA

Gregor Heinrich. Parameter estimation for text analysis. 2008, Technique Report.

Integrate Out Φ

- The basic idea is to integrate parameters Θ and Φ, and only sample the hidden variables $\mathcal{Z}$.
- We observe that, given $\mathcal{W}$, parameters Θ and Φ are conditional independent. We integrate out each ϕ_k in Eq. 13 with respect to Φ, because different $\phi_k$with regards to different topic z_k are independent:

$$
\begin{aligned}
P(\mathcal{W}|\mathcal{Z}) &= \prod_{k=1}^{K} \int P(\vec{\phi}_k|\vec{\beta}) \prod_{m=1}^{M} \prod_{n=1}^{N_m} P(w_{m,n}|\vec{\phi}_{z_{m,n}}) d\vec{\phi}_k \quad (14) \\
&= \prod_{k=1}^{K} \int \frac{1}{\Delta(\vec{\beta})} \prod_{t=1}^{V} \phi_{k,t}^{n_{\cdot,k,t}+\beta_t-1} d\vec{\phi}_k \\
&= \prod_{k=1}^{K} \frac{\Delta(\vec{n}_k+\vec{\beta})}{\Delta(\beta)}, \Delta(\vec{n}_k) = \{n_{\cdot,k,t}\}_{t=1}^{V}
\end{aligned}
$$

Integrate Out Θ

- In the same way, we integrate out each θ_m in Eq. 13 with respect to Θ, because different $\theta_m$with regards to different document d_m are independent:

$$
\begin{aligned}
P(\mathcal{Z}) &= \prod_{m=1}^{M} \int P(\vec{\theta}_m|\vec{\alpha}) \prod_{n=1}^{N_m} P(z_{m,n}|\vec{\theta}_m) d\vec{\theta}_m \qquad (15) \\
&= \prod_{m=1}^{M} \int \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{K} \theta_{m,k}^{n_{m,k,\cdot}+\alpha_k-1} d\vec{\theta}_m \\
&= \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\alpha)}, \Delta(\vec{n}_m) = \{n_{m,k,\cdot}\}_{k=1}^{K}
\end{aligned}
$$

Collapsed Joint Probability

- Thus we get:

$$P(\mathcal{W}, \mathcal{Z}) = \prod_{k=1}^{K} \frac{\Delta(\vec{n}_k + \vec{\beta})}{\Delta(\beta)} \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\alpha)}$$

LDA Gibbs Sampler

- The goal is to calculate $P(z_{m,n} = k|\mathcal{Z}^{\neg m,n}, \mathcal{W})$.
- Assume $w_{m,n} = t$.
- Because we know $P(z_{m,n} = k|\mathcal{Z}^{\neg m,n}, \mathcal{W}) = \frac{P(\mathcal{Z},\mathcal{W})}{P(\mathcal{Z}^{\neg m,n},\mathcal{W})}$
- Based on $P(\mathcal{W}, \mathcal{Z}) = \prod_{k=1}^{K} \frac{\Delta(\vec{n}_k+\vec{\beta})}{\Delta(\beta)} \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\alpha)}$, we can get :

$$P(z_{m,n} = k|\mathcal{Z}^{\neg m,n}, \mathcal{W}) = \frac{\Delta(\vec{n}_k + \vec{\beta})}{\Delta(\vec{n}_k^{\neg m,n} + \vec{\beta})} \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{n}_m^{\neg m,n} + \vec{\alpha})} \tag{16}$$

$$= \frac{\Gamma(n_{\cdot,k,t} + \beta_t)\Gamma(\sum_{t=1}^{V} n_{\cdot,k,t}^{\neg m,n} + \beta_t)}{\Gamma(n_{\cdot,k,t}^{\neg m,n} + \beta_t)\Gamma(\sum_{t=1}^{V} n_{\cdot,k,t} + \beta_t)} \frac{\Gamma(n_{m,k,\cdot} + \alpha_k)\Gamma(\sum_{k=1}^{K} n_{m,k,\cdot}^{\neg m,n} + \alpha_k)}{\Gamma(n_{m,k,\cdot}^{\neg m,n} + \alpha_k)\Gamma(\sum_{k=1}^{K} n_{m,k,\cdot} + \alpha_k)}$$

$$= \frac{n_{\cdot,k,t} + \beta_t - 1}{[\sum_{t=1}^{V} n_{\cdot,k,t} + \beta_t] - 1} \frac{n_{m,k,\cdot} + \alpha_k - 1}{[\sum_{k=1}^{K} n_{m,k,\cdot} + \alpha_k] - 1}$$

$$\propto \frac{n_{\cdot,k,t} + \beta_t - 1}{[\sum_{t=1}^{V} n_{\cdot,k,t} + \beta_t] - 1} (n_{m,k,\cdot} + \alpha_k - 1)$$

Gregor Heinrich. Parameter estimation for text analysis. 2008, Technique Report.

Solve Parameters

- The final objective is to estimate parameters Θ and Φ.
- According to Bayes' rules and Dirichlet prior, we can get :

$$P(\vec{\theta}_m|\vec{z}_m,\vec{\alpha}) = \frac{1}{Z_{\theta_m}}\prod_{n=1}^{N_m} P(z_{m,n}|\vec{\theta}_m)P(\vec{\theta}_m|\vec{\alpha}) = \mathrm{Dir}(\vec{\theta}_m|\vec{n}_m+\vec{\alpha})$$

$$P(\vec{\phi}_k|\mathcal{Z},\mathcal{W},\vec{\beta}) = \frac{1}{Z_{\phi_k}}\prod_{m,n:z_{m,n}=k} P(w_{m,n}|\vec{\phi}_k)P(\vec{\phi}_k|\vec{\beta}) = \mathrm{Dir}(\vec{\phi}_k|\vec{n}_k+\vec{\beta})$$

- Then we solve the expectation of Dirichlet distribution and get:

$$\theta_{m,k} = \frac{n_{m,k,\cdot}+\alpha_k}{\sum_{k=1}^{K} n_{m,k,\cdot}+\alpha_k}, \phi_{k,t} = \frac{n_{\cdot,k,t}+\beta_t}{\sum_{t=1}^{V} n_{\cdot,k,t}+\beta_t} \qquad (17)$$

Algorithm

1. Initialize $n_{m,k,\cdot}$ and $n_{\cdot,k,t}$ as 0
2. Repeat
3. For $m = 1$ to M
4. For $n = 1$ to N_m
5. $n_{m,z_{m,n},\cdot} = n_{m,z_{m,n},\cdot} - 1$
6. $n_{\cdot,z_{m,n},w_{m,n}} = n_{\cdot,z_{m,n},w_{m,n}} - 1$
7. Sample topic z_k according to Eq (16).
8. $n_{m,k,\cdot} = n_{m,k,\cdot} + 1$
9. $n_{\cdot,k,w_{m,n}} = n_{\cdot,k,w_{m,n}} + 1$
10. Until convergence;
11. Update Θ and Φ according to Eq. (17).

Inference

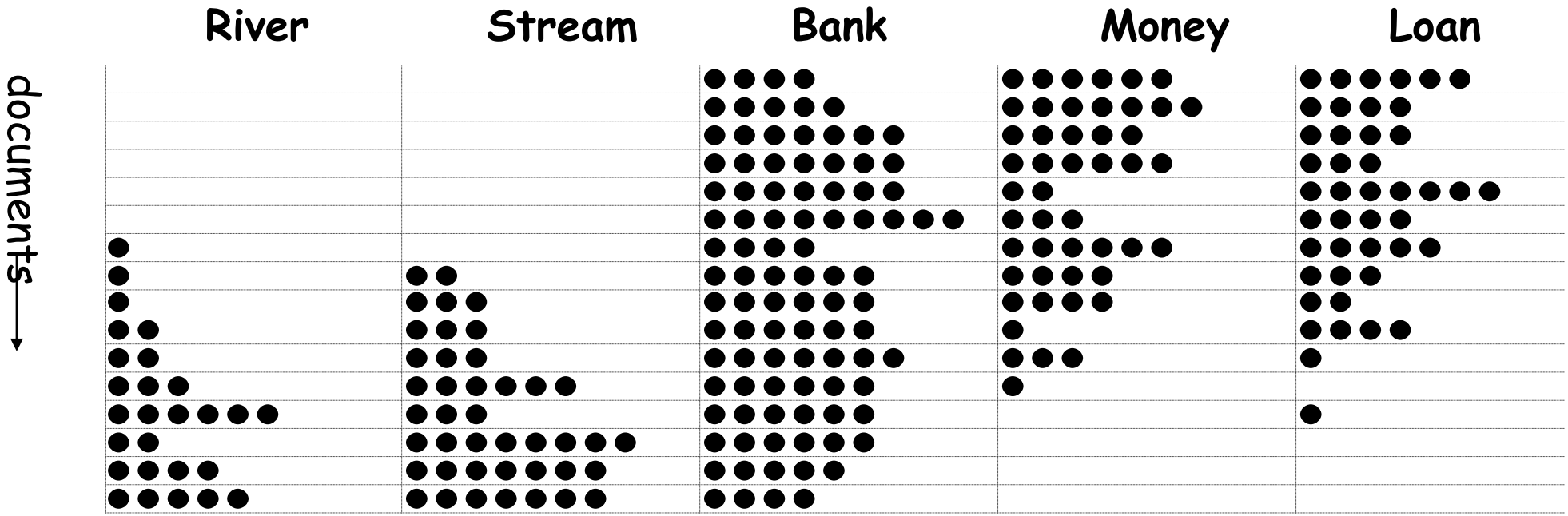
Given a new document d_m, we perform inference by

$$Assume \qquad w_{m,n} = t$$

$$P(z_{m,n} = k \mid Z^{\neg(m,n)}, W) \propto \frac{n_{\cdot,k,t} + n'_{\cdot,k,t} + \beta_t - 1}{\sum_{t=1}^{V}(n_{\cdot,k,t} + n'_{\cdot,k,t} + \beta_t - 1)} \cdot \frac{n_{m,k,\cdot} + \alpha_k - 1}{\sum_{k=1}^{K}(n_{m,k,\cdot} + \alpha_k - 1)}$$

Where notations with ’ denotes the number in the new document.

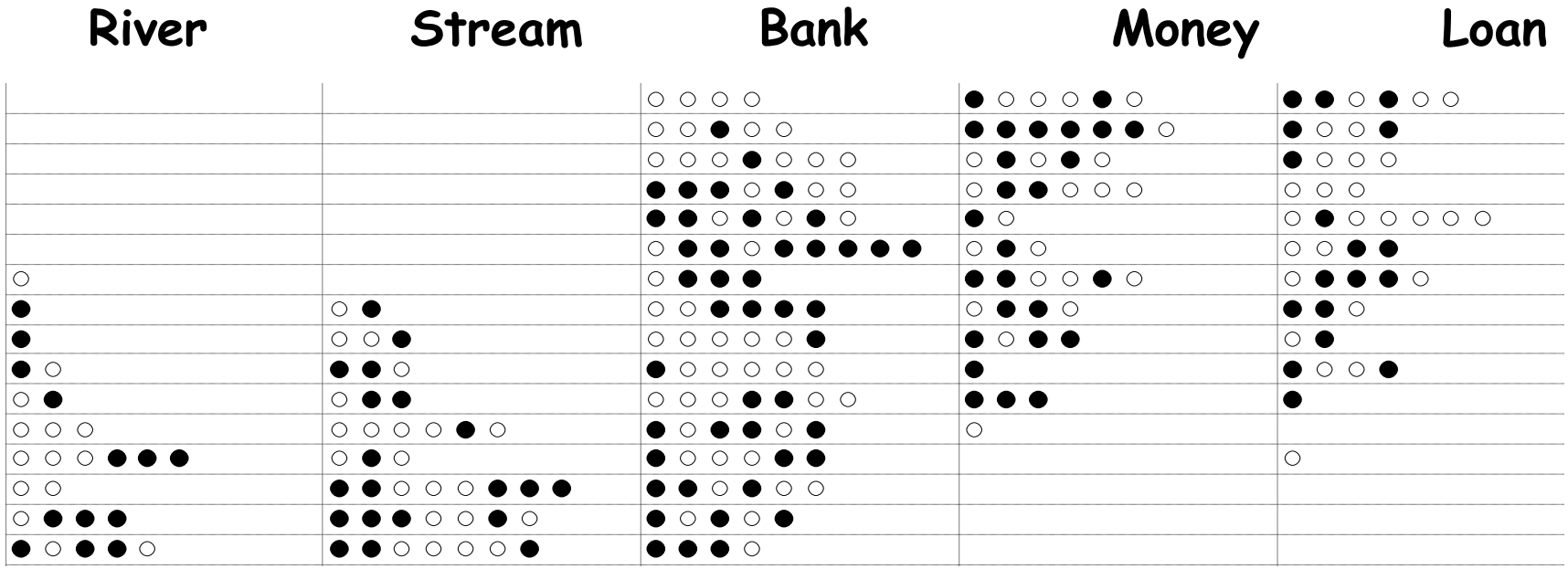
Topic Model: Inference



Can we recover the original topics and topic mixtures from this data?

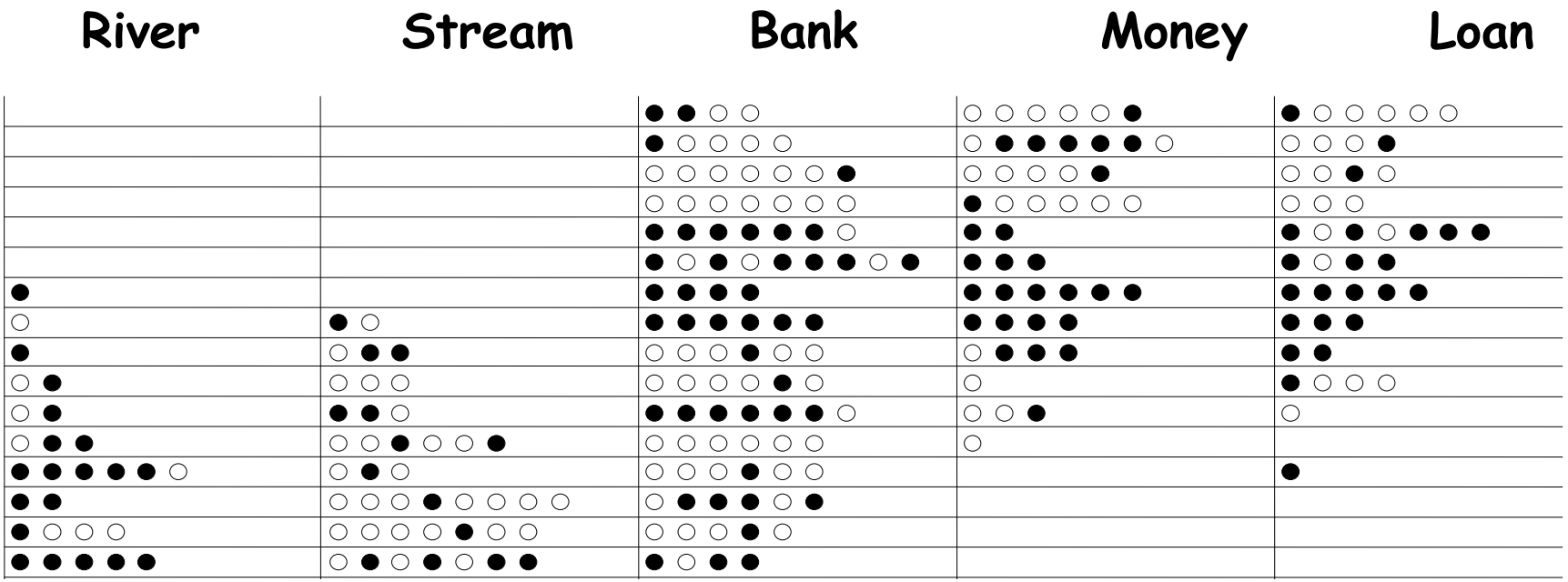
Example of Gibbs Sampling

- Assign word tokens randomly to topics (●=topic 1; ●=topic 2)

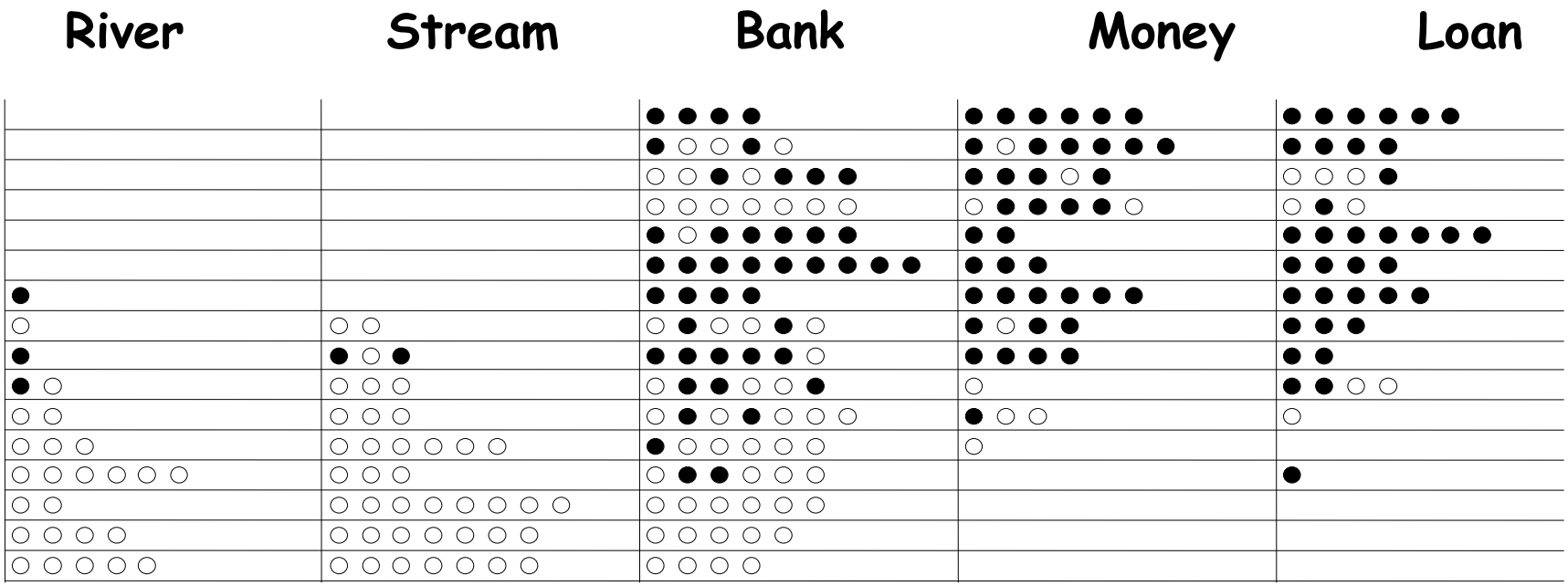


After 1 iteration

- Apply sampling equation to each word token



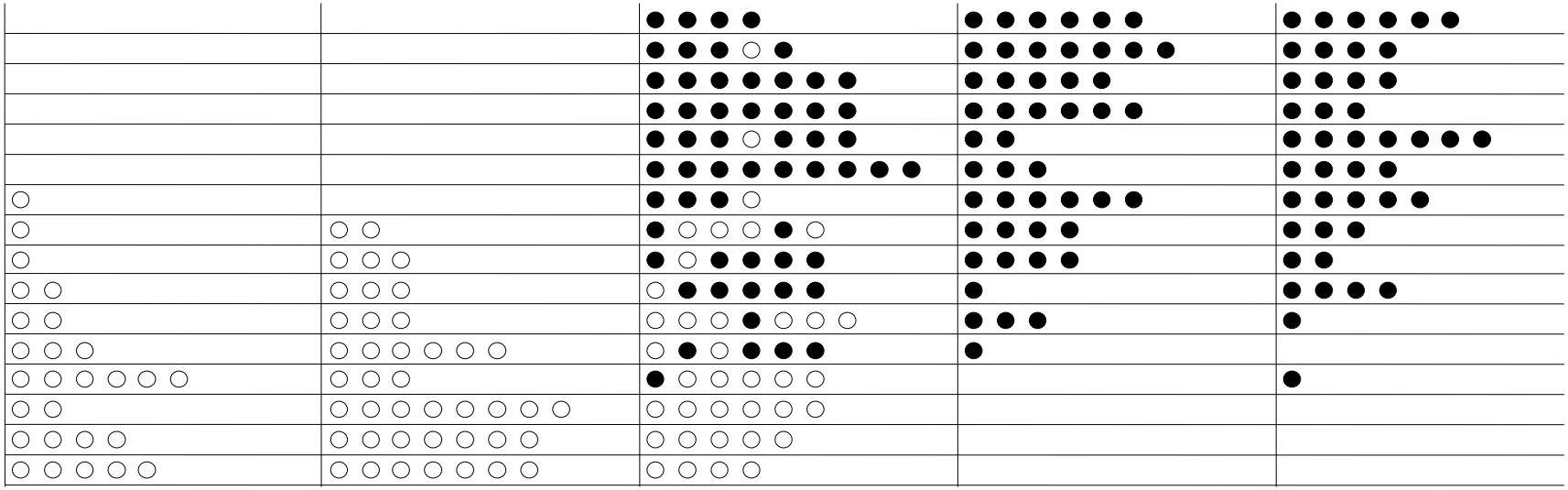
After 4 iterations



After 32 iterations

topic 1	
stream	.40
bank	.35
river	.25

topic 2	
bank	.39
money	.32
loan	.29



Hyper-parameter Estimation

- Stochastic EM (StEM) algorithm

Iteration(MCMC Gibbs sampling step)
  For each Iteration
    For each document
      For each item
        assign a topic from full conditional probability with
          input of α and β
        update count matrices.
      End for
    End for
  End for
  Estimate hyper-parameter α and β with input of updated count matrix
  Calculate document-topic and topic-item distribution with count matrices.

E-step

M-step

Hyper-parameter Estimation Methods

- Sampling Based Method

  Sampling new value of hyper-parameter from the posteriori distribution
  - Random Walk Metropolis-Hastings sampling
  - Adaptive Rejection Sampling

- MAP (Maximum A Posteriori) Based Method

  Obtain the new value by maximum a posterior probability using iteration methods.
  - Fixed-point iteration
  - Fixed-point by Leave-one-out Likelihood
  - Newton-Raphson iteration

The sampling based method

The full conditional for Gibbs sampling of α can be obtained by

$$p(\alpha|\vec{z}) \propto p(\vec{z}|\alpha)p(\alpha)$$
$$\propto \prod_m \frac{\Delta(\vec{n}_m + \alpha)}{\Delta_K(\alpha)} \alpha^{a-1} e^{-b\alpha}$$

Then for each parameter, we have

$$p(\alpha_i \mid \vec{\alpha}_{-i}, \vec{z}) = \frac{p(\alpha_i, \vec{\alpha}_{-i}, \vec{z})}{p(\vec{\alpha}_{-i}, \vec{z})} \propto p(\alpha_i, \vec{\alpha}_{-i}, \vec{z})$$

$$\propto p(\vec{z} \mid \vec{\alpha}) p(\vec{\alpha})$$

$$\propto \prod_m \left(\frac{\Gamma(n_{mi} + \alpha_i)}{\Gamma(\sum_k (n_{mk} + \alpha_k))} \cdot \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\alpha_i)} \right) \cdot \alpha_i^{a-1} e^{-b\alpha_i}$$

Analogously,

Analogously, the full conditional for Gibbs sampling of β can be obtained by

$$p(\beta|\vec{z};\vec{w}) \propto p(\vec{z}|\beta;\vec{w})p(\beta)$$
$$\propto \prod_k \frac{\Delta(\vec{n}_k+\beta)}{\Delta_V(\beta)}\beta^{a-1}e^{-b\beta}$$

For each parameter, we have

$$p(\beta_i \mid \vec{\beta}_{-i},\vec{z};\vec{w}) = \frac{p(\beta_i,\vec{\beta}_{-i},\vec{z};\vec{w})}{p(\vec{\beta}_{-i},\vec{z};\vec{w})} \propto p(\beta_i,\vec{\beta}_{-i},\vec{z};\vec{w})$$

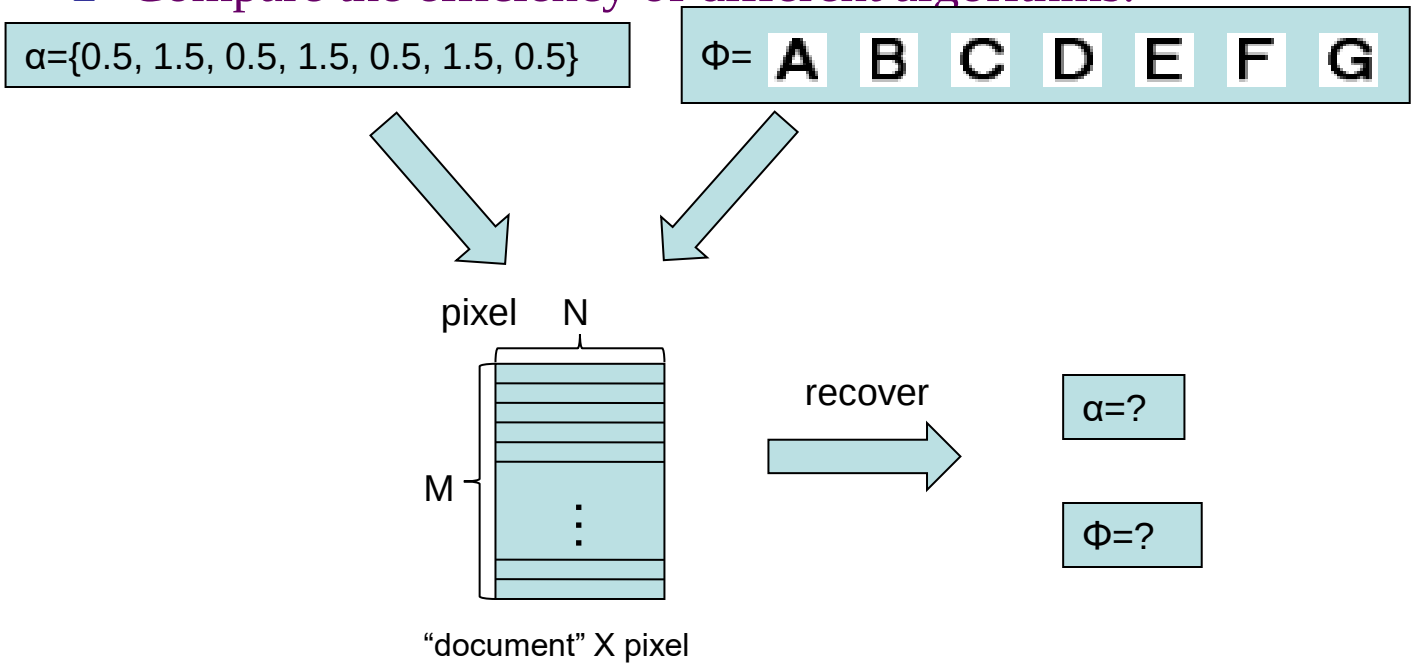
$$\propto p(\vec{z} \mid \vec{\beta};\vec{w})p(\vec{\beta};\vec{w})$$

$$\propto \prod_k \left(\frac{\Gamma(n_{ki}+\beta_i)}{\Gamma(\sum_v (n_{kv}+\alpha_v))} \cdot \frac{\Gamma(\sum_v \beta_v)}{\Gamma(\beta_i)} \right) \cdot \beta_i^{a-1}e^{-b\beta_i}$$

There are also many other methods (some non-sampling based method). For details, please refer to [Mink, 2000].

Experimental Results

- Experiment Design
  - Generate a synthetic data set using a predefined α and Φ.
  - Use different algorithms to recover both Φ and the hyperparameters α.
  - Compare the effectiveness of different learned algorithms.
  - Compare the efficiency of different algorithms.

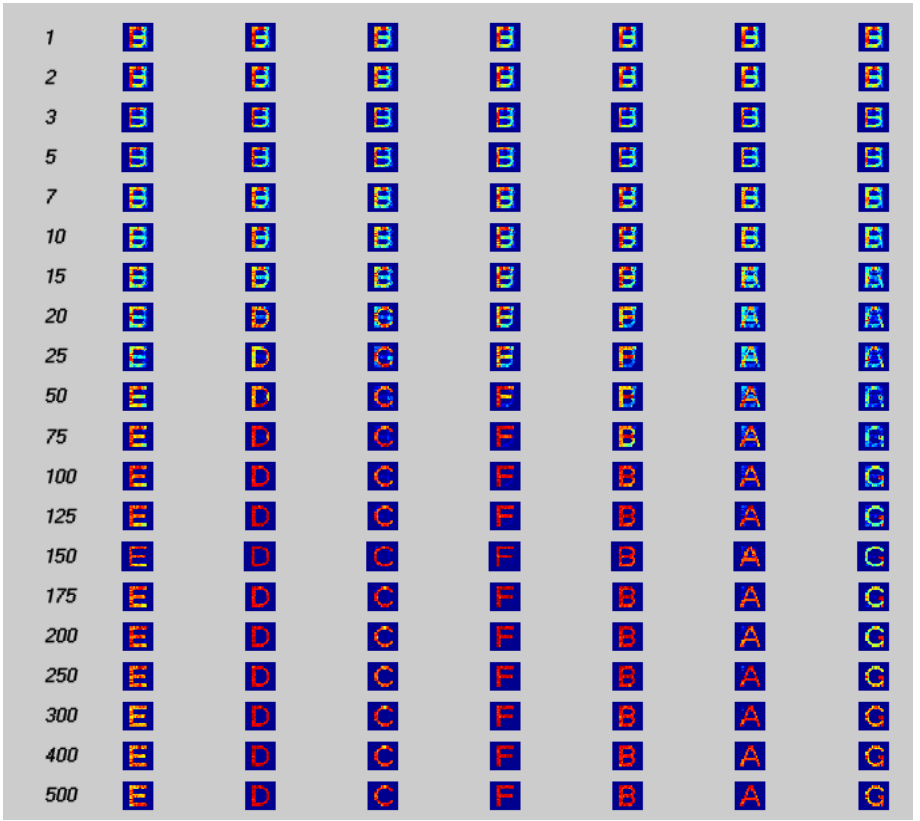


Performance of Recovering Φ

- Normal convergence process of topic-word (Φ) distribution



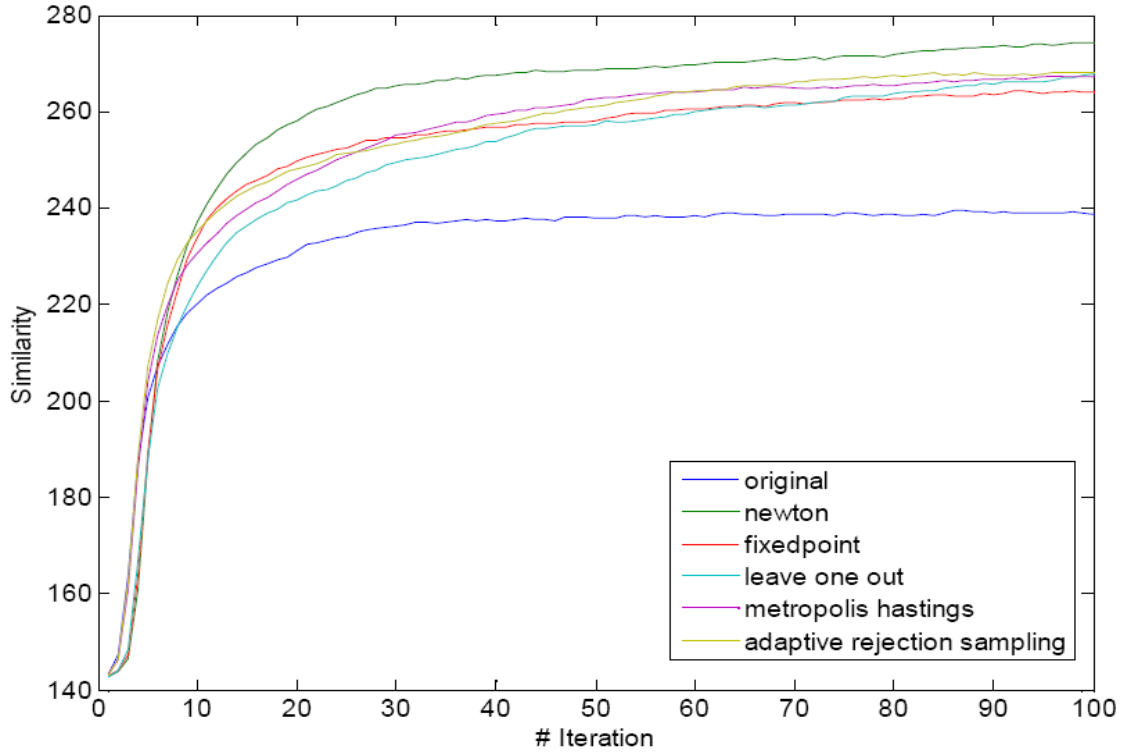
Original Φ



Recovered Φ

Performance of Recovering Φ

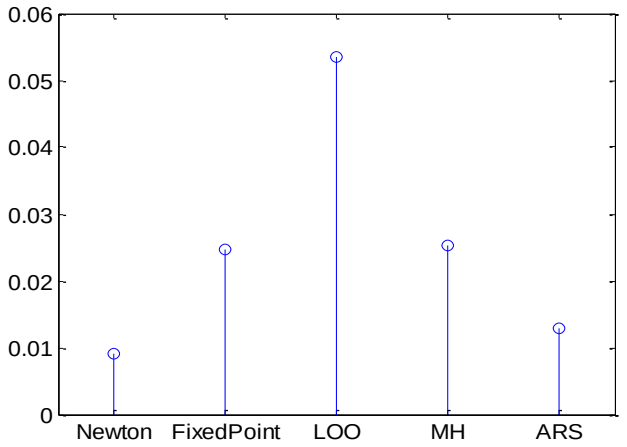
- Topic-item similarity traces over Gibbs sampling steps.



Performance of Recovering α

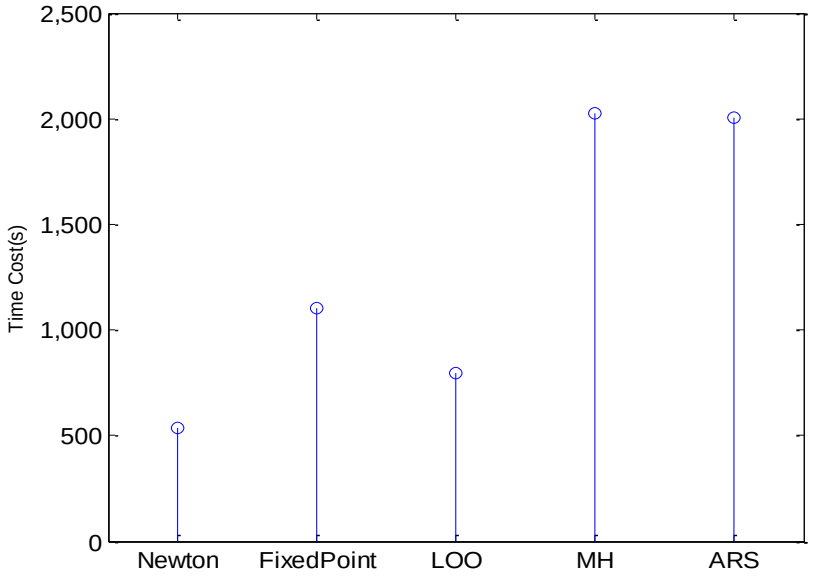
- Average estimated value with different estimation methods & the difference of between estimated value and real value.

	1	2	3	4	5	6	7
Real	0.5	1.5	0.5	1.5	0.5	1.5	0.5
Newton	0.553	1.564	0.656	1.410	0.481	1.643	0.555
FixedPoint	0.488	1.476	0.533	1.374	0.462	1.478	0.892
LOO	0.495	1.195	0.444	1.258	0.576	1.076	0.683
MH	0.563	1.474	0.675	1.625	0.368	1.700	0.764
ARS	0.548	1.321	0.689	1.516	0.534	1.496	0.639



Efficiency Comparison

- Time Consumption comparison of each estimation methods.



Conclusion

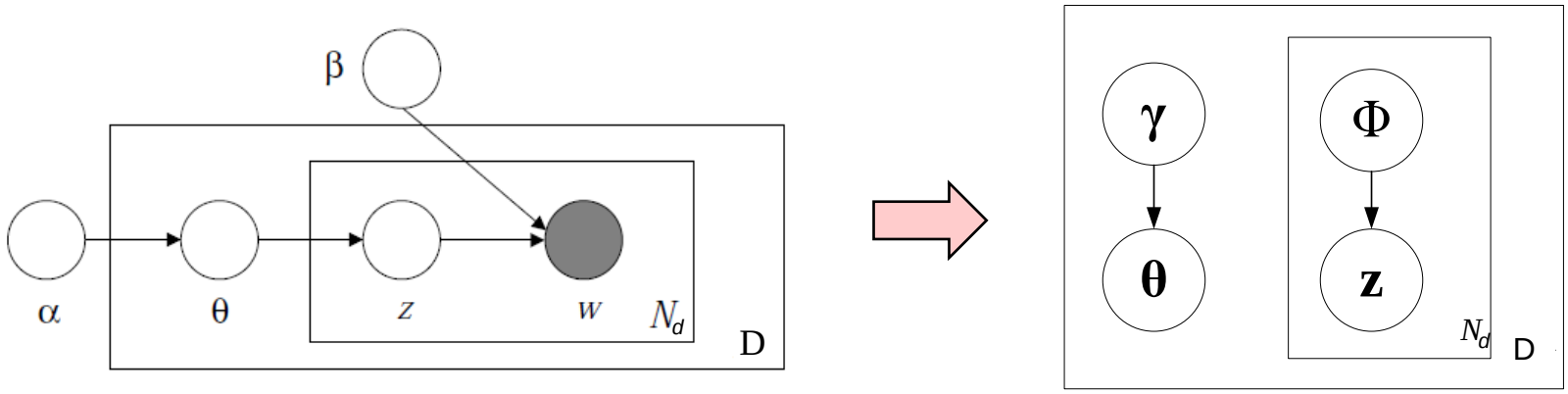
- Newton-Raphson method tends to be the best one in the experiment.
- In two sampling base methods, Metropolis-Hasting random walk is easier to implement, but the performance of adaptive rejection sampling seems a little better.
- In all the methods, the fixed point iteration using leave one out likelihood method doesn’t need to calculate the special gamma function, which make it more scalable.

Variational EM for Inference

- Another alternative algorithm for learning LDA

Recall Variational Inference (Variational Mean-field) Refer to another slides

Variational Inference Derivation for LDA



Variational Inference

- The objective is to compute posterior distribution of the parameters Θ, Φ and hidden variables $\mathcal{Z}$:

$$P(\Theta, \Phi, \mathcal{Z}|\mathcal{W}, \vec{\alpha}, \vec{\beta}) = \frac{P(\Theta, \Phi, \mathcal{Z}, \mathcal{W}|\vec{\alpha}, \vec{\beta})}{P(\mathcal{W}|\vec{\alpha}, \vec{\beta})}$$

- Unfortunately, the normalization contains integrals of Θ and Φ, and summation of $\mathcal{Z}$, making the coupling between Θ and Φ in the summation over $\mathcal{Z}$, which is intractable to solve.
- The basic idea of Variational Inference is use a variational distribution to approximate the original posterior distribution to drop the dependences, i.e.,

$$q(\Theta, \Phi, \mathcal{Z}|\vec{\gamma}, \vec{\lambda}, \vec{\nu}) = q(\Theta|\vec{\gamma})q(\Phi|\vec{\lambda}) \prod_{m=1}^{M} \prod_{n=1}^{N_m} q(z_{m,n}|\nu_{m,n})$$

Lower Bound

- We begin by bounding the log likelihood of a corpus by using Jensen's inequality:

$$
\begin{aligned}
\mathcal{L} &= \log P(\mathcal{W}|\vec{\alpha},\vec{\beta}) = \log \int\int \sum_z P(\Theta,\Phi,\mathcal{Z}|\vec{\alpha},\vec{\beta}) d\Theta d\Phi \qquad (18)\\
&= \log \int\int \sum_z \frac{P(\Theta,\Phi,\mathcal{Z}|\vec{\alpha},\vec{\beta}) q(\Theta,\Phi,\mathcal{Z})}{q(\Theta,\Phi,\mathcal{Z})} d\Theta d\Phi \\
&\geq \int \sum_z q(\Theta,\Phi,\mathcal{Z}) \log P(\Theta,\Phi,\mathcal{Z}) - \int \sum_z q(\Theta,\Phi,\mathcal{Z}) \log q(\Theta,\Phi,\mathcal{Z}) \\
&= E_q[\log P(\Theta,\Phi,\mathcal{Z},\mathcal{W}|\vec{\alpha},\vec{\beta})] - E_q[\log q(\Theta,\Phi,\mathcal{Z})]
\end{aligned}
$$

Continue

$$
\begin{aligned}
\mathcal{L} &= \sum_{m=1}^{M} E_q[\log P(\vec{\theta}_m|\vec{\alpha})] \\
&+ \sum_{k=1}^{K} E_q[\log P(\vec{\phi}_k|\vec{\beta})] \\
&+ \sum_{m=1}^{M}\sum_{n=1}^{N_m} E_q[\log P(z_{m,n}|\vec{\theta}_m)] \\
&+ \sum_{m=1}^{M}\sum_{n=1}^{N_m} E_q[\log P(w_{m,n}|\vec{\phi}_{z_{m,n}})] \\
&- \sum_{m=1}^{M} E_q[\log q(\vec{\theta}_m|\vec{\gamma})] \\
&- \sum_{k=1}^{K} E_q[\log q(\vec{\phi}_k|\vec{\lambda})] \\
&- \sum_{m=1}^{M}\sum_{n=1}^{N_m} E_q[\log q(z_{m,n}|\vec{\nu}_{z_{m,n}})]
\end{aligned}
$$

Continue

$$
\begin{aligned}
\mathcal{L} &= \sum_{m=1}^{M}\left\{\log\Gamma(\sum_{k=1}^{K}\alpha_k)-\sum_{k=1}^{K}\log\Gamma(\alpha_k)+\sum_{k=1}^{K}(\alpha_k-1)\left[\Psi(\gamma_k)-\Psi(\sum_{k=1}^{K}\gamma_k)\right]\right\} \\
&+ \sum_{k=1}^{K}\left\{\log\Gamma(\sum_{t=1}^{V}\beta_t)-\sum_{t=1}^{V}\log\Gamma(\beta_t)+\sum_{t=1}^{V}(\beta_t-1)\left[\Psi(\lambda_t)-\Psi(\sum_{t=1}^{V}\lambda_t)\right]\right\} \\
&+ \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K}\nu_{m,n}^{(k)}(\Psi(\gamma_k)-\Psi(\sum_{k=1}^{K}\gamma_k)) \\
&+ \sum_{m=1}^{M}\sum_{k=1}^{K}\nu_{m,n}^{(k)}\sum_{t=1}^{V}w_{m,n}^{(t)}(\Psi(\lambda_t)-\Psi(\sum_{t=1}^{V}\lambda_t)) \\
&- \sum_{m=1}^{M}\left\{\log\Gamma(\sum_{k=1}^{K}\gamma_k)-\sum_{k=1}^{K}\log\Gamma(\gamma_k)+\sum_{k=1}^{K}(\gamma_k-1)\left[\Psi(\gamma_k)-\Psi(\sum_{k=1}^{K}\gamma_k)\right]\right\} \\
&- \sum_{k=1}^{K}\left\{\log\Gamma(\sum_{t=1}^{V}\lambda_t)-\sum_{t=1}^{V}\log\Gamma(\lambda_t)+\sum_{t=1}^{V}(\lambda_t-1)\left[\Psi(\lambda_t)-\Psi(\sum_{t=1}^{V}\lambda_t)\right]\right\} \\
&- \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K}\nu_{m,n}^{(k)}\log\nu_{m,n}^{(k)}
\end{aligned}
$$

Appendix

We know the expected value of the log of a single probability component under the Dirichlet as:

$$E[\log \theta_i | \alpha] = \Psi(\alpha_i) - \Psi(\sum_{k=1}^{K} \alpha_k)$$

where Ψ is the first derivative of the log Gamma function.

Appendix (Cont.)

$$\sum_{m=1}^{M}\sum_{n=1}^{N_m} E_q[\log P(z_{m,n}|\vec{\theta}_m)] \tag{21}$$

$$= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\int\int q(z_{m,n},\vec{\theta}_m|\mu_{m,n},\vec{\gamma})\log P(z_{m,n}|\vec{\theta}_m)d\vec{\theta}_m dz_{m,n}$$

$$= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K} q(z_{m,n}=k|\nu_{m,n})\int q(\theta_{m,k}|\gamma_k)\log P(z_{m,n}=k|\theta_{m,k})d\theta_{m,k}$$

$$= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K} \nu_{m,n}^{(k)} E_q[\log\theta_{m,k}|\vec{\gamma}]$$

$$= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{K} \nu_{m,n}^{(k)}(\Psi(\gamma_k)-\Psi(\sum_{k=1}^{K}\gamma_k))$$

Appendix (Cont.)

$$\sum_{m=1}^{M}\sum_{n=1}^{N_m} E_q[\log P(w_{m,n}|\vec{\phi}_{z_{m,n}})] \tag{22}$$

$$= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\int\int q(z_{m,n},\vec{\phi}_k|\nu_{m,n},\vec{\lambda}_k)\log P(w_{m,n}|\vec{\phi}_{z_{m,n}})d\vec{\phi}_{z_{m,n}}dz_{m,n}$$

$$= \sum_{m=1}^{M}\sum_{k=1}^{K} q(z_{m,n}=k|\nu_{m,n})\sum_{t=1}^{V} w_{m,n}^{(t)}\int q(\phi_{k,t}|\lambda_{k,t})\log P(w_{m,n}=t|\phi_{k,t})d\phi_{k,t}$$

$$= \sum_{m=1}^{M}\sum_{k=1}^{K}\nu_{m,n}^{(k)}\sum_{t=1}^{V} w_{m,n}^{(t)}E_q(\log\phi_{k,t}|\vec{\lambda})$$

$$= \sum_{m=1}^{M}\sum_{k=1}^{K}\nu_{m,n}^{(k)}\sum_{t=1}^{V} w_{m,n}^{(t)}(\Psi(\lambda_t)-\Psi(\sum_{t=1}^{V}\lambda_t))$$

Solve Variational Parameters

Let

$$G(\gamma_k) = \psi(\gamma_k) - \psi\left(\sum_{k=1}^{K} \gamma_k\right), G(\lambda_t) = \psi(\lambda_t) - \psi\left(\sum_{t=1}^{V} \lambda_t\right)$$

Where

$$\psi(x) = \frac{d}{dx}\log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$

So

$$G(\gamma_k) = \psi(\gamma_k) - \psi\left(\sum_{k=1}^{K} \gamma_k\right) = \frac{\Gamma'(\gamma_k)}{\Gamma(\gamma_k)} - \frac{\Gamma'(\sum_k \gamma_k)}{\Gamma(\sum_k \gamma_k)}$$

Rearrange the Lower Bound

$$L = \sum_{m=1}^{M} \log\Gamma\left(\sum_{k=1}^{K} \alpha_k\right) - \log\Gamma\left(\sum_{k=1}^{K} \gamma_k\right) - \sum_{k=1}^{K} \log\Gamma(\alpha_k) + \sum_{k=1}^{K} \log\Gamma(\gamma_k) + \sum_{k=1}^{K} (\alpha_k - \gamma_k)\, G(\gamma_k)$$

$$+ \sum_{k=1}^{K} \log\Gamma\left(\sum_{t=1}^{V} \beta_t\right) - \log\Gamma\left(\sum_{t=1}^{V} \lambda_t\right) - \sum_{t=1}^{V} \log\Gamma(\beta_t) + \sum_{t=1}^{V} \log\Gamma(\lambda_t) + \sum_{t=1}^{V} (\beta_t - \lambda_t) G(\lambda_t)$$

$$+ \sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{k=1}^{K} \nu_{m,n}^{(k)} \left(G(\gamma_k) - \log \nu_{m,n}^{(k)}\right)$$

$$+ \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{t=1}^{V} \nu_{m,n}^{(k)} \omega_{m,n}^{(t)} G(\lambda_t)$$

Partial Derivative of $\gamma_{m,k}$

$$\frac{\partial L}{\partial \gamma_{m,k}} = G'(\gamma_k)\left(\alpha_k - \gamma_k + \sum_{n=1}^{N_m} \nu_{m,n}^{(k)}\right) - \frac{\Gamma'(\sum_k \gamma_k)}{\Gamma(\sum_k \gamma_k)} + \frac{\Gamma'(\gamma_k)}{\Gamma(\gamma_k)} - G(\gamma_k)$$

$$= G'(\gamma_k)\left(\alpha_k - \gamma_k + \sum_{n=1}^{N_m} \nu_{m,n}^{(k)}\right) + G(\gamma_k) - G(\gamma_k)$$

$$= G'(\gamma_k)\left(\alpha_k - \gamma_k + \sum_{n=1}^{N_m} \nu_{m,n}^{(k)}\right) = 0$$

Since $G'(\gamma_k) \neq 0$, we obtain that

$$\gamma_{k,m} = \alpha_k + \sum_{n=1}^{N_m} \nu_{m,n}^{(k)}$$

Partial Derivative of $\lambda_{k,t}$

$$\frac{\partial L}{\partial \lambda_{k,t}} = G'(\lambda_t)\left(\beta_t - \lambda_t + \sum_{m=1}^{M}\sum_{n=1}^{N_m} \nu_{m,n}^{(k)}\omega_{m,n}^{(t)}\right) - \frac{\Gamma'(\sum_t \lambda_t)}{\Gamma(\sum_t \lambda_t)} + \frac{\Gamma'(\lambda_t)}{\Gamma(\lambda_t)} - G(\lambda_t)$$

$$= G'(\lambda_t)\left(\beta_t - \lambda_t + \sum_{m=1}^{M}\sum_{n=1}^{N_m} \nu_{m,n}^{(k)}\omega_{m,n}^{(t)}\right) = 0$$

Since $G'(\lambda_t) \neq 0$, we obtain that

$$\lambda_{k,t} = \beta_t + \sum_{m=1}^{M}\sum_{n=1}^{N_m} \nu_{m,n}^{(k)} w_{m,n}^{(t)}$$

Partial Derivative of $\nu_{m,n}^{(k)}$

$$\frac{\partial L}{\partial \nu_{m,n}^{(k)}} = G(\gamma_k) - \log \nu_{m,n}^{(k)} - \frac{\nu_{m,n}^{(k)}}{\nu_{m,n}^{(k)}} + \sum_{t=1}^{V} \omega_{m,n}^{(t)} G(\lambda_t) = G(\gamma_k) - \log \nu_{m,n}^{(k)} - 1 + \sum_{t=1}^{V} \omega_{m,n}^{(t)} G(\lambda_t)$$

$$= 0$$

$$\nu_{m,n}^{(k)} = e^{\log \nu_{m,n}^{(k)}} = \exp\left(G(\gamma_k) + \sum_{t=1}^{V} \omega_{m,n}^{(t)} G(\lambda_t) - 1 \right) \propto \exp\left(G(\gamma_k) + \sum_{t=1}^{V} \omega_{m,n}^{(t)} G(\lambda_t) \right)$$

$$= \exp\left[\psi(\gamma_k) - \psi\left(\sum_{\mathrm{k}=1}^{\mathrm{K}} \gamma_k \right) + \sum_{t=1}^{V} \left(\omega_{m,n}^{(t)} \left(\psi(\lambda_t) - \psi\left(\sum_{\mathrm{t}=1}^{\mathrm{V}} \lambda_t \right) \right) \right) \right]$$

Algorithm

1. Repeat
2. Initialize $\nu_{m,n}^{k} = 1/K$ for all m, n, k
3. Initialize $\gamma_{m,k} = \alpha_k + \frac{N_m}{K}$ for all m, k
4. Initialize $\lambda_{k,t} = \beta_t + \frac{N}{V}$ for all k, t, where N is the number of words in corpus.
5. For $m = 1$ to M:
6. For $n = 1$ to N_m:
7. For $k = 1$ to K:
8. $$\nu_{m,n}^{(k)} \propto \exp\left[\left(\Psi(\gamma_k) - \Psi(\sum_{k=1}^{K} \gamma_k)\right) + \sum_{t=1}^{V} w_{m,n}^{(t)} \left(\Psi(\lambda_t) - \Psi(\sum_{t=1}^{V} \lambda_t)\right)\right]$$
9. Normalize $\nu_{m,n}^{(k)}$ to sum to 1.
10. Until Convergence.
11. Update $\gamma_{m,k} = \alpha_k + \sum_{n=1}^{N_m} \nu_{m,n}^{(k)}$ for all m, k
12. Update $\lambda_{k,t} = \beta_t + \sum_{m=1}^{M} \sum_{n=1}^{N_m} \nu_{m,n}^{(k)} w_{m,n}^{(t)}$ for all k, t

Collapsed Variational Inference Derivation for LDA

Yee Whye Teh, David Newman, and Max Welling. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocationt. NIPS 2007

Collapsed Variational Inference

- The basic idea of Collapsed Variational Inference is to first integrate out Θ and Φ, and then only use one distribution variational parameter ν for z, i.e., $q(\mathcal{Z}|\nu)$.
- By introducing independence assumption of $\{z_{m,n}\}$, we factorize $q(\mathcal{Z}|\nu)$ into $\prod_{m=1}^{M}\prod_{n=1}^{N_m} q(z_{m,n}|\nu_{m,n})$.
- By setting $\frac{\partial \mathrm{KL}(q(\mathcal{Z}|\nu)||P(\mathcal{Z}|\mathcal{W}))}{\partial \nu_{m,n}^{(k)}} = 0$, we get (Refer to PRML chapter 10.1):

$$\nu_{m,n}^{(k)} = \frac{\exp\left[E_q\left(\ln P(z_{m,n}=k, \mathcal{Z}^{\neg m,n}, \mathcal{W})\right)\right]}{\sum_{k'=1}^{K}\exp\left[E_q\left(\ln P(z_{m,n}=k', \mathcal{Z}^{\neg m,n}, \mathcal{W})\right)\right]}$$

Collapsed Variational Inference (Cont.)

$$\nu_{m,n}^{(k)} = \frac{\exp\left[E_q\left(\ln P(z_{m,n}=k, \mathcal{Z}^{\neg m,n}, \mathcal{W})\right)\right]}{\sum_{k'=1}^{K}\exp\left[E_q\left(\ln P(z_{m,n}=k', \mathcal{Z}^{\neg m,n}, \mathcal{W})\right)\right]}$$

- For any value of $z_{m,n}$, the joint probability $P(\mathcal{Z}^{\neg m,n}, \mathcal{W})$ is a constant. Thus we divide both the nominator and the denominator by $P(\mathcal{Z}^{\neg m,n}, \mathcal{W})$ and get:

$$\nu_{m,n}^{(k)} = \frac{\exp\left[E_q\left(\ln \frac{P(z_{m,n}=k, \mathcal{Z}^{\neg m,n}, \mathcal{W})}{P(\mathcal{Z}^{\neg m,n}, \mathcal{W})}\right)\right]}{\sum_{k'=1}^{K}\exp\left[E_q\left(\ln \frac{P(z_{m,n}=k', \mathcal{Z}^{\neg m,n}, \mathcal{W})}{P(\mathcal{Z}^{\neg m,n}, \mathcal{W})}\right)\right]}$$

- According to the collapsed joint probability in Eq. 16, we get :

$$\frac{P(z_{m,n}=k, \mathcal{Z}^{\neg m,n}, \mathcal{W})}{P(\mathcal{Z}^{\neg m,n}, \mathcal{W})} \propto \frac{n_{\cdot,k,t}^{\neg m,n} + \beta_t}{\sum_{t=1}^{V} n_{\cdot,k,t}^{\neg m,n} + \beta_t}(n_{m,k,\cdot}^{\neg m,n} + \alpha_k)$$

Collapsed Variational Inference (Cont.)

Then we get :

$$\nu_{m,n}^{(k)} \approx \exp E_q \left\{ \ln \frac{n_{\cdot,k,t}^{\neg m,n} + \beta_t}{\sum_{t=1}^{V} n_{\cdot,k,t}^{\neg m,n} + \beta_t} + \ln(n_{m,k,\cdot}^{\neg m,n} + \alpha_k) \right\}$$

- We further approximate the function $E[\ln(x)]$ by first-order Taylor expansion[Teh, 06]:

$$E_q \ln(n_{m,k,\cdot}^{\neg m,n} + \alpha_k) \approx \ln\left(\alpha_k + E_q(n_{m,k,\cdot}^{\neg m,n})\right)$$

  where $E_q(n_{m,k,\cdot}^{\neg m,n})$ is the aggregation of $\nu^{\neg m,n}$, i.e., $E_q(n_{m,k,\cdot}^{\neg m,n}) = \sum_{n' \neq n} \nu_{m,n'}^{(k)}$.
- Finally we get:

$$\nu_{m,n}^{(k)} \propto \frac{E_q(n_{\cdot,k,t}^{\neg m,n}) + \beta_t}{\sum_{t=1}^{V} E_q(n_{\cdot,k,t}^{\neg m,n}) + \beta_t} (E_q(n_{m,k,\cdot}^{\neg m,n}) + \alpha_k) \qquad (23)$$

Solve Original Parameters

$$\theta_{m,k} = \frac{E_q(n_{m,k,\cdot}) + \alpha_k}{\sum_{k=1}^{K} E_q(n_{m,k,\cdot}) + \alpha_k}, \phi_{k,t} = \frac{E_q(n_{\cdot,k,t}) + \beta_t}{\sum_{t=1}^{V} E_q(n_{\cdot,k,t}) + \beta_t} \quad (24)$$

Algorithm

1. Randomly initialize $\nu_{m,n}^{(k)}$ for each m, n, k;
2. Randomly initialize $n_{m,k,\cdot}$ and $n_{\cdot,k,t}$ by adding corresponding $\nu_{m,n}^{(k)}$ together;
3. Repeat
4. For $m = 1$ to M
5. For $n = 1$ to N_m
6. $n_{m,k,\cdot} = n_{m,k,\cdot} - \nu_{m,n}^{(k)}$
7. $n_{\cdot,k,t} = n_{\cdot,k,t} - \nu_{m,n}^{(k)}$
8. Calculate $\nu_{m,n}^{(k)}$ according to Eq. (23).
9. $n_{m,k,\cdot} = n_{m,k,\cdot} + \nu_{m,n}^{(k)}$
10. $n_{\cdot,k,t} = n_{\cdot,k,t} + \nu_{m,n}^{(k)}$
11. Until convergence;
12. Update Θ and Φ according to Eq. (24).

Applications

Document Modeling

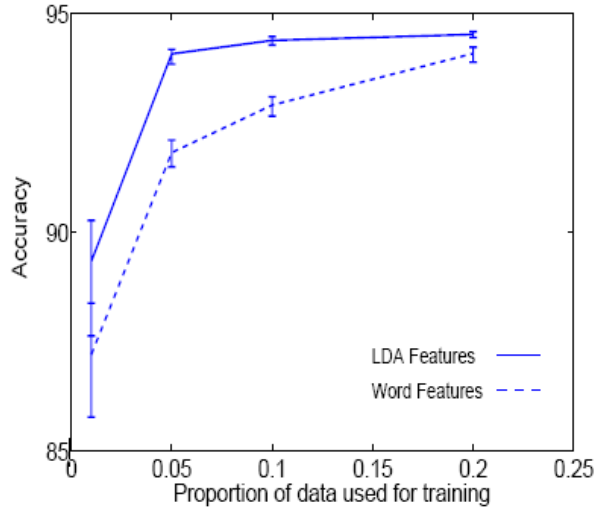
- Perplexity is used to indicate the generalization performance of a method.
- Specifically, we estimate a document modeling and use this model to describe the new data set.

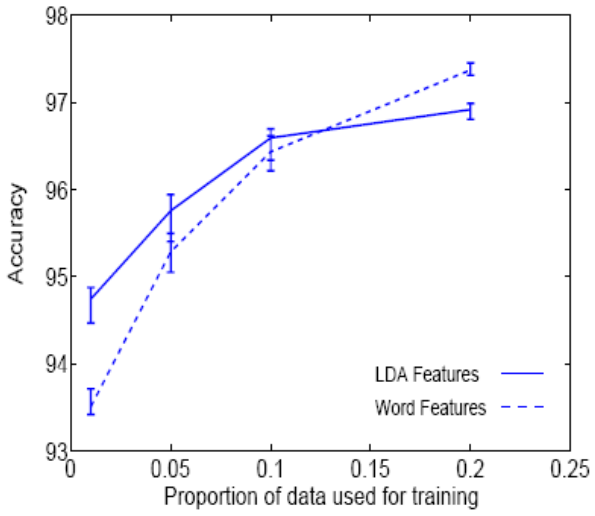
$$perplexity(D_{test}) = \exp\{-\frac{\sum_{m=1}^{M} \log p(\mathbf{w}_m)}{\sum_{m=1}^{M} N_m}\}$$

- LDA outperforms the other models including pLSI, Smoothed Unigram, and Smoothed Mixt. Unigrams.

Document Classification

- We can use the LDA model results as the features for classifiers. In this way, say 50 topics, we can reduce the feature space by 99.6%.
- The experimental results show that such feature reduction may decrease the accuracy only a little.





Document Clustering

- We can also use the LDA model results as the features for clustering algorithms.

Collaborative Filtering

- We can learn a model on a fully observed set of users. Then for each unobserved user, we are shown all but one of the movies preferred by that user and are asked to predict what the held-out movie is.
- Precisely, define the predictive perplexity on M test uses as:

$$predictive-perplexity(D_{test}) = \exp\{-\frac{\sum_{m=1}^{M} \log p(w_{m,N_m} \mid \mathbf{w}_{m,1:N_m-1})}{M}\}$$

Our Work related to Topic Model

Topic Browser

Topic 13: Computing programming



Commun. ACM , IFIP Congress , IFIP Congress (2) , A
Comput. Surv. , Advances in Computers , IFIP Congre

Topic 14: Computer networks and communication



INFOCOM , IEEE/ACM Trans. Netw. , IEEE Journal on
Areas in Communications , SIGCOMM , Computer Netw
ICNP

Topic 15: Automatic reasoning



CADE , J. Autom. Reasoning , J. Log. Comput. , JELIA
Math. Artif. Intell. , KR

Topic 16: Knowledge discovery and data mining



KDD , PAKDD , ICDM , PKDD , IEEE Trans. Knowl. Da
, SDM

Topic 17: Human computer interaction



CHI Extended Abstracts , CHI , Interactions , INTERA
Interacting with Computers , CHI Conference Compani

Topic 18: Evolvable Hardware



Evolvable Hardware , ICES , AHS , BioADIT , Conf.
Computing Frontiers , UC

200 topics have been discovered automatically from the academic network

Hot words in **Topic 16: Knowledge discovery and data mining**

mining:	0.041622	**data:**	0.041032
clustering:	0.014825	**discovery:**	0.010956
learning:	0.010630	**classification:**	0.010569
rules:	0.008899	**association:**	0.008655
patterns:	0.007962	**knowledge:**	0.007820
approach:	0.007636	**algorithm:**	0.007596
frequent:	0.007473	**based:**	0.007311
analysis:	0.007209	**databases:**	0.006211
efficient:	0.006129	**web:**	0.005417
text:	0.004908	**detection:**	0.004887



Hot conferences/jounals in **Topic 16: Knowledge discovery and data mining**

KDD: 0.143204
PAKDD: 0.106370
ICDM: 0.100902
PKDD: 0.082549
IEEE Trans. Knowl. Data Eng.: 0.064238
SDM: 0.047920
Intell. Data Anal.: 0.044338
SIGKDD Explorations: 0.037705
IDA: 0.033127
Knowl. Inf. Syst.: 0.030880
Data Min. Knowl. Discov.: 0.027447
DaWaK: 0.024374
CIKM: 0.022848
ICDM Workshops: 0.022806
Discovery Science: 0.020771
MLDM: 0.017677
CIDM: 0.016194
J. Intell. Inf. Syst.: 0.008055
KDID: 0.007081
DMKD: 0.007059

Active authors in **Topic 16: Knowledge discovery and data mining**


Jiawei Han


Eamonn J. Keogh


Philip S. Yu


Mohammed Javeed Zaki


Ming-Syan Chen


Wei Wang


Srinivasan Parthasarathy



Charu C. Aggarwal


Qiang Yang


Shusaku Tsumoto

Data Mining
data stream
Data Mining Techniques
Evolving Data Streams
Mining Time Series Data

Association Rules
Association Mining
Data Streams
Mining Association Rules
data mining application

Topic Trend in **Topic 16: Knowledge discovery and data mining**



Publications in **Topic 16: Knowledge discovery and data mining**

Data Mining Techniques for Associations, Clustering and Classification
Authors: Charu C. Aggarwal, Philip S. Yu.
In Proceedings: PAKDD
Published year: 1999

Data Mining for the Web
Authors: Myra Spiliopoulou.
In Proceedings: PKDD
Published year: 1999

On string classification in data streams
Authors: Charu C. Aggarwal, Philip S. Yu.
In Proceedings: KDD
Published year: 2007

Matching in Time Series Databases
Authors: Eamonn J. Keogh, Padhraic Smyth.
In Proceedings: KDD
Published year: 1997

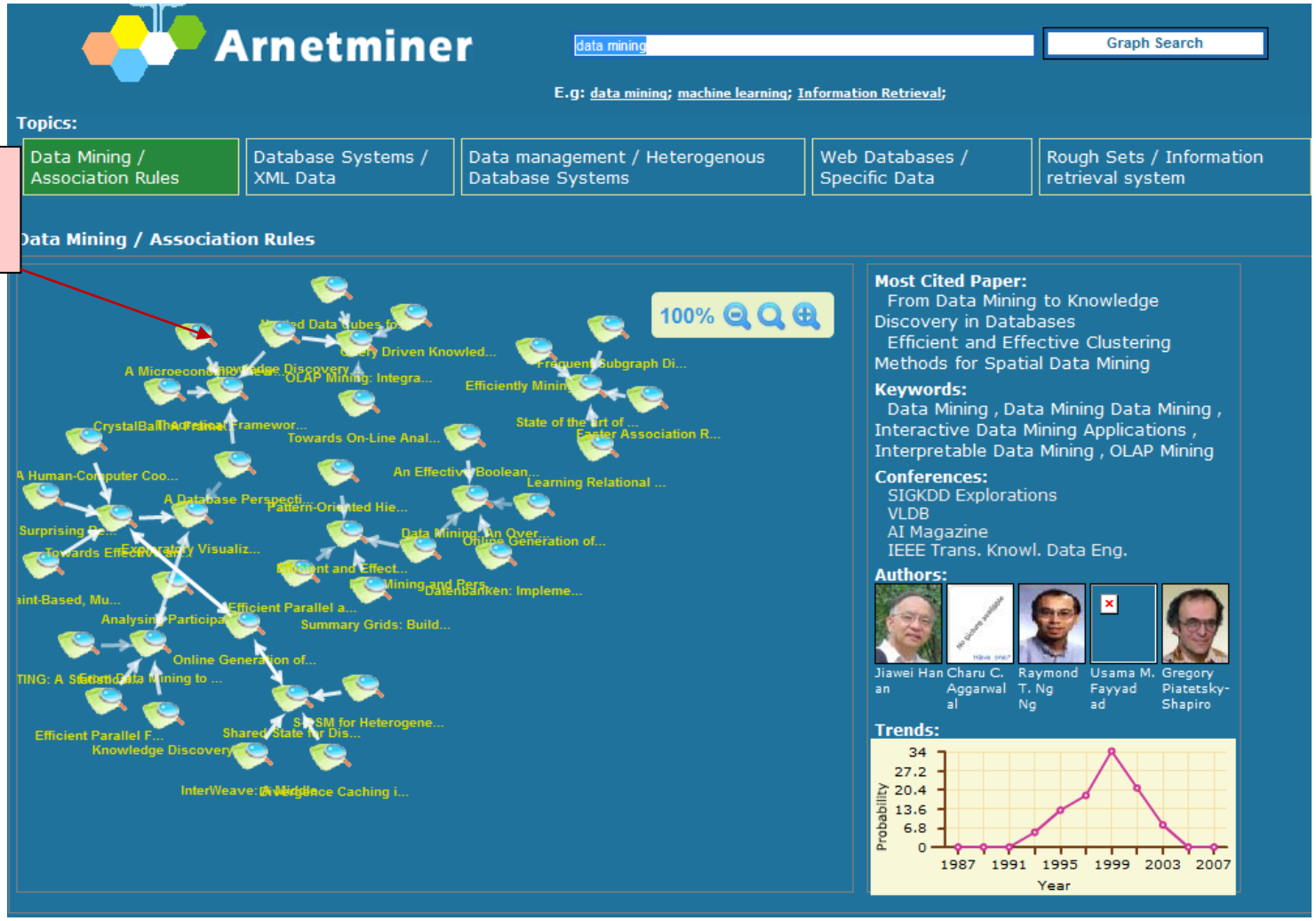
Knowledge Discovery and Data Mining: Towards a Unifying Framework
Authors: Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth.
In Proceedings: KDD
Published year: 1996

Efficiently mining frequent trees in a forest
Authors: Mohammed Javeed Zaki.
In Proceedings: KDD
Published year: 2002

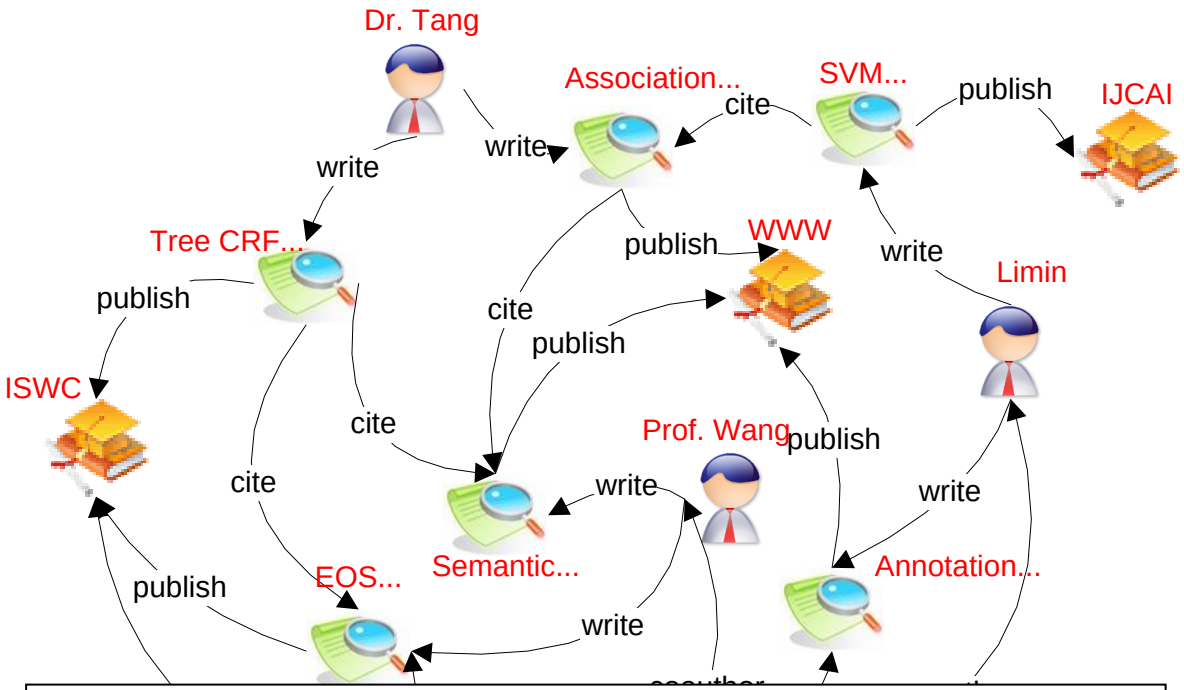
Mining of Evolving Data Streams with Privacy Preservation
Authors: Philip S. Yu.
In Proceedings: PAKDD

Graph Search

Sub graphs



Heterogeneous Academic Networks



In Academic Research Network

Heterogeneous objects:

Paper, Person, Conf./Journal

Relationships:

- Conf./Journal *publish* paper
- Paper *cite* paper
- Person *write* paper
- Person is *PC member* of Conf./Journal
- Person is *coauthor* of person

Challenges:

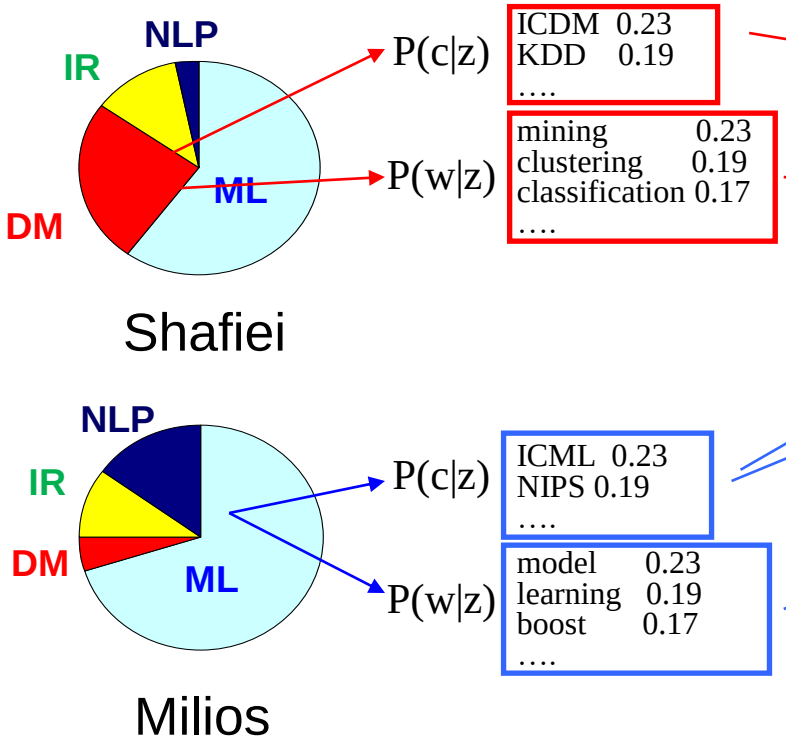
- How to model the heterogeneous objects in a unified model?

- How to model the relationships between the heterogeneous data?

- How to apply the proposed model to different applications?

Work1: Modeling social network --Generative Story

- Generative process



Paper

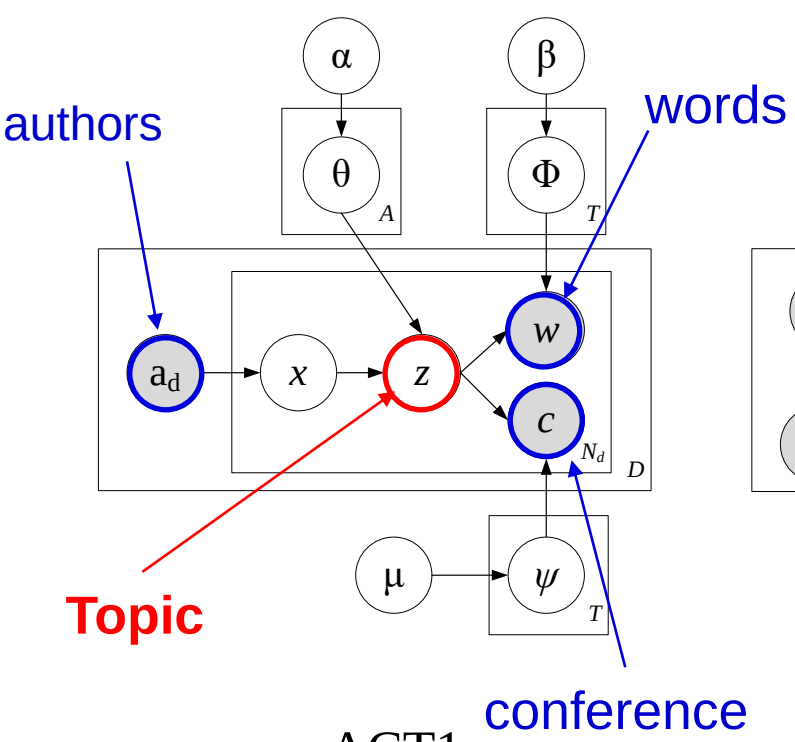
Latent Dirichlet Co-clustering

Shafiei and Milios

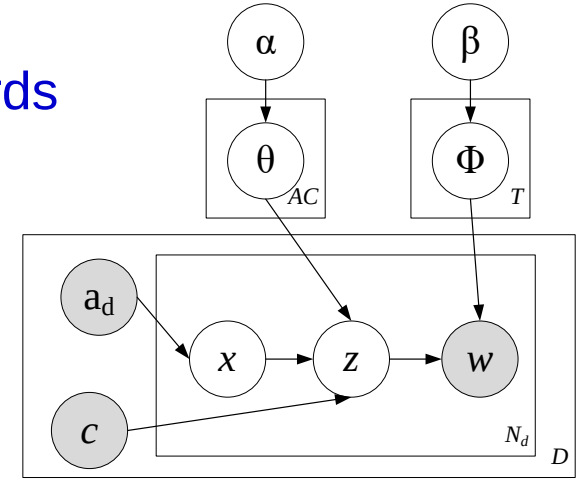
ICDM ***NIPS***

We present a generative model for clustering documents and terms. Our model is a four hierarchical bayesian model. We present efficient inference techniques based on Markow Chain Monte Carlo. We report results in document modeling, document and terms clustering …

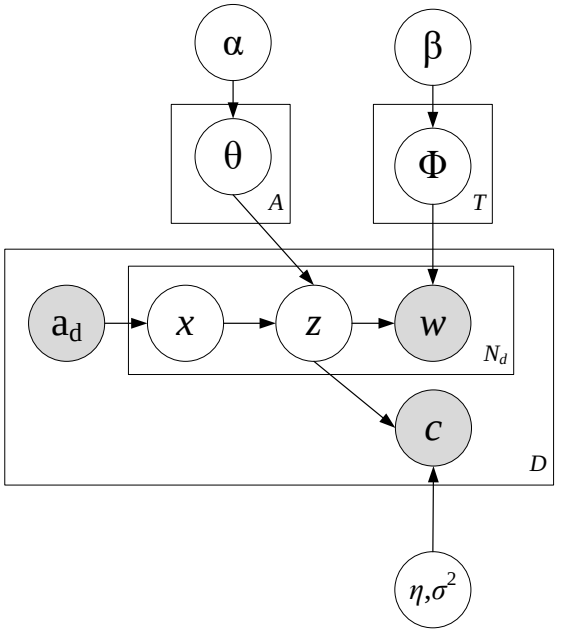
Modeling Multiple Objects in Academic network (KDD'08)



ACT1

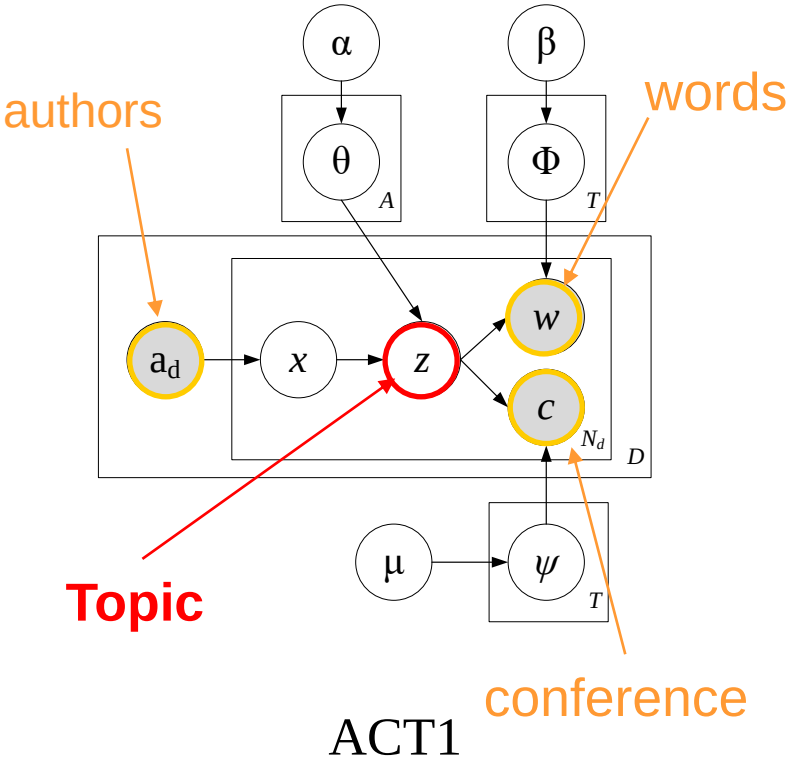


ACT2



ACT3

ACT Model 1



ACT1

Generative process:

1. For each topic z, draw ϕ_z and ψ_z respectively from Dirichlet priors β_z and μ_z;

2. For each word w_{di} in document d:

   - draw an author x_{di} from $\mathbf{a}_d$ uniformly;
   - draw a topic z_{di} from a multinomial distribution $\theta_{x_{di}}$ specific to author x_{di}, where θ is generated from a Dirichlet prior α;
   - draw a word w_{di} from multinomial $\phi_{z_{di}}$;
   - draw a conference stamp c_{di} from multinomial $\psi_{z_{di}}$.

$$P(z_{di}, x_{di} | \mathbf{z}_{-di}, \mathbf{x}_{-di}, \mathbf{w}, \mathbf{c}, \alpha, \beta, \mu) \propto \frac{m^{-di}_{x_{di} z_{di}} + \alpha_{z_{di}}}{\sum_z (m^{-di}_{x_{di} z} + \alpha_z)} \frac{n^{-di}_{z_{di} w_{di}} + \beta_{w_{di}}}{\sum_v (n^{-di}_{z_{di} v} + \beta_v)} \frac{n^{-d}_{z_{di} c_d} + \mu_{c_d}}{\sum_c (n^{-d}_{z_{di} c} + \mu_c)}$$

Expertise Search Results

Arnetminer data:
- 14,134 authors
- 10,716 papers
- 1,434 confs/journals

Evaluation measures:
- pooled relevance
- + human judgement

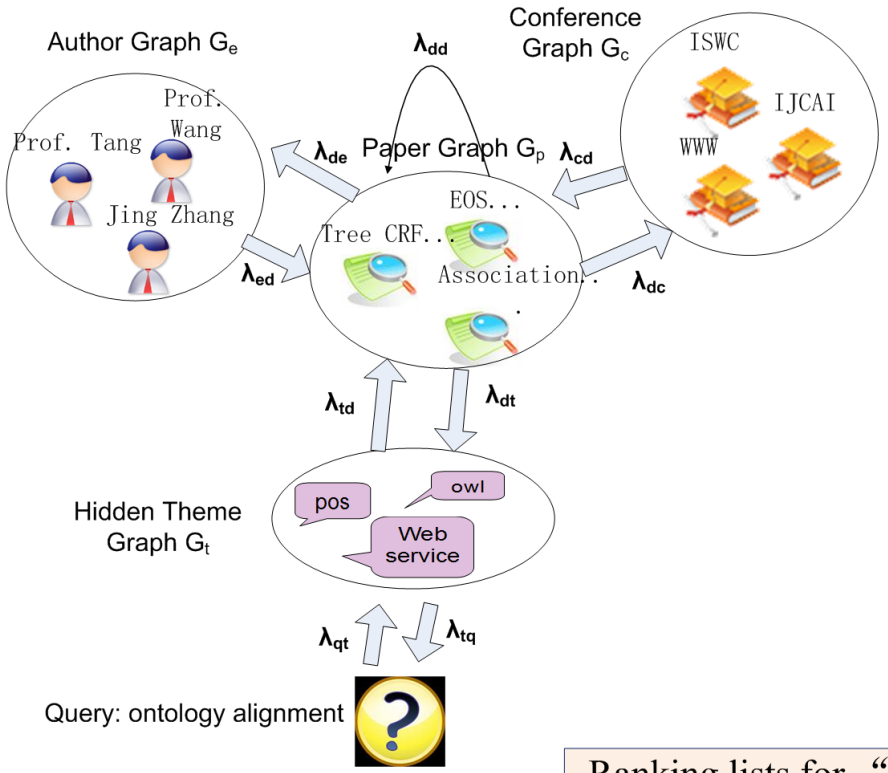
Baselines:
- Language Model (LM)
- LDA
- Author Topic (AT)

Method	Object	P@5	P@10	P@20	R-pre	MAP
LM	Paper	40.0	38.6	37.1	10.0	46.4
	Author	65.7	44.3	25.0	58.8	73.4
	Conference	51.4	32.9	21.4	47.6	63.1
	Average	52.4	38.6	27.9	38.8	61.0
LDA	Paper	31.4	48.6	42.9	13.5	45.8
AT	Paper	42.9	48.6	42.9	13.1	49.3
	Author	82.9	45.7	25.7	73.5	78.1
	Average	62.9	47.1	34.3	43.3	63.7
ACT1	Paper	42.9	45.7	43.6	16.6	51.0
	Author	91.4	50.0	26.4	80.0	89.6
	Conference	62.9	41.4	23.6	60.7	72.3
	Average	**65.7**	**45.7**	**31.2**	**52.4**	**71.0**
ACT2	Paper	42.9	47.1	39.3	15.0	47.7
	Author	74.3	50.0	25.7	69.4	80.1
	Conference	54.3	41.4	22.1	54.2	63.9
	Average	57.1	46.2	29.1	46.2	63.9
ACT3	Paper	42.9	38.6	41.4	17.1	47.0
	Author	71.4	47.1	25.7	70.0	78.7
	Conference	57.1	38.6	23.6	58.3	65.7
	Average	57.1	41.4	30.2	48.5	63.8

Work2: Integration of Links

—A Unified Probabilistic Ranking Model (ICDM'08)

Unified Probabilistic Model Combing Random Walk over Heterogeneous Network



- The unified model integrates different kinds of objects, called **heterogeneous graph**.
- The weights $p(d_i|d_j)$, $p(e_i|d_i)$, $p(c_i|d_i)$, $p(d_i|e_i)$, and $p(d_i|c_i)$ are similarly defined as uniform probability like PageRank.
- The **key difference** here is that the unified model integrates hidden theme nodes and a virtual query node into the heterogeneous graph. Hence, $p(\theta_i|d_i)$, $p(\theta_i|q)$, $p(d_i|\theta_i)$, and $p(q|\theta_i)$ are calculated based on Mixture Model.
- The rank vector is: (where M is the transition probability matrix)

$$r = Ar, A = M^T$$

- Finally, the model can rank different objects simultaneously.

$$r(d_i) = \lambda_{dd} \times \sum_{(d_j,d_i)\in E_{dd}} p(d_i \mid d_j)\cdot r(d_j) + \lambda_{ed} \times \sum_{(e_i,d_i)\in E_{de}} P(d_i \mid e_i)\cdot r(e_i)$$
$$+ \lambda_{cd} \times \sum_{(c_i,d_i)\in E_{dc}} P(d_i \mid c_i)\cdot r(c_i) + \lambda_{td} \times \sum_{(\theta_i,d_i)\in E_{dt}} P(d_i \mid \theta_i)\cdot r(\theta_i)$$

$$r(e_i) = \lambda_{de} \times \sum_{(d_j,e_i)\in E_{de}} p(e_i \mid d_j)\cdot r(d_j)$$

$$r(c_i) = \lambda_{dc} \times \sum_{(d_j,c_i)\in E_{dc}} p(c_i \mid d_j)\cdot r(d_j)$$

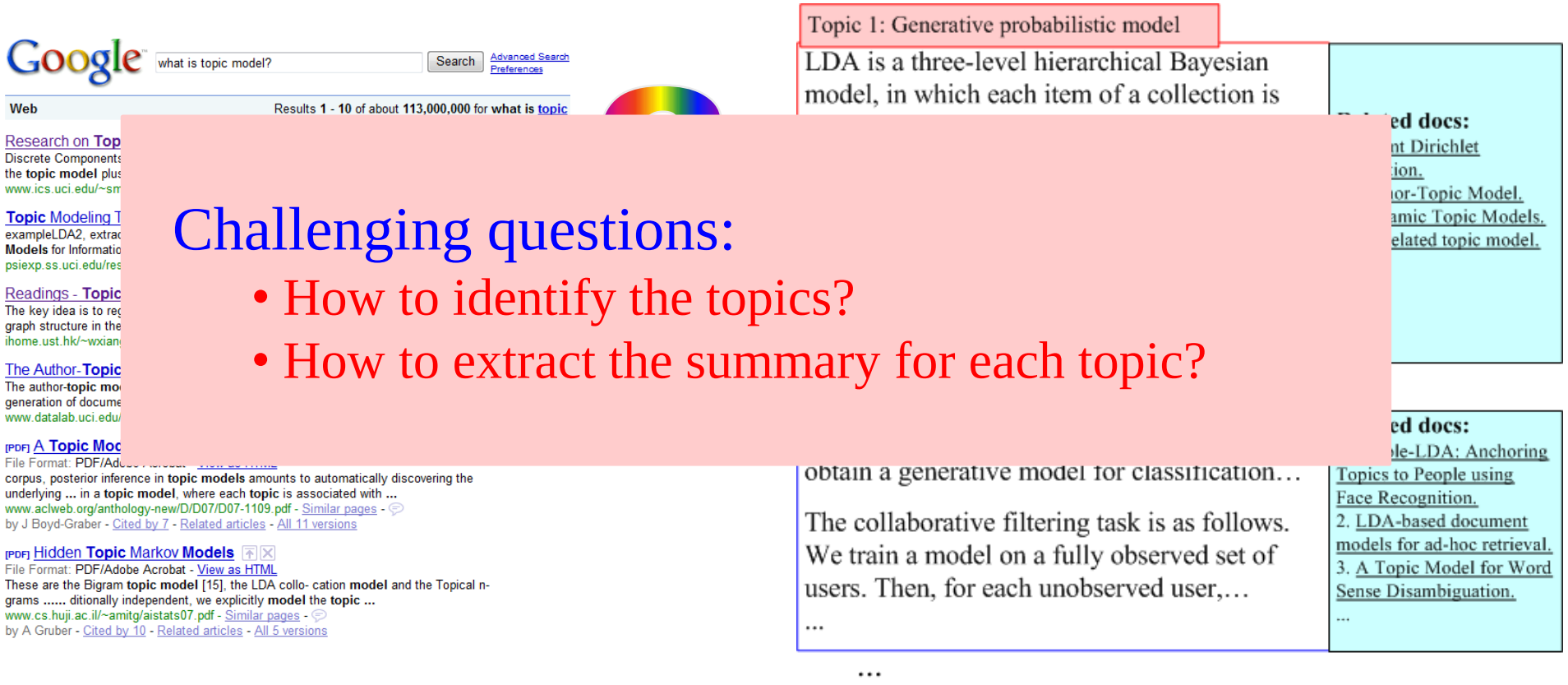
Ranking lists for "Information Extraction"

Persons	Conferences	Papers
• Raymond J. Mooney • Andrew McCallum • Fabio Ciravegna • Nicholas Kushmerick	• AAAI • ACL • IJCAI • KDD	• Conditional random fields: Probabilistic models for segmenting and labeling sequence data. • Maximum entropy markov models for information extraction and segmentation. • Wrapper induction for information extraction. • Learning information extraction rules for semi-structured and free text

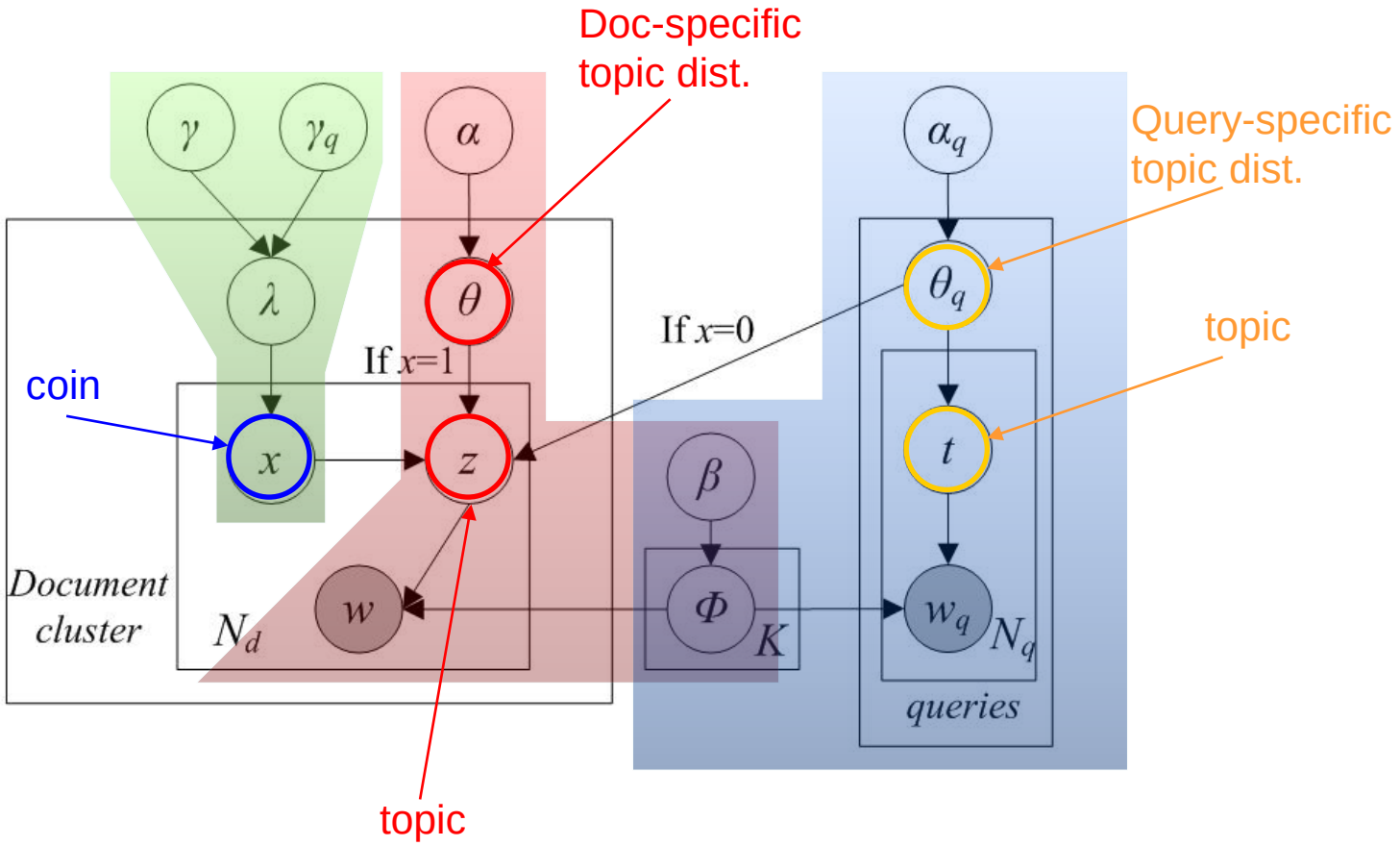
Expertise Search Results

Method	Object	P@5	P@10	P@20	R-pre	MAP
BM25	Paper	42.9	45.7	41.4	12.0	47.2
	Author	77.1	47.1	26.4	67.5	85.5
	Conference	51.4	38.6	22.9	48.8	66.0
	Average	57.1	43.8	30.2	42.8	66.2
LM	Paper	40.0	38.6	37.1	10.0	46.4
	Author	65.7	44.3	25.0	58.8	73.4
	Conference	51.4	32.9	21.4	47.6	63.1
	Average	52.4	38.6	27.9	38.8	61.0
LM+RW1	Paper	62.9	55.7	44.3	12.9	65.3
	Author	71.4	48.6	25.7	64.6	83.8
	Conference	60.0	35.7	22.1	53.6	64.6
	Average	64.8	46.7	30.7	43.7	71.2
pLSI	Paper	32.5	33.8	30	9.7	40.4
	Author	65.0	40.0	22.5	60.4	75.5
	Conference	47.5	36.3	21.3	45.1	54.1
	Average	48.3	36.7	24.6	38.4	56.7
LDA	Paper	31.4	48.6	42.9	13.5	45.8
AT	Paper	42.9	48.6	42.9	13.1	49.3
	Author	82.9	45.7	25.7	73.5	78.1
	Average	62.9	47.1	34.3	43.3	63.7
ACT1	Paper	42.9	45.7	43.6	16.6	51.0
	Author	91.4	50.0	26.4	80.0	89.6
	Conference	62.9	41.4	23.6	60.7	72.3
	Average	65.7	45.7	31.2	52.4	71.0
ACT1+RW1	Paper	68.6	61.4	50.7	17.1	66.6
	Author	80.0	51.4	27.1	77.6	87.4
	Conference	62.9	42.9	23.6	59.5	72.0
	Average	**70.5**	**51.9**	**33.8**	**51.4**	**75.4**

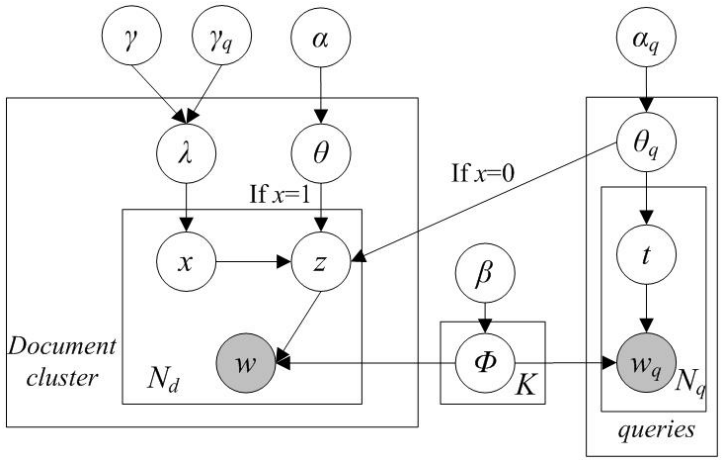
Work3: Multi-topic based Query-oriented Summarization (SDM'09)



qLDA – Query Latent Dirichlet Allocation



qLDA



1. For each topic t, draw the word distribution Φ_t from Dirichlet(β)
2. For each query q
   a) Draw a topic distribution θ_q from Dirichlet(α_q)
   b) For each word w_{qi} in query,
      i) Draw a topic t_{qi} from θ_q
      ii) Draw a word w_{qi} from Φ_{tqi}
3. For each document d
   a) Draw a topic mixture θ_d from Dirichlet(α)
   b) Draw the proportion between words associated with a specific query and those associated with its general topic $\lambda=p(x=0|d)\sim beta(\gamma_q, \gamma)$.
   c) For each word w_{di},
      i) Toss a coin x_{di} from λ with $x=bernoulli(\lambda)$
      ii) if $x=0$,
         a) Draw a topic z_{di} from θ_q
         b) Draw a word w_{di} from Φ_{zdi}
      iii) else
         a) Draw a topic z_{di} from θ_d
         b) Draw a word w_{di} from Φ_{zdi}

$$P(t_{qi}|\mathbf{t}_{-qi},\mathbf{w}_q,\alpha_q)=\frac{n_{qt_{qi}}^{-qi}+\alpha_q}{\sum_t(n_{qt}^{-qi}+\alpha_q)}\frac{n_{t_{qi}w_{qi}}^{-qi}+\beta}{\sum_v(n_{t_{qi}v}^{-qi}+\beta)}$$

$$P(z_{w_{di}},x_{w_{di}}=0|\mathbf{w}_d,\mathbf{w}_q,\mathbf{z}_{-di},\gamma,\gamma_q,\alpha_q,\beta)=\frac{n_{d0}^{-di}+\gamma_q}{n_{d0}^{-di}+n_{d1}^{-di}+\gamma_q+\gamma}\frac{n_{d0z_{w_{di}}}^{-di}+n_{qz_{w_{di}}}+\alpha_q}{\sum_z(n_{d0z}^{-di}+n_q+\alpha_q)}\frac{n_{z_{w_{di}}w_{di}}^{-di}+n_{z_{w_{di}}w_{di}}^{q}+\beta}{\sum_v(n_{z_{w_{di}}v}^{-di}+n_{z_{w_{di}}v}^{q}+\beta)}$$

$$P(z_{w_{di}},x_{w_{di}}=1|\mathbf{w}_d,\mathbf{w}_q,\mathbf{z}_{-di},\gamma,\gamma_q,\alpha,\beta)=\frac{n_{d1}^{-di}+\gamma}{n_{d0}^{-di}+n_{d1}^{-di}+\gamma_q+\gamma}\frac{n_{d1z_{w_{di}}}^{-di}+\alpha}{\sum_z(n_{d1z}^{-di}+\alpha)}\frac{n_{z_{w_{di}}w_{di}}^{-di}+n_{z_{w_{di}}w_{di}}^{q}+\beta}{\sum_v(n_{z_{w_{di}}v}^{-di}+n_{z_{w_{di}}v}^{q}+\beta)}$$

Topic Modeling with Regularization

The new objective function:

$$O_\xi(C, q) = -\xi L(C) + (1 - \xi) R(C, q)$$

with

$$R(C, q) = \frac{1}{2} \sum_{d \in C} w(d, q) \sum_{z} (\theta_{qz} - \theta_{dz})^2$$

Algorithm 1: Parameter estimation

1. initialize the parameters (θ, θ_q, ϕ) randomly;
2. run 200 burn-in sampling iterations using LDA [5];
3. run a two-stage training process iteratively:

(a) train the model parameters (θ, θ_q, ϕ) using the objective function $O_1(C, q) = -L(C)$ with a standard Gibbs sampling procedure by setting the Dirichlet prior for each document as

$$\alpha_{dz} = \alpha + \eta \theta_{qz}$$

where θ_{qz} is the z-th dimension of the multinomial of query q with $1 \leq z \leq T$; α_{dz} is a parameter specific to the topic of document d; and η is a parameter.

(b) fix ϕ, and re-estimate the multinomial θ to minimize O_ξ; we employ the algorithm proposed in [20] to optimize O_ξ, (i.e. run an iterative process to obtain the new θ for each document d by minimizing the objective function again.)

$$\theta_{dz}^{(n+1)} = \mu \theta_{dz}^{(n)} + (1 - \mu) w(d, q) \theta_{qz}$$

where μ is a coefficient to smooth the topic distribution between documents and the query.

Results

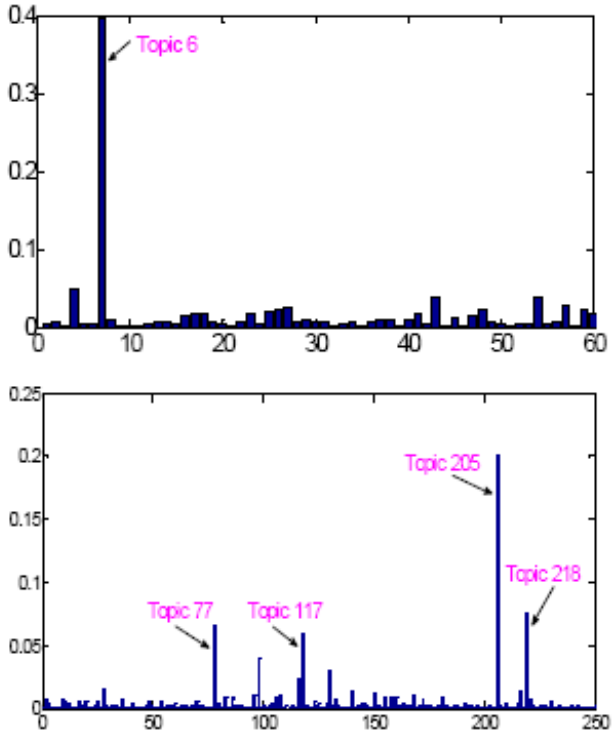


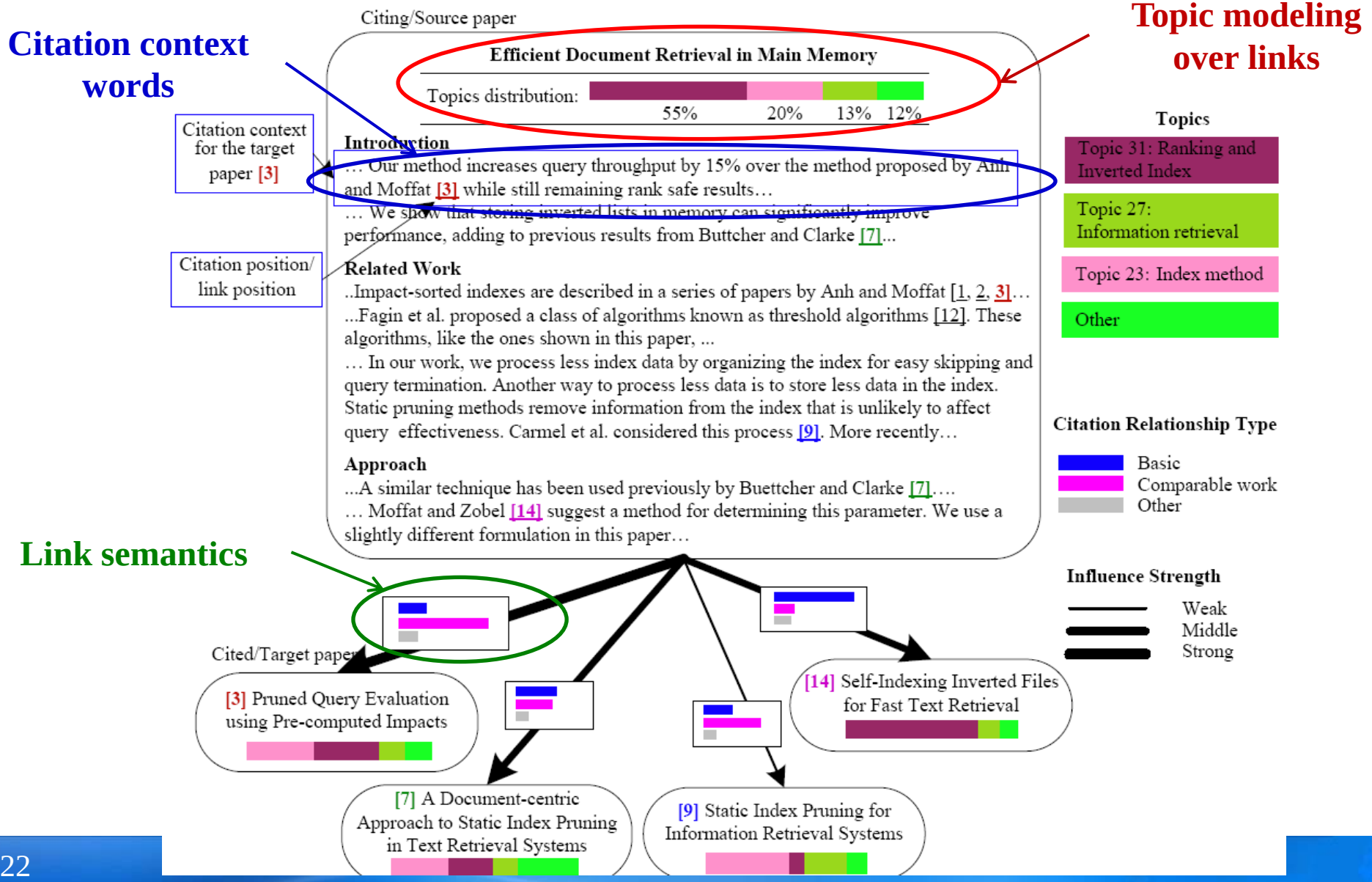
Figure 3: Topic distribution for in D357 (T=60 and T=250). The x axis denotes topics and the y axis denotes the occurrence probability of each topic in D357.

Table 6: Results of different methods on the Epinions data.

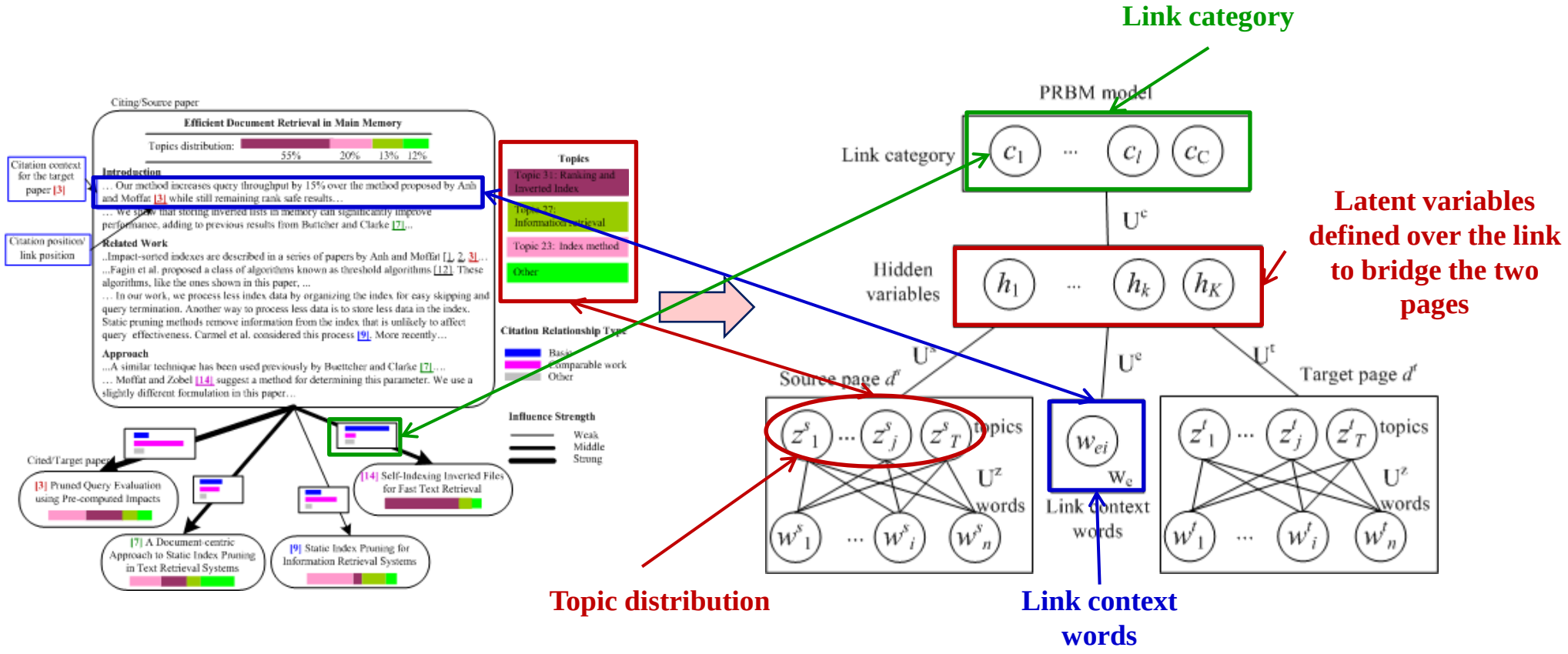
Method	Rouge1	Rouge2	RougeSU4
TF	0.28486	0.07579	0.10588
pLSI (Max Score)	0.28763	0.08031	0.10215
pLSI (Sum Score)	0.28291	0.07628	0.10072
pLSI+TF (Max Score)	0.28653	0.07956	0.10572
pLSI+TF (Sum Score)	0.29104	0.07832	0.10721
LDA (Max Score)	0.29155	0.08218	0.11245
LDA (Sum Score)	0.28761	0.07827	0.10671
LDA+TF (Max Score)	0.28483	0.07228	0.10403
LDA+TF (Sum Score)	0.29305	0.07802	0.10796
qLDA (Max Score)	0.29113	0.07765	0.10662
qLDA (Sum Score)	0.28815	0.08189	0.10803
qLDA+TF (Max Score)	0.29698	0.07430	0.10741
qLDA +TF (Sum Score)	0.30326	0.08774	0.11611
TMR (Max Score)	0.30453	0.08349	0.11590
TMR (Sum Score)	0.31271	0.08960	0.11931
TMR+TF (Max Score)	0.31647	0.08815	0.11758
TMR +TF (Sum Score)	**0.32375**	**0.09081**	**0.12894**

DUC 05 (document understanding conference)

Work4: Topic Distributions over Links on Web (ICDM'09)



Pairwise Restricted Boltzmann Machines (PRBMs)



Example

Pairwise Restricted Boltzmann Machines (PRBMs)

Results: Accuracy of Link Categorization

Approach	Type	Precision	Recall	F1-measure	Accuracy
SVM	drill down	64.09	53.60	58.24	88.50
	similar	77.66	83.34	80.33	76.59
	other	64.63	59.58	61.77	79.71
SVM+LDA	dı				
SVM+RBM	dı				
gPRBM	dı				
dPRBM	dı				
hPRBM	dı				

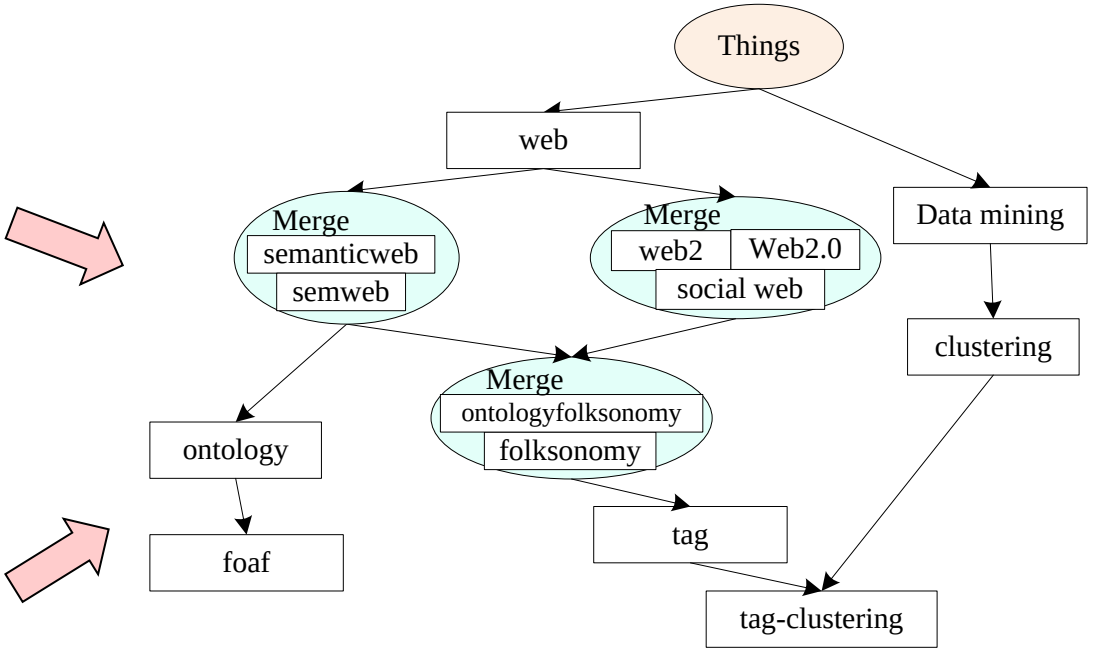
Topics (z^s) of Citing Papers	Topics (z^t) of Cited Papers	Influential Strength
Drill down (Basic theory)		
Topic 37 (Probabilistic model)	Topic 51 (Maximum entropy)	5.05
Topic 37 (Probabilistic model)		
Topic 13 (Semantic Web)		
Sin		
Topic 7 (Recommendation)		
Topic 21 (Frequent pattern learning)		
Topic 26 (Ranking & Inverted Index)		
Topic 26 (Ranking & Inverted Index)		
Topic 26 (Ranking & Inverted Index)		
Topic 9 (Clustering)		

proach

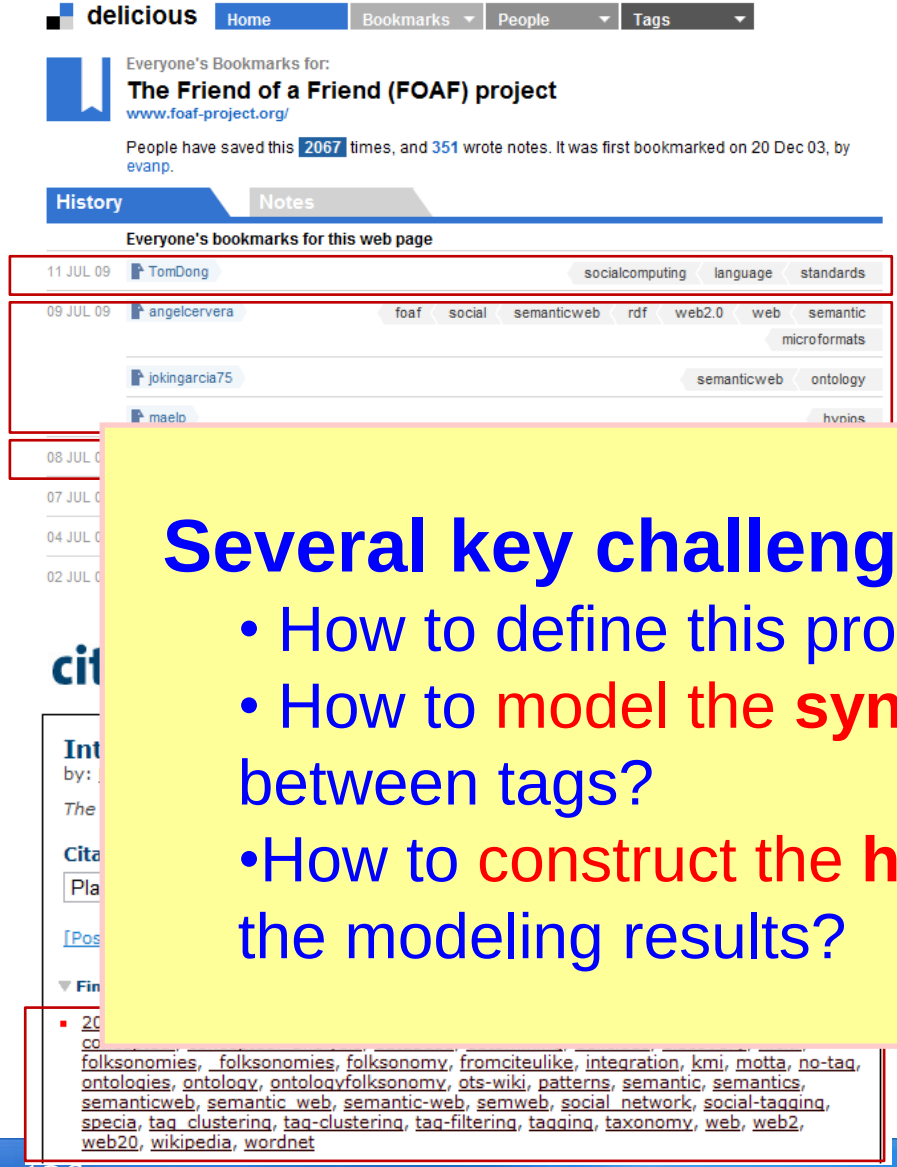
Source Paper	Target Paper	Influence
Drill Down (Basic theory)		
Latent dirichlet allocation	A variational bayesian framework for graphical models	2.36
A sentiment-aware model for predicting sales	Probabilistic latent semantic analysis	1.93
Mining and summarizing customer reviews	Foundations of statistical natural language processing	3.25
Similar (Comparable work)		
Constraint-driven clustering	Max-min d-cluster formation in wireless ad hoc networks	2.00
Hmms and coupled hmms for multi-channel eeg classification	Gaussian observation hidden markov models for eeg analysis	0.51
Scalable collaborative filtering using cluster-based smoothing	Evaluating collaborative filtering recommender systems	1.51
Other		
Latent dirichlet allocation	Overview of the first text retrieval conference	3.93
Utility scoring of product reviews	Introduction to modern information retrieval	3.16
Sentiment analyzer: extracting sentiments about a given topic	Finding parts in very large corpora	2.71

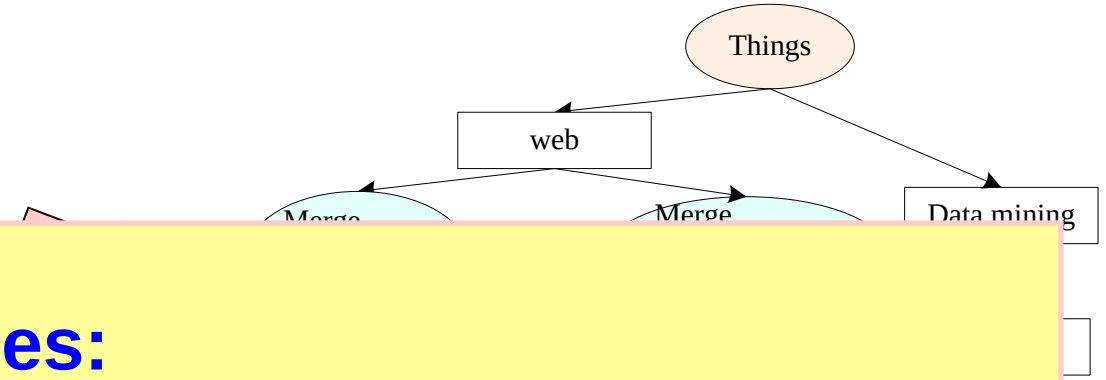
Work5: Towards Ontology Learning from Folksonomies (IJCAI'09)





Towards Ontology Learning from Folksonomies



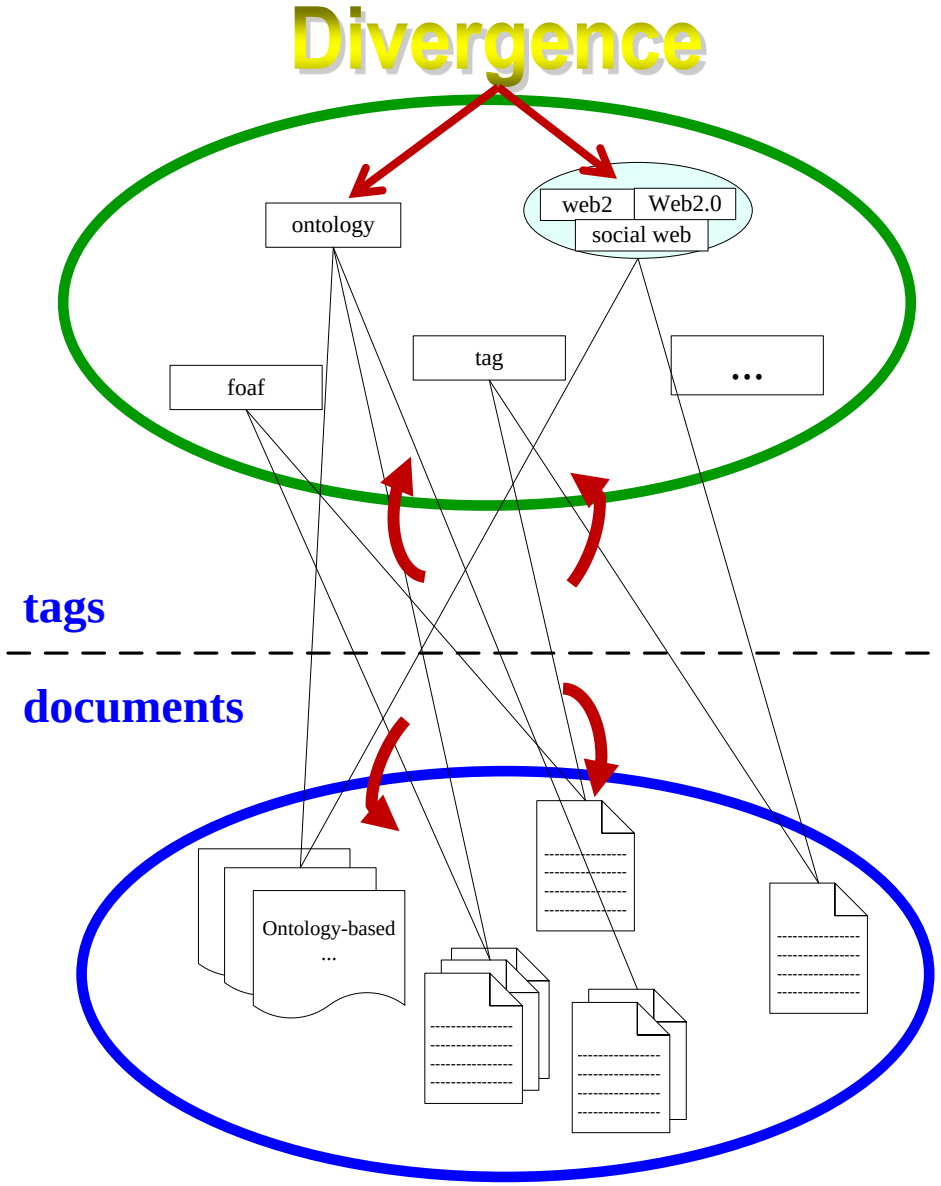


Several key challenges:

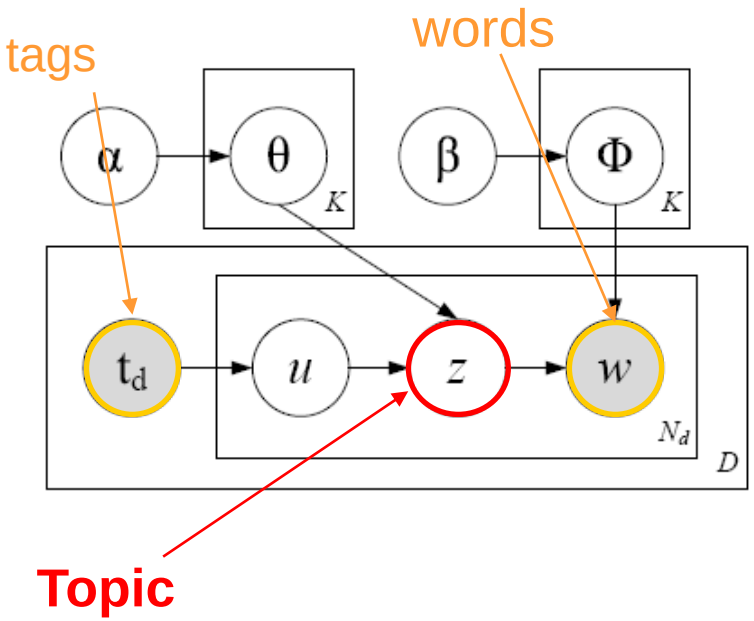
- How to define this problem in a **principled** way?
- How to model the **synonym/hypernym/homonym** between tags?
- How to construct the **hierarchical ontology** according to the modeling results?

Our Solution

1. Use topic to model tags and documents.
2. Define four divergence measures to estimate the difference between tags.
3. Present an algorithm to construct the hierarchical structure from the tags.



Tag-Topic (TT) Models



Tag-Topic (TT) Models

Generative process:

1. For each topic z, draw a multinomial distribution ϕ_z from a Dirichlet prior β_z;
2. For each word w_{di} in the document d:
   (a) draw a tag u_{di} from t_d uniformly;
   (b) draw a topic from a multinomial distribution $\theta_{u_{di}}$ specific to the tag u_{di}, where θ is generated from a Dirichlet prior α;
   (c) draw a word w_{di} from the multinomial distribution $\phi_{z_{di}}$.

$$P(z_{di}, u_{di}|\mathbf{z}_{-di}, \mathbf{u}_{-di}, \mathbf{w}, \mathbf{t}, \alpha, \beta) = \frac{P(\mathbf{z}, \mathbf{u}, \mathbf{w}, \mathbf{t}|\alpha, \beta)}{P(\mathbf{z}_{-di}, \mathbf{u}_{-di}, \mathbf{w}, \mathbf{t}|\alpha, \beta)}$$
$$\propto \frac{m_{uz}^{-di} + \alpha_z}{\sum_z (m_{uz}^{-di} + \alpha_z)} \frac{n_{zv}^{-di} + \beta_v}{\sum_v (n_{zv}^{-di} + \beta_v)}$$

Topic Smoothing

The new objective function:

$$O_\epsilon(D,G) = -\epsilon L(D) + (1-\epsilon) R(D,G)$$

with

$$R(D,G) = \frac{1}{2} \sum_{(t_i,t_j)\in G} w(t_i,t_j) \sum_{k=1}^{K} (\theta_{t_i Z_k} - \theta_{t_j Z_k})^2$$

Smoothing term

Log-likelihood of the tag-topic (TT) model.

Algorithm: Parameter estimation

1. initialize the parameters (θ, ϕ) randomly;
2. run 200 burn-in iterations using Equation (4);
3. run a two-stage training iteratively:

(a) train the model parameters (θ, ϕ) using the objective function $O_1(D,G) = -L(D)$ with a standard Gibbs sampling procedure (Equation (4)) by setting the Dirichlet prior for each tag t_i as

$$\alpha_{t_i k} = \epsilon\alpha + (1-\epsilon)\frac{K}{|G_{t_i}|} \sum_{(t_i,t_j)\in G} \theta_{t_j k}$$

where θ_{tk} is the k-th dimension of the multinomial of tag t with $1 \le k \le K$, $|G_t|$ is the number of neighbors of tag t in the tag network G;

(b) fix ϕ, and re-estimate the multinomial θ to minimize O_ϵ; we employ the algorithm proposed in [Mei *et al.*, 2008] to optimize O_ϵ, (i.e. run an iterative process to obtain the new θ for each tag t by minimizing the objective function again.)

$$\theta_{t_i k}^{(n+1)} = \gamma\theta_{t_i k}^{(n)} + (1-\gamma)\frac{\sum_{(t_i,t_j)\in G} w(t_i,t_j)\theta_{t_j k}^{(n)}}{\sum_{(t_i,t_j)\in G} w(t_i,t_j)}$$

where γ is a coefficient to smooth the topic distribution among the neighbors.

Divergence Estimation

- Tag divergence

$$diverg(t_i, t_j) = D_{KL}(\theta_{t_i} \| \theta_{t_j}) = \sum_{z=1}^{K} \theta_{t_i z} log \frac{\theta_{t_i z}}{\theta_{t_j z}}$$

Estimated topic distribution

- Hypernym-divergence

$$hyper - diverg(t_i, t_j) = \sum_{z=1}^{K} \frac{P(t_i|z_k) - P(t_j|z_k)}{P(t_i|z_k)}$$

Posterior probability derived from the topic modeling results

- Merging-divergence

$$merg - diverg(t_i, t_j) = \frac{1}{2}(diverg(t_i, t_j) + diverg(t_j, t_i))$$

- Keep-divergence

$$keep - diverg(t_i, t_j) = \frac{1}{2}(merg - diverg(t_i, t_j) + (max_k(diverg(t_i, t_k) - diverg(t_i, t_k))))$$

Hierarchical Structure Construction

1. initialize the tree **O** by taking all tags as the leaf nodes;
2. initialize a concept set **A** with all tags;
3. do{

Step 1.

(a) find a pair of tags (or virtual tag/concepts) (t_i, t_j) with an operation from A that minimizes the objective function (10);

Step 2.

(b) execute the operation:

i. for Subordinate, assign t_i as the hypernym of t_j, remove t_j from **A**, and move ti as the parent node of t_j in **O**;

ii. for Merge, create a virtual concept c_{ij} by combining t_i and t_j, add it to **A** as well to **O** as a leaf node, remove t_i and t_j from both **A** and **O**, and calculate the divergences of the new concept c_{ij} with all others tags in **A**.

iii. for Keep, create a virtual concept c_{ij} as the hypernym of both t_i and t_j, add c_{ij} as the parent node of t_i and t_j in **O**, add c_{ij} to **A**, remove t_i and t_j from **A**, and calculate the divergences of the new concept c_{ij} with all others tags in **A**.

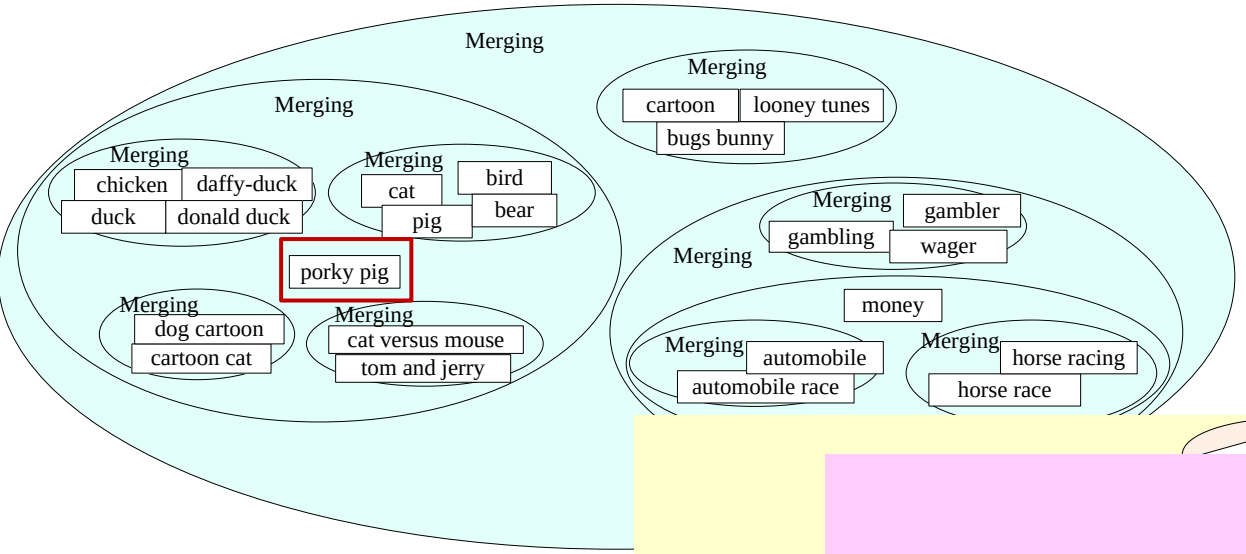
4. }until $|\mathbf{A}| \leq 1$; //$|\mathbf{A}|$ is the number of elements in **A**.
5. return the hierarchical tree **O**.

Correspond to a divergence

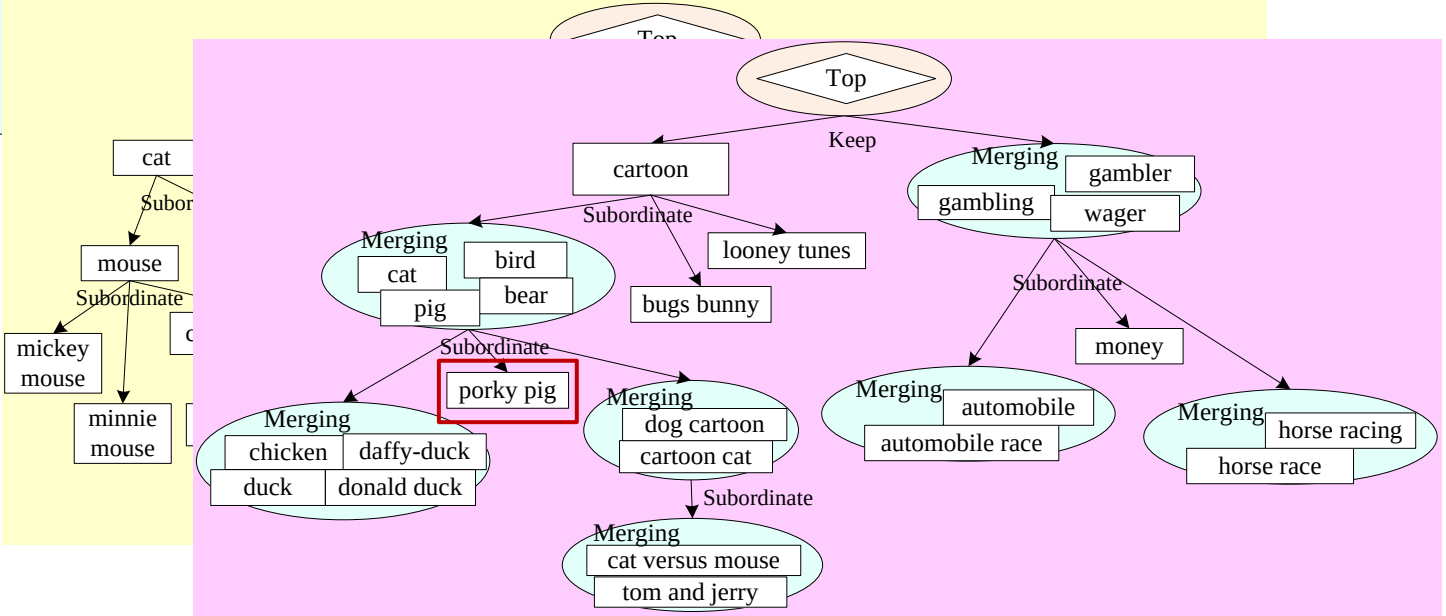
$$argmin_{i \neq j} Operation(t_i, t_j) + \lambda log(N)$$

Penalty to the complex of the generated hierarchy

Case Study—Movie



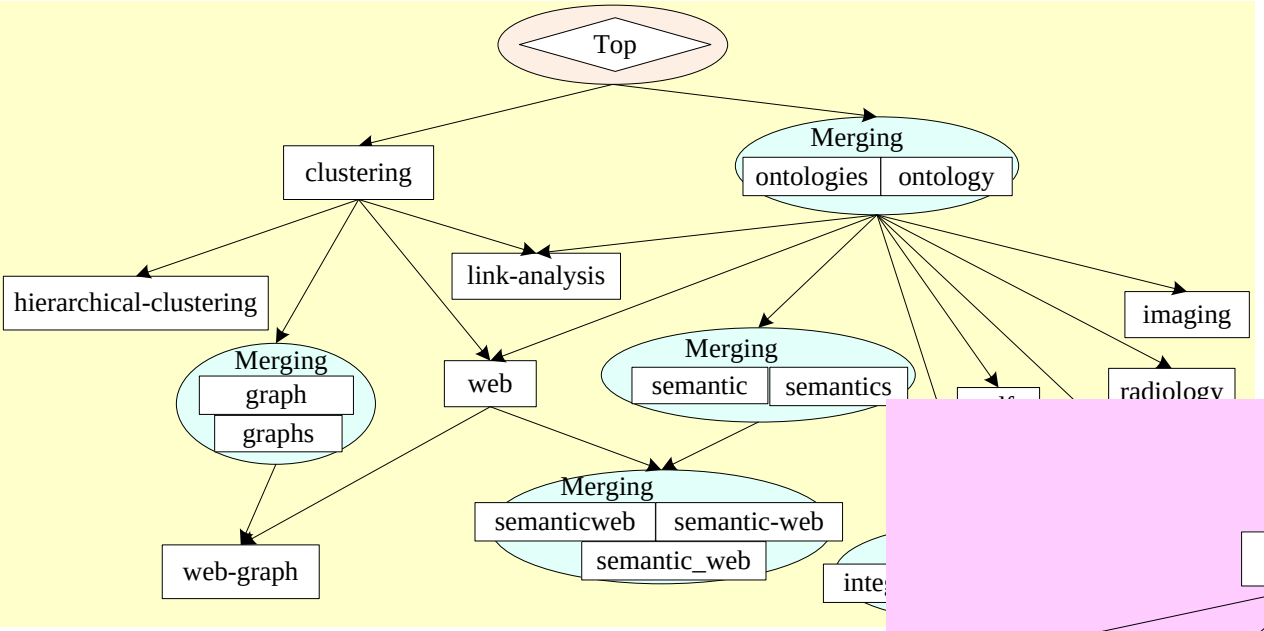
By clustering



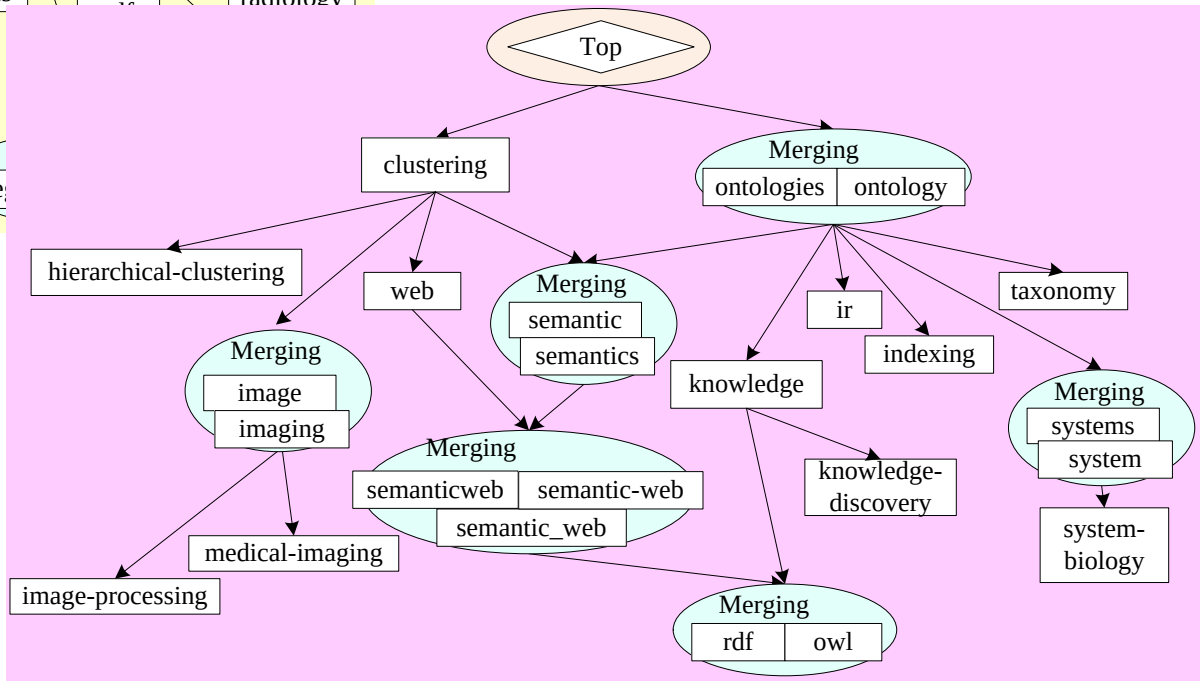
By TT

By TT with smoothing

Case Study—Paper



By TT



By TT with smoothing

Suggested Reading List

Reading List

Red color-good theoretical analysis

Blue color-good applications.

@ARTICLE{Blei:03,
AUTHOR = "David~M. Blei and Andrew~Y. Ng and Michael~I. Jordan",
TITLE = "Latent Dirichlet Allocation",
JOURNAL = "Journal of Machine Learning Research",
VOLUME = {3},
PAGES = {993-1022},
YEAR = {2003} }

@INPROCEEDINGS{ l. griffiths:integrating,
AUTHOR = "Thomas L. Griffiths and Mark Steyvers and David M. Blei and Joshua B. Tenenbaum",
TITLE = "Integrating Topics and Syntax",
booktitle = "NIPS'04“
YEAR = {2004} }

@INPROCEEDINGS{Blei2:06,
AUTHOR = "David~M. Blei and John~D. Lafferty",
TITLE = "Dynamic Topic Models",
BOOKTITLE = "Proceedings of the 23th International Conference on Machine Learning (ICML'06)",
PAGES = "113-120",
YEAR = {2006} }

@INPROCEEDINGS{Blei:07,
AUTHOR = "David~M. Blei and Jon~D. McAuliffe",
TITLE = "Supervised Topic Models",
BOOKTITLE = "Proceedings of Advances in Neural Information Processing Systems",
YEAR = {2007} }

Reading List (cont.)

```
@INPROCEEDINGS{Griffiths:04PNAS,
   AUTHOR = "Thomas~L. Griffiths and Mark Steyvers",
   TITLE = "Finding Scientific Topics",
   BOOKTITLE = "Proceedings of the National Academy of Sciences",
   PAGES = "5228-5235",
   YEAR = {2004}   }

@INPROCEEDINGS{Griffiths:04NIPS,
   AUTHOR = "Thomas~L. Griffiths and Mark Steyvers and David~M. Blei and Joshua~B",
   TITLE = "Integrating Topics and Syntax",
   BOOKTITLE = "Proceedings of Advances in Neural Information Processing Systems",
   YEAR = {2004}   }

@INPROCEEDINGS{Gruber:07,
   AUTHOR = "Amit Gruber and Michal Rosen-Zvi and Yair Weiss",
   TITLE = "Hidden Topic Markov Models",
   BOOKTITLE = "Proceedings of Artificial Intelligence and Statistics (AISTATS'07)",
   YEAR = {2007}   }
@ARTICLE{McCallum:07,
   AUTHOR = "Andrew McCallum and Xuerui Wang and Andres Corrada-Emmanuel",
   TITLE = "Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email",
   JOURNAL = {Journal of Artificial Intelligence Research (JAIR)},
   VOLUME = {30},
   PAGES = {249-272},
   YEAR = {2007} }
```

Reading List (cont.)

```
@inproceedings{Mei:07,
    author = {Qiaozhu Mei and Xu Ling and Matthew Wondra and Hang Su and ChengXiang Zhai},
    title = {Topic sentiment mixture: modeling facets and opinions in weblogs},
    year = {2007},
    pages = {171--180},
    booktitle = {Proceedings of WWW'07},
%    booktitle = {Proceedings of the 16th international conference on World Wide Web (WWW'07)},
 }

@INPROCEEDINGS{Mei:08WWW,
   AUTHOR = "Qiaozhu Mei and Deng Cai and Duo Zhang and ChengXiang Zhai",
   TITLE = "Topic Modeling with Network Regularization",
   PAGES = "101-110",
   YEAR = {2008},
   BOOKTITLE = "Proceedings of the 17th International World Wide Web Conference (WWW'08)",
}

@INPROCEEDINGS{Mei:08ACL,
   AUTHOR = "Qiaozhu Mei and ChengXiang Zhai",
   TITLE = "Generating impact-cased summaries for scientific literature",
   BOOKTITLE ={Proceedings of the 46nd Annual Meeting of the Association for Computational Linguistics (ACL'08)},
   PAGES = "816-824",
   YEAR = {2008}}
```

Reading List (cont.)

```
@INPROCEEDINGS{Mika:05,
        author = {Peter Mika},
        title = {Ontologies Are Us: A Unified Model of Social Networks and Semantics},
        BOOKTITLE = {Proceedings of the 4th International Semantic Web Conference (ISWC'05)},
        pages = {522--536},
        year = {2005}
}

@INPROCEEDINGS{Mimno:07,
   AUTHOR = "David Mimno and Andrew McCallum",
   TITLE = "Expertise Modeling for Matching Papers with Reviewers",
   PAGES = "500-509",
   YEAR = {2007},
   BOOKTITLE = "KDD'07",
}

@INPROCEEDINGS{Minka:02,
   AUTHOR = "Thomas Minka and John Lafferty",
   TITLE = "Expectation-Propagation for the Generative Aspect Model",
   BOOKTITLE = "Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI'2002)",
   PAGES = "352-359",
   YEAR = {2002}}
```

Reading List (cont.)

@INPROCEEDINGS{Minka:03,
AUTHOR = "Thomas Minka",
TITLE = "Estimating a Dirichlet Distribution",
BOOKTITLE = "Technique Report, http://research.microsoft.com/~minka/papers/dirichlet/",
YEAR = {2003} }

@INPROCEEDINGS{Nallapati:07,
AUTHOR = "Ramesh~M. Nallapati and Susan Ditmore and John~D. Lafferty and Kin Ung",
TITLE = "Multiscale Topic Tomography",
PAGES = "520-529",
YEAR = {2007},
BOOKTITLE = "KDD'2007",
}
@INPROCEEDINGS{Nallapati:08,
AUTHOR = "Ramesh~M. Nallapati and Amr Ahmed and Eric Xing and William~W. Cohen",
TITLE = "Joint Latent Topic Models for Text and Citations",
BOOKTITLE = "Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2008)",
Year = “2008”}

@inproceedings{Newman:07,
Author = {David Newman and Arthur Asuncion and Padhraic Smyth and Max Welling},
Title = {Distributed inference for Latent Dirichlet Allocation},
Year = {2007},
Booktitle = {Proceedings of NIPS'07},
}

Reading List (cont.)

```
@inproceedings{Porteous:08,
    author = {Ian Porteous and David Newman and Alexander Ihler and Arthur Asuncion and Padhraic Smyth and Max Welling},
    title = {Fast collapsed gibbs sampling for latent dirichlet allocation},
    booktitle = {Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'08)},
    year = {2008},
    pages = {569--577},
 }
@INPROCEEDINGS{Nie:06,
   AUTHOR = "Lan Nie and Brian~D. Davison and Xiaoguang Qi",
   TITLE = "Topical Link Analysis for Web Search",
   PAGES = "91-98",
   YEAR = {2006},
   BOOKTITLE = "SIGIR'2006",
}
@INPROCEEDINGS{Nodelman:02,
   AUTHOR = "Uri Nodelman and Christian~R. Shelton and Daphne Koller",
   TITLE = "Continuous Time Bayesian Networks",
   BOOKTITLE = "Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI'2002)",
   PAGES = "378-387",
   YEAR = {2002}}
@INPROCEEDINGS{Rosen-Zvi:04,
   AUTHOR = "Michal Rosen-Zvi and Thomas Griffiths and Mark Steyvers and Padhraic Smyth",
   TITLE = "The Author-Topic Model for Authors and Documents",
   YEAR = {2004},
   BOOKTITLE = "UAI'04“,}
```

Reading List (cont.)

```
@INPROCEEDINGS{Steyvers:04,
   AUTHOR = "Mark Steyvers and Padhraic Smyth and Thomas Griffiths",
   TITLE = "Probabilistic Author-Topic Models for Information Discovery",
   YEAR = {2004},
   PAGES = {306-315},
   BOOKTITLE = "KDD'04“,}
@INPROCEEDINGS{Teh:04,
   AUTHOR = "Yee~Whye Teh and Michael~I. Jordan and Matthew~J. Beal and David~M. Blei",
   TITLE = "Hierarhical Dirichlet Processes",
   BOOKTITLE = "Technical Report 653, Department of Statistics, UC Berkeley",
   YEAR = {2004}}
@INPROCEEDINGS{Ahmed:09KDD,
   AUTHOR = "Amr Ahmed and Eric P. Xing and William W. Cohen and Robert F. Murphy",
   TITLE = "Structured correspondence topic models for mining captioned figures in biological literature",
   booktitle = "Proceeding of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'09)"
   PAGES = {39-48},
   YEAR = {2009}
}
@inproceedings{Leskovec:09KDD,
 author = {Leskovec, Jure and Backstrom, Lars and Kleinberg, Jon},
 title = {Meme-tracking and the dynamics of the news cycle},
 booktitle = {Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'09)},
 year = {2009},
 pages = {497--506},
 }
```

Reading List (cont.)

@INPROCEEDINGS{Wallach:06,
AUTHOR = "Hanna~M. Wallach",
TITLE = "Topic Modeling: Beyond Bag-of-Words",
BOOKTITLE = "Proceedings of the 23rd international conference on Machine learning (ICML'06)",
PAGES = "977-984",
YEAR = {2006}}

@INPROCEEDINGS{Wang:06,
AUTHOR = "Xuerui Wang and Andrew McCallum",
TITLE = "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends",
BOOKTITLE = "Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'06)",
PAGES = "424-433",
YEAR = {2006}}

@inproceedings{Wang:07,
author = {Xuerui Wang and Chris Pal and Andrew McCallum},
title = {Generalized component analysis for text with heterogeneous attributes},
booktitle = {Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'07)},
year = {2007},
pages = {794--803},}

@INPROCEEDINGS{Wei:06,
AUTHOR = "Xing Wei and W. Bruce Croft",
TITLE = "LDA-Based Document Models for Ad-hoc Retrieval",
BOOKTITLE = "Proceedings of the 29th ACM SIGIR International Conference on Information Retrieval (SIGIR'06)",
PAGES = "178-185",
YEAR = {2006}}

Reading List (cont.)

```
@inproceedings{Zhai:04,
    author = {ChengXiang Zhai and Atulya Velivelli and Bei Yu},
    title = {A cross-collection mixture model for comparative text mining},
    year = {2004},
    pages = {743--748},
    booktitle = {KDD'04},
}
@INPROCEEDINGS{Cohn:01,
   AUTHOR = {David Cohn and Thomas Hofmann},
   TITLE = {The Missing Link - A Probabilistic Model of Document Content and Hypertext Connectivity},
   BOOKTITLE = {Proceedings of the 13th Neural Information Processing Systems (NIPS'01)},
   YEAR = {2001}}
@INPROCEEDINGS{Hofmann:99,
   AUTHOR = "Thomas Hofmann",
   TITLE = "Probabilistic Latent Semantic Indexing",
   PAGES = "50-57",
   YEAR = {1999},
   BOOKTITLE = "SIGIR'99“,}
@INPROCEEDINGS{Hofmann:03,
   AUTHOR = "Thomas Hofmann",
   TITLE = "Collaborative Filerting via Gaussian Probabilistic Latent Semantic Analysis",
   BOOKTITLE = "Proceedings of the 26nd International Conference on Research and Development in Information Retrieval (SIGIR'03)",
   PAGES = "259-266",
YEAR = {2003}   }
```

Reading List (cont.)

@INPROCEEDINGS{Tang:08ICDM,
AUTHOR = "Jie Tang and Ruoming Jin and Jing Zhang",
TITLE = "A Topic Modeling Approach and its Integration into the Random Walk Framework for Academic Search",
YEAR = {2008},
BOOKTITLE = "ICDM'08",}
@INPROCEEDINGS{Tang:09SDM,
AUTHOR = "Jie Tang and Limin Yao and Dewei Chen",
TITLE = "Multi-topic based Query-oriented Summarization",
YEAR = {2009},
BOOKTITLE = "Proceedings of 2009 SIAM International Conference on Data Mining (SDM'2009)“,}
@INPROCEEDINGS{Tang:09KDD,
AUTHOR = "Jie Tang and Jimeng Sun and Chi Wang and Zi Yang",
TITLE = "Social Influence Analysis in Large-scale Networks",
YEAR = {2009},
PAGES = "807--816",
BOOKTITLE = "SIGKDD'09“,}
@INPROCEEDINGS{Tang:09ICDM,
AUTHOR = "Jie Tang and Jing Zhang and Jeffrey Xu Yu and Zi Yang and Keke Cai and Rui Ma and Li Zhang and Zhong Su",
TITLE = "Topic Distributions over Links on Web",
YEAR = {2009},
BOOKTITLE = "Proceedings of 2009 IEEE International Conference on Data Mining (ICDM'09)“, }
@inproceedings{Tang:09IJCAI,
author = {Tang, Jie and Leung, Ho-fung and Luo, Qiong and Chen, Dewei and Gong, Jibin},
title = {Towards ontology learning from folksonomies},
booktitle = {Proceedings of the 21st international jont conference on Artifical intelligence (IJCAI'09)},
year = {2009},
pages = {2089--2094}, }

Thanks!

Q&A

HP: http://keg.cs.tsinghua.edu.cn/jietang/

http://aminer.org

Appendix

Variational Inference

- Suppose all observations are X, and all hidden variables are Z. Our goal is to find an approximation for the posterior distribution $P(Z|X)$ as well as for the model evidence $P(X)$.
- We can decompose the log marginal probability :

$$\ln P(X) = \mathcal{L} + \mathrm{KL}(q||P)$$

where :

$$\mathcal{L}(q) = \int q(Z) \ln \frac{P(X,Z)}{q(Z)} dZ$$

$$\mathrm{KL}(q||P) = -\int q(Z) \ln \frac{p(Z|X)}{q(Z)} dZ$$

- The maximum of the lower bound occurs when the KL divergence vanishes, which occurs when $q(Z)$ equals the true posterior distribution $P(Z|X)$. However, the true posterior distribution is intractable.
- Our goal is to restrict the family of $q(Z)$ sufficiently that they comprise only tractable distributions.

Factorized Distributions

- We assume q distribution can factorize as:

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i)$$

- The factorized form of variational inference is called mean field theory.
- We wish to make a free form optimization of $\mathcal{L}(q)$ with respect to all the distributions $q_i(Z_i)$, which we do by optimizing with respect to each of the factors in turn.

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left(\ln P(X,Z) - \sum_i \ln q_i \right) dZ \qquad (8) \\
&= \int q_j \left(\int \ln P(X,Z) \prod_{i \neq j} q_i dZ_i \right) dZ_j - \int q_j \ln q_j dZ_j + \text{const} \\
&= \int q_j \ln \widetilde{P}(X,Z_j) dZ_j - \int q_j \ln q_j dZ_j + \text{const}
\end{aligned}$$

where $\ln \widetilde{P}(X,Z_j) = E_{i \neq j}[\ln P(X,Z)] + \text{const}$

Factorized Distributions (Cont.)

- The maximizing $\mathcal{L}(q)$ is equivalent to minimizing the KL divergence, and the minimum occurs when $q_j(Z_j) = \widetilde{P}(X, Z_j)$. Thus we obtain the optimal solution as:

$$\ln q_j^*(Z_j) = E_{i \neq j}[\ln P(X, Z)] + \text{const}$$

- It says the log of the optimal solution for factor q_j is obtained simply by considering the log of the joint distribution over all hidden and visible variables and then taking the expectation with respect to all of the other factors $\{q_i\}$ for $i \neq j$.
- We seek a consistent solution by first initializing all of the factors $q_i(Z_i)$ appropriately and then cycling through the factors and replacing each in turn with a revised estimate given by the above equitation evaluated using the current estimates for all of the other factors.