Probabilistic Topic Model

Jie Tang

Tsinghua University
2012
Why Topic?

• Math & Feature Space: SVD -> PCA
• Information Retrieval: Tfidf->lsi->plsi->lda
• More:
  – Lda->topic correlation model
  – Static->dynamic
  – Bow->word-order
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?

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

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

- Bayesian approach: use priors
  - Mixture weights ~ Dirichlet(\(\alpha\))
  - Mixture components ~ Dirichlet(\(\beta\))
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

\[
\begin{align*}
\text{Doc 1} & \rightarrow \theta \\
\alpha & \rightarrow T_1, T_2, \ldots, T_T \\
Z & \rightarrow \phi \\
\phi & \rightarrow w_1, w_2, \ldots, w_v \\
Z & \rightarrow w_1, w_2, \ldots, w_v \\
W & \rightarrow w_1, w_2, \ldots, w_v \\
\end{align*}
\]
PLSI(SIGIR’99)
LDA (JMLR’03)

Document specific distribution over topics

Topic distribution over words

Document

Topic

Word

α

θ

z

w

N_d

D

β

φ

T

β

φ
Modeling Authors with Words
(AAAI’09 workshop)

Uniform distribution over authors of doc

Distribution of authors over words
Author-Topic Model (UAI’04, KDD’05)

- **Uniform distribution of documents over authors**
- **Distribution of authors over topics**
- **Topic distribution over words**

The diagram shows the model with the following nodes:

- **Ad**: Author
- **X**: Document
- **Z**: Topic
- **W**: Word
- **Nd**: Number of documents
- **D**: Total number of documents

The model consists of the following distributions:

- **Author distribution over topics**
- **Topic distribution over words**
- **Document distribution of documents over authors**
Two state-of-the-art topic models:

- PLSI
- LDA
Probabilistic Latent Semantic Indexing (PLSI)
A generative model for generating the co-occurrence of documents $d \in D = \{d_1, \ldots, d_D\}$ and terms $w \in W = \{w_1, \ldots, w_W\}$, which associates latent variable $z \in Z = \{z_1, \ldots, z_Z\}$.

The generative processing is:

- $P(w|z)$
- $P(z|d)$
- $P(d)$
Matrix Interpretation

In the topic model, the word-document co-occurrence matrix is split into two parts: a topic matrix $\Phi$ and a document matrix $\Theta$. 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 e.g., P5 → C2(0.4); P4 → C1 (0.3)
  - More
Example 2: images

\[ X \approx U \times V^T \]

- **X**: input images
- **U**: face patterns
- **V^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)

\[
\begin{array}{c|c}
\text{court} & \text{president} \\
\text{government} & \text{served} \\
\text{council} & \text{governor} \\
\text{culture} & \text{secretary} \\
\text{supreme} & \text{senate} \\
\text{constitutional} & \text{congress} \\
\text{rights} & \text{presidential} \\
\text{justice} & \text{elected} \\
\hline
\text{flowers} & \text{disease} \\
\text{leaves} & \text{behaviour} \\
\text{plant} & \text{glands} \\
\text{perennial} & \text{contact} \\
\text{flower} & \text{symptoms} \\
\text{plants} & \text{skin} \\
\text{growing} & \text{pain} \\
\text{annual} & \text{infection} \\
\end{array}
\]

input documents \[\times\] \approx \quad \text{latent topics} \quad \times \quad \text{docs represented with topics}

\[X \times U \approx VT\]
Example 4: videos

$$\text{Video } D = \text{Low-rank appx. } A + \text{Sparse error } E$$
PLSI as a Mixture Model

\[ p_d(w) = \lambda_B p(w | \theta_B) + (1 - \lambda) \sum_{j=1}^{k} \pi_{d,j} p(w | \theta_j) \]

\[ \log p(d) = \sum_{w \in d} c(w, d) \log[\lambda_B p(w | \theta_B) + (1 - \lambda) \sum_{j=1}^{k} \pi_{d,j} p(w | \theta_j)] \]

Parameters:
\[ \lambda_B = \text{noise-level (manually set)} \]
\[ \theta \text{'s and } \pi \text{'s are estimated with Maximum Likelihood} \]
But how can we obtain the Eqs?
Maximum-likelihood

- Definition
  - We have a density function $P(x|\Theta)$ that is governed by the set of parameters $\Theta$, e.g., $P$ might be a set of Gaussians and $\Theta$ 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 assume these data vectors are i.i.d. with $P$.
  - Then the log-likelihood 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 $\Theta$ where the data $X$ is fixed. Our goal is to find the $\Theta$ that maximizes $L$. That is
    \[
    \Theta^* = \arg \max_{\Theta} L(\Theta \mid X)
    \]
PLSI Model

- Following the likelihood principle, we determine \( P(d) \), \( P(z|d) \), and \( P(w|d) \) by maximization of the log-likelihood function:

\[
L(\Theta | d, w, z) = \log \prod_d \prod_w P(d, w)^n \]

\[
= \sum_{d \in D} \sum_{w \in W} n(d, w) \log P(d, w)
\]

\[
= \sum_{d \in D} \sum_{w \in W} n(d, w) \log \left( \sum_{z \in Z} P(w | z)P(d | z)P(z) \right)
\]

- The co-occurrence times of \( d \) and \( w \) are obtained according to the multi-distribution.
Recall Jensen Inequality
Jensen’s Inequality

- Recall that $f$ is a **convex function** if $f''(x) \geq 0$, 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 $X = EX$ 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 \( L(\Theta) \) explicitly might be difficult, and the strategy is to instead repeatedly construct a lower-bound on \( L(\text{E-step}) \), 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, \( f(x) = \log x \) is a concave function, since \( f''(x) = -1/x^2 < 0 \)
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

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

– 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 \(\Theta\), try to make the lower-bound tight at the value of \(\Theta\).
  – To make the bound tight for a particular value of \(\Theta\), 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 \(\sum Q_i(z) = 1\), this further tells us that

\[
Q_i(z^{(i)}) = \frac{p(x^{(i)},z^{(i)};\Theta)}{\sum_z p(x^{(i)},z;\Theta)} = \frac{p(x^{(i)},z^{(i)};\Theta)}{p(x^{(i)};\Theta)} = p(z^{(i)}|x^{(i)};\Theta)
\]

– 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 \(\Theta\)
Basic EM Algorithm (cont’)

Repeat until convergence 

(E-step) For each $i$, set 

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

(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.

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

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

\[
= L(\Theta^{(t)})
\]

- 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 $\Theta$, 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

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

s.t. \[ \int_z q(z) = 1 \quad \Theta = \{P(w \mid z), p(z \mid d), p(z)\} \]

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

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

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

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

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

When \( \frac{dL}{dq(z)}=0 \) the lower bound \( L \) equates to the original problem \( f \)
By KL-Divergence

\[ L = \int q(z) \log p(X, z, \Theta) - q(z) \log q(z) \, dz \]

\[ = \int q(z) \log \frac{p(X, z, \Theta)}{q(z)} \, dz \]

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

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

\[ = -D(q(z) \| p(z | X, \Theta)) + \log p(X, \Theta) \]

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

Recall what is \( X \) in our problem.
Using EM to Learn PLSI
Parameter Estimation-Using EM

• According to Basic EM:

\[ Q_i(z^{(i)}) = p(z^{(i)} | x^{(i)}; \Theta) \]

• Then we define

\[ Q_i(z^{(i)}) = p(z | d, w) \]

• Thus according to Jensen’s inequality

\[
L(\Theta) = \sum_{d \in D} \sum_{w \in W} n(d, w) \log \sum_{z \in \mathcal{Z}} p(z | d, w) \frac{p(w | z)p(d | z)p(z)}{p(z | d, w)} \\
\leq \sum_{d \in D} \sum_{w \in W} n(d, w) \sum_{z \in \mathcal{Z}} p(z | d, w) \log \frac{p(w | z)p(d | z)p(z)}{p(z | d, w)}
\]
(1) Solve $P(w|z)$

- We introduce Lagrange multiplier $\lambda$ with the constraint that $\sum_w P(w|z) = 1$, and solve the following equation:

$$\frac{\partial}{\partial P(w|z)} \left\{ \sum_{d \in D} \sum_{w \in W} n(d, w) \sum_{z \in Z} p(z|d,w) \log \frac{p(w|z)p(d|z)p(z)}{p(z|d,w)} + \lambda \left[ \sum_z P(w|z) - 1 \right] \right\} = 0$$

$$\sum_{d \in D} n(d, w) P(z|d,w) \frac{P(w|z)}{P(w|z)} + \lambda = 0,$$

$$\therefore P(w|z) = - \frac{\sum_{d \in D} n(d, w) P(z|d,w)}{\lambda},$$

$$\sum_w P(w|z) = 1,$$

$$\therefore \lambda = - \sum_{w \in W} \sum_{d \in D} n(d, w) P(z|d,w),$$

$$\therefore P(w|z) = \frac{\sum_{d \in D} \sum_{w \in W} n(d, w) P(z|d,w)}{\sum_{w \in W} \sum_{d \in D} n(d, w) P(z|d,w)}$$
(2) Solve $P(d|z)$

- We introduce Lagrange multiplier $\lambda$ with the constraint that $\sum_d P(d|z) = 1$, and get the following result:

$$
\therefore P(d|z) = \frac{\sum_{w \in W} n(d, w)P(z|d, w)}{\sum_{d \in D} \sum_{w \in W} n(d, w)P(z|d, w)}
$$
(3) Solve $P(z)$

- We introduce Lagrange multiplier $\lambda$ with the constraint that $\sum_z P(z) = 1$, and solve the following equation:

$$\frac{\partial}{\partial P(z)} \left\{ \sum_{d \in D} \sum_{w \in W} n(d, w) \sum_{z} p(z \mid d, w) \log \frac{p(w \mid z)p(d \mid z)p(z)}{p(z \mid d, w)} + \lambda \left[ \sum_z P(z) - 1 \right] \right\} = 0$$

$$\sum_{d \in D} \sum_{w \in W} n(d, w) P(z \mid d, w) \frac{\partial P(z)}{P(z)} + \lambda = 0,$$

$$\therefore P(z) = - \frac{\sum_{d \in D} \sum_{w \in W} n(d, w) P(z \mid d, w)}{\lambda}$$

$$\sum_z P(z) = 1,$$

$$\therefore \lambda = - \sum_{d \in D} \sum_{w \in W} n(d, w) \sum_z P(z \mid d, w) = - \sum_{d \in D} \sum_{w \in W} n(d, w),$$

$$\therefore P(z) = \frac{\sum_{d \in D} \sum_{w \in W} n(d, w) P(z \mid d, w)}{\sum_{w \in W} \sum_{d \in D} n(d, w)}$$
(4) Solve $P(z|d,w) - M$ step

$$P(z | d, w) = \frac{P(d, w, z)}{P(d, w)}$$

$$= \frac{P(w, d | z)P(z)}{P(d, w)}$$

$$= \frac{P(w | z)P(d | z)P(z)}{\sum_{z \in Z} P(w | z)P(d | z)P(z)}$$
The final update Equations

- **E-step:**

\[
P(z \mid d, w) = \frac{P(w \mid z)P(d \mid z)P(z)}{\sum_{z \in Z} P(w \mid z)P(d \mid z)P(z)}
\]

- **M-step:**

\[
P(w \mid z) = \frac{\sum_{d \in D} n(d, w)P(z \mid d, w)}{\sum_{w \in W} \sum_{d \in D} n(d, w)P(z \mid d, w)}
\]

\[
P(d \mid z) = \frac{\sum_{w \in W} n(d, w)P(z \mid d, w)}{\sum_{d \in D} \sum_{w \in W} n(d, w)P(z \mid d, w)}
\]

\[
P(z) = \frac{\sum_{d \in D} \sum_{w \in W} n(d, w)P(z \mid d, w)}{\sum_{w \in W} \sum_{d \in D} n(d, w)}
\]
Coding Design

- Variables:
  - \texttt{double[][] p\_dz\_n} // \( p(d|z), |D|*|Z| \)
  - \texttt{double[][] p\_wz\_n} // \( p(w|z), |W|*|Z| \)
  - \texttt{double[] p\_z\_n} // \( p(z), |Z| \)

- Running Processing:
  1. Read dataset from file
     
     \texttt{ArrayList<DocWordPair> doc; // all the docs}
     
     \hspace{1cm} 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 ( \( |\text{Log-likelihood} – \text{old Log-likelihood}| < \) threshold)
  4. Output \( p\_dz\_n, p\_wz\_n \) and \( p\_z\_n \)
• **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} = \frac{\text{nominator}[\( j \)]}{\text{denominator}}; \)
      nominator_p_dz_n[d][z] += tf_w*d*P_{\_z\_condition\_d\_w};  
      denominator_p_dz_n[z] += tf_w*d*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 \( z \) {  
    \( p_{dz\_n\_new}[d][z] = \frac{\text{nominator}_p_{dz\_n}[d][z]}{\text{denominator}_p_{dz\_n}[z]}; \)
  } // 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[$z$] = $p_{dz\_n}[d][z] \times p_{wz\_n}[w][z] \times p_{z\_n}[z]$
        - denominator += nominator[$z$];
    - } // end for each topic $z$
  - For each topic $z$
    - $P_{z\_condition\_d\_w} =$ nominator[$z$]/denominator;
    - nominator$_{p_{wz\_n}}[w][z]$ += $tf_{wd} \times P_{z\_condition\_d\_w}$;
    - denominator$_{p_{wz\_n}}[z]$ += $tf_{wd} \times P_{z\_condition\_d\_w}$;
  - } // 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 $p_{z\cdot n}$**
  
  For each doc $d$ {
  
  For each word $w$ included in $d$ {
  
  denominator = 0;
  
  nominator = new double[Z];
  
  For each topic $z$ {
  
  nominator[z] = $p_{dz\cdot n}[d][z]$ * $p_{wz\cdot n}[w][z]$ * $p_{z\cdot n}[z]$
  
  denominator += nominator[z];
  
  } // end for each topic $z$
  
  For each topic $z$ {
  
  $P_{z\cdot condition\cdot d\cdot w} = nominator[z]/denominator$
  
  nominator_p_z_n[z] += tf_{wd} * $P_{z\cdot condition\cdot d\cdot w}$;
  
  } // 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 $z$
  
  $p_{dz\cdot n\cdot new}[d][z] = nominator_p_z_n[z]/denominator_p_z_n$;
  
  } // end for each topic $z$
Latent Dirichlet Allocation (LDA)
Model

Generative Process:
For each document $d$ in the corpus, the LDA model
1. picks a multinomial distribution $\theta^d = [\theta^d_1 \ldots \theta^d_K]^T$ from the dirichlet distribution $\alpha = [\alpha_1 \ldots \alpha_K]^T$
2. assigns a topic $z_{di} = k$ to the $i^{th}$ word in the document $d$ according to the multinomial distribution $\theta^d$.
3. given a topic $z_{di} = k$, the model then picks a word $w_{di}$ from the vocabulary of $V$ words according to the multinomial distribution $[\Phi^k_1 \ldots \Phi^k_V]^T$ which is generated from the Dirichlet distribution $[\beta^k_1 \ldots \beta^k_V]^T$.
Joint Distribution

- Given the parameters $\alpha$ and $\beta$, the joint distribution of a topic mixture $\theta$, a set of $N$ topics $z$, and a set of $N$ words $w$ is given by:

$$p(\theta, z, w \mid \alpha) = p(\theta \mid \alpha) \prod_{i=1}^{N_d} p(z_{di} \mid \theta)p(w_{di} \mid z_{di}, \beta)$$

- where $p(z_{di} \mid \theta)$ is simply $\theta_z$ for the unique $z$. Integrating over $\theta$ and summing over $z$ we obtain the marginal distribution of a document

$$p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left( \prod_{i=1}^{N_d} \sum_{z_{di}} p(z_{di} \mid \theta)p(w_{di} \mid z_{di}, \beta) \right) d\theta$$

- Taking the products of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{D} \int p(\theta_d \mid \alpha) \left( \prod_{i=1}^{N_d} \sum_{z_{di}} p(z_{di} \mid \theta_d)p(w_{di} \mid z_{di}, \beta) \right) d\theta_d$$
Three-level Representation

- $\alpha, \beta, \Phi$ are corpus-level parameters
  - Sampled once when generating a corpus
- $\theta_d$ are document-level variables
  - Sampled once per document
- $z_{di}, w_{di}$ are word-level variables
  - Sampled once for each word in each document

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

Unigram

\[ p(w) = \prod_{n=1}^{N} p(w_n) \]

Mixture of unigrams

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

PLSI/aspect model

\[ p(d, w_n) = p(d) \sum_{z} p(w_n \mid z) p(z \mid d) \]

Problems: (1) \( p(z \mid d) \) 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.
Inference and Parameter Estimation

\[ p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left( \prod_{i=1}^{N_d} \sum_{z=1}^{K} p(z \mid \theta) p(w_i \mid z, \beta) \right) d\theta \]

• The key inferential problem is that of computing the posterior distribution of the hidden variables given a document:

\[ p(\theta, z \mid w, \alpha, \beta) = \frac{p(\theta, z, w \mid \alpha, \beta)}{p(w \mid \alpha, \beta)} \quad (a) \]

• However, such distribution is intractable to compute in general. For normalization in the above distribution, we have to marginalize over the hidden variables and write the Equation (a) in terms of the model parameters:

\[ p(w \mid \alpha, \beta) = \frac{\Gamma(\sum_z \alpha_z)}{\prod_z \Gamma(\alpha_z)} \int \left( \prod_{z=1}^{K} \theta^{\alpha_z - 1} \right) \left( \prod_{i=1}^{N_d} \sum_{z=1}^{K} \prod_{\nu=1}^{V} (\theta z \beta_{z\nu})^{w_{iz}} \right) d\theta \]
Inference and Parameter Estimation (cont.)

\[ p(w | \alpha, \beta) = \frac{\Gamma(\sum_z a_z)}{\prod_z \Gamma(a_z)} \int \left( \prod_{z=1}^{K} \theta_z^{a_z - 1} \right) \left( \prod_{i=1}^{N_d} \sum_{z=1}^{K} \prod_{v=1}^{V} (\theta_z \beta_{zv})^{w_{iz}} \right) d\theta \]

- This function is intractable due to the coupling between \( \theta \) and \( \beta \) in the summation over latent topics (Dickey, 1983).

- Rather than the intractable exact inference, we can use some other approximate inference algorithms, e.g., Laplace approximation, variational approximation, and Markov chain Monte Carlo (Jordan, 1999).
Gibbs Sampling Derivation for LDA
Notations

- $D$: #documents in the whole corpus
- $K$: #topics
- $V$: #vocabulary
- $w, z$: bold font means words or topics in the whole corpus
- $d$: the index of document, $d = 1$ to $D$
- $N_d$: #words in $d^{th}$ document
- $i$: the index of word in a document, $i=1$ to $N_d$
  - $w_{di}$: $i^{th}$ word of $d^{th}$ document
  - $z_{di}$: $i^{th}$ word of $d^{th}$ document is denoted by topic $z_{di}$
- $z$: the index of topic, $z=1$ to $K$
- $v$: the index of word in the vocabulary, $v=1$ to $V$
- $n_{z,v}$: the number of times word $v$ is generated by topic $z$
- $n_{d,z}$: the number of times topic $z$ occurs in document $d$
Notations

- $\phi_z$: the topic-words distribution corresponding to topic $z$
- $\theta_d$: the document-topic distribution corresponding to document $d$
- $\beta$: $K$ vector, $\phi \sim \text{Dirichlet}(\beta)$
- $\alpha$: $K$ vector, $\theta \sim \text{Dirichlet}(\alpha)$
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 \(p(Z|X=x_{i+1}, Y=y_{i+1})\)
Gibbs Sampling Derivation for LDA

- To apply Gibbs Sampling, we need the full conditional distribution \( p(z_{di}|w,z_{-di}) \). Then \( w_{di} \) and \( z_{di} \) can be sampled based on the probability. When the sampling completes, each \( w_{di} \) has assigned a topic, and each document also has assigned topics. So the document-topics distribution \( \theta_d \) and topic-words distribution \( \Phi_z \) can be easily estimated.

- According to and bayesian formula, we can get:

\[
p(z_{di}, z_{-di}) = p(z) \quad \text{and bayesian formula, we can get:}
\]

\[
p(z_{di} | w, z_{-di}) = \frac{p(z, w)}{p(z_{-di}, w)} \propto \frac{p(z, w)}{p(z_{-di}, w_{-di})}
\]

\( w_{di} \) is generated by \( z_{di} \), so we can ignore it if \( z_{di} \) is not considered.

- The key point is to calculate the joint distribution:

\[
p(z, w) = p(w | z, \beta)p(z | \alpha)
\]

Then we calculate \( p(z|\alpha) \) and \( p(w|z,\beta) \) respectively.
First $p(z|\alpha)$

Import $\Theta$:

$$p(z|\alpha) = \int p(z|\Theta)p(\Theta|\alpha)d\Theta$$  \hspace{1cm} (1)

Expand all the bold and capital character:

$$p(z|\alpha) = \int \prod_{d=1}^{D} \prod_{i=1}^{N_d} p(z_{di}|\theta_d)p(\theta_d|\alpha)d\theta_d$$  \hspace{1cm} (2)

Because $z_n \sim \text{Multinomial}(\theta)$ and $\theta \sim \text{Dir}(\alpha)$:

$$p(z|\alpha) = \int \prod_{d=1}^{D} \prod_{i=1}^{K} \theta_{di,z}^{n_{di,z}} \prod_{d=1}^{D} \left( \frac{\Gamma\left(\sum_{z=1}^{K} \alpha_z\right)}{\prod_{z=1}^{K} \Gamma(\alpha_z)} \prod_{z=1}^{K} \theta_{d,z}^{\alpha_z-1} \right)d\theta_d$$  \hspace{1cm} (3)

Because $\frac{\Gamma\left(\sum_{z=1}^{K} \alpha_z\right)}{\prod_{z=1}^{K} \Gamma(\alpha_z)}$ is unrelated with $\theta$:

$$p(z|\alpha) = \left( \frac{\Gamma\left(\sum_{z=1}^{K} \alpha_z\right)}{\prod_{z=1}^{K} \Gamma(\alpha_z)} \right)^D \int \prod_{d=1}^{D} \prod_{i=1}^{K} \theta_{di,z}^{n_{di,z}+\alpha_z^{-1}}d\theta_d$$  \hspace{1cm} (4)

For different document, its document-topics distribution $\theta_d$ is independent, so:

$$p(z|\alpha) = \left( \frac{\Gamma\left(\sum_{z=1}^{K} \alpha_z\right)}{\prod_{z=1}^{K} \Gamma(\alpha_z)} \right)^D \int \prod_{d=1}^{D} \prod_{i=1}^{K} \theta_{d,z}^{n_{d,z}+\alpha_z^{-1}}d\theta_d$$  \hspace{1cm} (5)
For different document, its document-topics distribution $\theta_d$ is independent, so:

$$
p(z | \alpha) = \left( \frac{\Gamma(\sum_{z=1}^{K} \alpha_z)}{\prod_{z=1}^{K} \Gamma(\alpha_z)} \right)^D \prod_{d=1}^{D} \prod_{z=1}^{K} \theta_d^{n_{d,z} + \alpha_z - 1} \ d\theta_d
$$

Because

$$
\int \text{Dir}(\theta | \alpha) d\theta = \int \prod_{z=1}^{K} \frac{\Gamma(\alpha_z)}{\Gamma(\sum_{z=1}^{K} \alpha_z)} \prod_{z=1}^{K} \theta_z^{\alpha_z - 1} d\theta = 1
$$

$$
\therefore \quad \frac{\prod_{z=1}^{K} \Gamma(\alpha_z)}{\Gamma(\sum_{z=1}^{K} \alpha_z)} = \int \prod_{z=1}^{K} \theta_z^{\alpha_z - 1} d\theta
$$

We get

$$
p(z | \alpha) = \left( \frac{\Gamma(\sum_{z=1}^{K} \alpha_z)}{\prod_{z=1}^{K} \Gamma(\alpha_z)} \right)^D \prod_{d=1}^{D} \prod_{z=1}^{K} \frac{\Gamma(n_{d,z} + \alpha_z)}{\Gamma(\sum_{z=1}^{K} (n_{d,z} + \alpha_z))}
$$
\[
\frac{p(z|\alpha)}{p(z_{-di}|\alpha)} = \frac{\frac{\left(\prod_{z=1}^{K} \Gamma(z)\right)^{D}}{\prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z)}}{\frac{\left(\prod_{z=1}^{K} \Gamma(z)\right)^{D}}{\prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z)}} = 1
\]

Ignoring \(d_i\) means ignoring \(z_{d_i}\) and \(w_{d_i}\). That is to say, when considering the times of co-occurrence of topic \(z\) and word \(v\) \((n_{z,v})\), ignore this co-occurrence of \(z_{d_i}\) and \(w_{d_i}\). This makes \(n_{z_{d_i},w_{d_i}}^{-1}\), \(n_{d_{d_i},z_{d_i}}^{-1}\), however, do not impact \(n_{z,w_d}\) and \(n_{d,z_d}\)

\[
= \frac{\prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z)}{\prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z)} = 1
\]

\[
\prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z) - \prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z)
\]

\[
\prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z) - \prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z)
\]

\[
\prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z) - \prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z)
\]

\[
\prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z) - \prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z)
\]

We get:

\[
\prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z) - \prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z)
\]

\[
\prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z) - \prod_{d=1}^{D} \prod_{z=1}^{K} \Gamma(n_{d,z} + \alpha_z)
\]

Because:

\[
\Gamma(x + 1) = x\Gamma(x)
\]
The Full Conditional Probability

The same as (1)-(9), we can get:

\[
\frac{p(w | z, \beta)}{p(w_{-di} | z_{-di}, \beta)} = \frac{n_{z_{di}, w_{di}} + \beta_{w_{di}} - 1}{\sum_{v=1}^{V} (n_{z_{di}, v} + \beta_v) - 1}
\]  

(10)

Finally, the full conditional probability is:

\[
p(z_{di} | w, z_{-di}) \propto \frac{p(z_{di}, w)}{p(z_{-di}, w_{-di})} = \frac{p(w | z, \beta)p(z | \alpha)}{p(w_{-di} | z_{-di}, \beta)p(z_{-di} | \alpha)}
\]

\[
= \frac{n_{z_{di}, w_{di}} + \beta_{w_{di}} - 1}{\sum_{v=1}^{V} (n_{z_{di}, v} + \beta_v) - 1} \cdot \frac{n_{d_{di}, z_{di}} + \alpha_{z_{di}} - 1}{\sum_{z=1}^{K} (n_{d_{di}, z} + \alpha_z) - 1}
\]

(11)

Then we can use the full conditional probability to sample \( z_{di} \).
Finally

Finally, we need to calculate the topic of a new observation (word $w$), hence we can evaluate the distribution of a new topic-word pair $(z, w)$ by

$$p(z = k, w = t | w, z) = \frac{p(z = k, w = t, w, z)}{p(w, z)}$$

$$= \frac{\Gamma(n_{z \delta_i, w \delta_i} + 1 + \beta_{w \delta_i}) \cdot \Gamma(n_{d \delta_i, z \delta_i} + 1 + \alpha_{z \delta_i})}{\Gamma(n_{z \delta_i, w \delta_i} + \beta_{w \delta_i}) \cdot \Gamma(n_{d \delta_i, z \delta_i} + \alpha_{z \delta_i})} \cdot \frac{\Gamma(\sum_{v=1}^{V} (n_{z \delta_i, v} + 1 + \beta_{v})) \cdot \Gamma(\sum_{z=1}^{K} (n_{d \delta_i, z} + 1 + \alpha_{z}))}{\Gamma(\sum_{v=1}^{V} (n_{z \delta_i, v} + \beta_{v})) \cdot \Gamma(\sum_{z=1}^{K} (n_{d \delta_i, z} + \alpha_{z}))}$$

$$= \frac{n_{z \delta_i, w \delta_i} + \beta_{w \delta_i}}{\sum_{v=1}^{V} (n_{z \delta_i, v} + \beta_{v})} \cdot \frac{n_{d \delta_i, z \delta_i} + \alpha_{z \delta_i}}{\sum_{z=1}^{K} (n_{d \delta_i, z} + \alpha_{z})}$$
Multinomial Parameters

During parameter estimation, the algorithm keeps track of a $D \times K$ (document by topic) count matrix, a $K \times V$ (topic by word) count matrix. Given these matrices, we can easily estimate the probability $\theta_{dz}$ of a topic given a document and the probability $\Phi_{zv}$ of a word given a topic by:

\[
\theta_d^z = \frac{n_{d,z} + \alpha_z - 1}{\sum_{z=1}^{K} (n_{d,z} + \alpha_z) - 1}, \quad \Phi_z^v = \frac{n_{z,v} + \beta_v - 1}{\sum_{v=1}^{V} (n_{z,v} + \beta_v) - 1}
\]

Given a new document $d$, we perform inference by

\[
P(z_i = j \mid z_{-i}, w_i, d_i, \cdot) \propto \frac{n_{k}^{(i)} + n_{k}^{(i)'} + \beta_i}{\sum_{v=1}^{V} (n_{k}^{(v)} + n_{k}^{(v)'} + \beta_v)} \cdot \frac{n_{m}^{(k)} + \alpha_k}{\sum_{z=1}^{Z} (n_{m}^{z} + \alpha_z)}
\]

where $n'$ denotes the number in the new document.
Coding

• Init()

• Estimate()
  – Foreach doc
    • Foreach word
      – Sampling()
        » Compute probability $P(z_i|\mathbf{w}, \mathbf{z}_{-i})$;
        » Sampling a topic $z$ based on $P(z_i|\mathbf{w}, \mathbf{z}_{-i})$;

  – Compute_theta()

  – Compute_phi()
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)

<table>
<thead>
<tr>
<th>River</th>
<th>Stream</th>
<th>Bank</th>
<th>Money</th>
<th>Loan</th>
</tr>
</thead>
<tbody>
<tr>
<td>○○○○○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>●○○●○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>○○○○○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>●○○●○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>○○○○○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>●○○●○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>○○○○○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>●○○●○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>○○○○○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>●○○●○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>○○○○○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>●○○●○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>○○○○○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>●○○●○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>○○○○○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>●○○●○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>○○○○○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
<tr>
<td>●○○●○</td>
<td>○●●○○</td>
<td>●○○●○</td>
<td>○●●●○</td>
<td>○●○○○</td>
</tr>
</tbody>
</table>
After 1 iteration

- Apply sampling equation to each word token

<table>
<thead>
<tr>
<th>River</th>
<th>Stream</th>
<th>Bank</th>
<th>Money</th>
<th>Loan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
After 4 iterations

<table>
<thead>
<tr>
<th>River</th>
<th>Stream</th>
<th>Bank</th>
<th>Money</th>
<th>Loan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
After 32 iterations

<table>
<thead>
<tr>
<th>River</th>
<th>Stream</th>
<th>Bank</th>
<th>Money</th>
<th>Loan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**topic 1**

<table>
<thead>
<tr>
<th>stream</th>
<th>.40</th>
</tr>
</thead>
<tbody>
<tr>
<td>bank</td>
<td>.35</td>
</tr>
<tr>
<td>river</td>
<td>.25</td>
</tr>
</tbody>
</table>

**topic 2**

<table>
<thead>
<tr>
<th>bank</th>
<th>.39</th>
</tr>
</thead>
<tbody>
<tr>
<td>money</td>
<td>.32</td>
</tr>
<tr>
<td>loan</td>
<td>.29</td>
</tr>
</tbody>
</table>
Parameter Estimation with Gibbs Sampling

Initialization:
  Randomly assign topic for each item in each document and initialize count matrices.

Iteration (MCMC Gibbs sampling step)
  For each Iteration
    For each document
      For each item
        assign a topic from full conditional probability with input of $\alpha$ and $\beta$
        update count matrices.
      End for
    End for
  End for

Calculate document-topic and topic-item distribution with count matrices.
End for
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 $\alpha$ and $\beta$
        update count matrices.
      End for
    End for
  End for
End for

Calculate document-topic and topic-item distribution with count matrices.

Estimate hyper-parameter $\alpha$ and $\beta$ with input of updated count matrix.

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 $\alpha$ can be obtained by

$$p(\alpha | \bar{z}) \propto p(\bar{z} | \alpha) p(\alpha)$$

$$\propto \prod_m \frac{\Delta(n_m + \alpha)}{\Delta_K(\alpha)} \alpha^{a-1} e^{-b\alpha}$$

Then for each parameter, we have

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

$$\propto p(\bar{z} | \bar{\alpha}) p(\bar{\alpha})$$

$$\propto \prod_m \left( \frac{\Gamma(n_{mi} + \alpha_i)}{\Gamma(\sum_k (n_{mk} + \alpha_k))} \cdot \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\alpha_i)} \right) \cdot \alpha_i^{a-1} e^{-b\alpha_i}$$
Analogously, the full conditional for Gibbs sampling of $\beta$ can be obtained by

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

For each parameter, we have

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

$$
\propto p(\bar{z} | \bar{\beta}; \bar{w})p(\bar{\beta}; \bar{w})
$$

$$
\propto \prod_k \left( \frac{\Gamma(n_{ki} + \beta_i)}{\Gamma(\sum_y (n_{ky} + \alpha_y))} \cdot \frac{\Gamma(\sum_y \beta_y)}{\Gamma(\beta_i)} \right) \beta_i^{a-1} e^{-b\beta_i}
$$

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 $\alpha$ and $\Phi$.
  - Use different algorithms to recover both $\Phi$ and the hyperparameters $\alpha$.
  - Compare the effectiveness of different learned algorithms.
  - Compare the efficiency of different algorithms.

$\alpha=\{0.5, 1.5, 0.5, 1.5, 0.5, 1.5, 0.5\}$  

$\Phi=\{A, B, C, D, E, F, G\}$

“document” X pixel

recover

$\alpha=?$

$\Phi=?$
Performance of Recovering $\Phi$

- Normal convergence process of topic-word ($\Phi$) distribution
Performance of Recovering $\Phi$

- Topic-item similarity traces over Gibbs sampling steps.
Performance of Recovering $\alpha$

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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Newton</td>
<td>0.53</td>
<td>1.56</td>
<td>0.65</td>
<td>1.41</td>
<td>0.48</td>
<td>1.64</td>
<td>0.56</td>
</tr>
<tr>
<td>FixedPoint</td>
<td>0.49</td>
<td>1.47</td>
<td>0.53</td>
<td>1.37</td>
<td>0.46</td>
<td>1.48</td>
<td>0.89</td>
</tr>
<tr>
<td>LOO</td>
<td>0.49</td>
<td>1.19</td>
<td>0.44</td>
<td>1.25</td>
<td>0.57</td>
<td>1.08</td>
<td>0.68</td>
</tr>
<tr>
<td>MH</td>
<td>0.56</td>
<td>1.47</td>
<td>0.67</td>
<td>1.62</td>
<td>0.36</td>
<td>1.70</td>
<td>0.76</td>
</tr>
<tr>
<td>ARS</td>
<td>0.54</td>
<td>1.32</td>
<td>0.69</td>
<td>1.51</td>
<td>0.53</td>
<td>1.49</td>
<td>0.64</td>
</tr>
</tbody>
</table>
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
Joint Distribution

For each word, we have

\[
p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left( \prod_{i=1}^{N_d} \sum_{z_{di}} p(z_{di} \mid \theta) p(w_{di} \mid z_{di}, \beta) \right) d\theta
\]

\[
= \frac{\Gamma(\sum_{z} \alpha_z)}{\prod_{z} \Gamma(\alpha_z)} \int \left( \prod_{i=1}^{K} \theta_{z_{di}}^{\alpha_z-1} \right) \left( \prod_{i=1}^{N_d} \sum_{z_{di}} \sum_{v=1}^{V} (\theta_z \beta_{zy}) w_{i}^{v} \right) d\theta
\]

- We can write the probability of a corpus:

\[
p(D \mid \alpha, \beta) = \prod_{d=1}^{D} \int p(\theta_d \mid \alpha) \left( \prod_{i=1}^{N_d} \sum_{z_{di}} p(z_{di} \mid \theta_d) p(w_{di} \mid z_{di}, \beta) \right) d\theta_d
\]
Variational Inference

- Here, we introduce a simple convexity-based variational algorithm for inference in LDA.

- The basic idea here is to make use of Jensen’s inequality

\[
\log p(w | \alpha, \beta) = \log \int \sum_z \frac{p(\theta, z, w | \alpha, \beta) q(\theta, z)}{q(\theta, z)} d\theta
\]

\[
\geq \log \int \sum_z \left( \frac{p(\theta, z, w | \alpha, \beta)}{q(\theta, z)} \right)^{q(\theta,z)} d\theta = \int \sum q(\theta, z) \log \frac{p(\theta, z, w | \alpha, \beta)}{q(\theta, z)} d\theta
\]

\[
= \int \sum q(\theta, z) \log p(\theta, z, w | \alpha, \beta) d\theta - \int \sum q(\theta, z) \log q(\theta, z) d\theta
\]

\[
= E_q \left[ \log p(\theta, z, w | \alpha, \beta) \right] - E_q \left[ \log q(\theta, z) \right]
\]

- A simple way to obtain a tractable family of lower bounds is to consider simple modifications of the original graphical model in which some of the edges and nodes are removed.
Variational Inference (cont.)

• Hence, by dropping edges between $\theta$, $z$, and $w$, and $w$ nodes, and also endow the resulting simplified graphical model with free variational parameters, we obtain a family of distributions on the latent variables:

$$q(\theta, z | \gamma, \phi) = q(\theta | \gamma) \prod_{i=1}^{N_d} q(z_i | \phi_i)$$

• where the Dirichlet parameter $\gamma$ and the multinomial parameters ($\Phi_1, \ldots, \Phi_N$) are the free variational parameters.
How to determine the variational parameters

- We can set up an optimization problem to determine the values of the variational parameters $\gamma$ and $\Phi$.
- We can define the optimization function as minimizing the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior $p(\theta, z | w, \alpha, \beta)$:

$$ (\gamma^*, \phi^*) = \arg \min_{(\gamma, \phi)} D(q(\theta, z | \gamma, \phi) \| p(\theta, z | w, \alpha, \beta) \quad (5) $$

- This minimization can be achieved by an iterative fixed-point method.
Variational Inference

- We now discuss how to set the parameter $\gamma$ and $\Phi$ via an optimization procedure.
- Following Jordan et al. (1999), we have a lower bound of the log likelihood of a document using Jensen’s inequality:

$$
p(w | \alpha, \beta) = \log \int \sum_z p(\theta, z, w | \alpha, \beta) d\theta$$

$$= \log \int \sum_z \frac{p(\theta, z, w | \alpha, \beta)q(\theta, z)}{q(\theta, z)} d\theta$$

$$> \int \sum_z q(\theta, z) \log p(\theta, z, w | \alpha, \beta) d\theta - \int \sum_z q(\theta, z) \log q(\theta, z) d\theta$$

$$= E_q[\log p(\theta, z, w | \alpha, \beta)] - E_q[\log q(\theta, z)]$$
Variational Inference (cont.)

- Then from the above formula, we see that the Jensen’s inequality provides us with a lower bound on the log likelihood for an arbitrary variational distribution \( q(\theta, z|\gamma, \Phi) \).

- It can be easily verified that the difference between the left-hand side and the right-hand side of the above equation is the KL divergence between the variational posterior probability and the true posterior probability.
Variational Inference (cont.)

• That is, letting $L(\phi, \gamma; \alpha, \beta)$ denote the right-hand side of the above equation we have:

$$\log p(w \mid \alpha, \beta) = L(\gamma, \phi; \alpha, \beta) + \text{D}(q(\theta, z \mid \gamma, \phi) \mid\mid p(\theta, z \mid w, \alpha, \beta))$$

• This means that maximizing the lower bound $L(\phi, \gamma; \alpha, \beta)$ w.r.t. $\gamma$ and $\Phi$ is equivalent to minimizing the KL divergence between the variational posterior probability and the true posterior probability, the optimization problem in the above equation.
Variational Inference (cont.)

- We can expand the above equation

\[
L(\phi, \gamma; \alpha, \beta) = \log p(w | \alpha, \beta) - D(q(\theta, z | \gamma, \phi) \| p(\theta, z | w, \alpha, \beta))
\]

\[
= E_q \left[ \log p(\theta, z, w | \alpha, \beta) \right] - E_q \left[ \log q(\theta, z) \right]
\]

\[
= E_q \left[ \log p(\theta | \alpha) \right] + E_q \left[ \log p(z | \theta) \right] + E_q \left[ \log p(w | z, \beta) \right]
\]

\[
- E_q \left[ \log q(\theta) \right] - E_q \left[ \log q(z) \right]
\]

- By extending it again, we can have

\[
L(\phi, \gamma; \alpha, \beta) = \log \Gamma(\sum_{z=1}^{K} \alpha_z) - \sum_{z=1}^{K} \log \Gamma(\alpha_z) + \sum_{z=1}^{K} (\alpha_z - 1)(\Psi(\gamma_z) - \Psi(\sum_{z=1}^{K} \gamma_z))
\]

\[
+ \sum_{i=1}^{N_d} \sum_{z=1}^{K} \phi_{zi} (\Psi(\gamma_z) - \Psi(\sum_{z=1}^{K} \gamma_z)) + \sum_{i=1}^{N_d} \sum_{z=1}^{K} \sum_{v=1}^{V} \phi_{zi} w_i^v \log \beta_{zv} - \log \Gamma(\sum_{z=1}^{K} \gamma_z)
\]

\[
+ \sum_{z=1}^{K} \log \Gamma(\gamma_z) - \sum_{z=1}^{K} (\gamma_z - 1)(\Psi(\gamma_z) - \Psi(\sum_{z=1}^{K} \gamma_z)) - \sum_{i=1}^{N_d} \sum_{z=1}^{K} \phi_{zi} \log \phi_{zi}
\]
Variational Multinomial

• We first maximize above Eq w.r.t. $\Phi_{ni}$, the probability that the $n$-th word is generated by topic $i$.

• We form the Lagrangian by isolating the terms which contain $\Phi_{ni}$ and adding the appropriate Lagrange multipliers. Let $\beta_{iv}$ be $p(w_{vn}=1|z^i=1)$ for the appropriate $v$. (recall that each $w_n$ is a vector of size $V$ with exactly one component equal to one; we can select the unique $v$ such that $w_{vn}=1$):

$$L_{[\phi_{ni}]} = \phi_{ni} \left( \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right) + \phi_{ni} \log \beta_{iv} - \phi_{ni} \log \phi_{ni} + \lambda_n \left(\sum_{j=1}^{k} \phi_{ni} - 1\right)$$
Variaitonal Multinominal (cont.)

• Taking derivatives w.r.t. $\Phi_{ni}$, we obtain:

$$\frac{\partial L}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) + \log \beta_{iv} - \log \phi_{ni} - 1 + \lambda$$

• Setting this to zero yields the maximizing value of the variational parameter $\Phi_{ni}$:

$$\phi_{ni} \propto \beta_{iv} \exp(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j))$$
Variational Dirichlet

• Next we maximize equation (15) w.r.t. $\gamma_i$, the $i$-th component of the posterior Dirichlet parameters, the terms containing $\gamma_i$ are:

$$L_{[\gamma]} = \sum_{i=1}^{k} (\alpha_i - 1) (\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)) + \sum_{n=1}^{N} \phi_{ni} (\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j))$$

$$- \log \Gamma(\sum_{j=1}^{k} \gamma_j) + \log \Gamma(\gamma_i) - \sum_{i=1}^{k} (\gamma_i - 1) (\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j))$$

• By simplifying

$$L_{[\gamma]} = \sum_{i=1}^{k} (\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)) - \log \Gamma(\sum_{j=1}^{k} \gamma_j) + \log \Gamma(\gamma_i)$$
Variational Dirichlet (cont.)

• Taking the derivative w.r.t. $\gamma_i$:

$$\frac{\partial L}{\partial \gamma_i} = \Psi'(\gamma_i)(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i) - \Psi'(\sum_{j=1}^{k} \gamma_j) \sum_{i=1}^{k} (\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i)$$

$$= \sum_{i=1}^{k} \left( \alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i \right) \left( \Psi'(\gamma_i) - \Psi'\left(\sum_{j=1}^{k} \gamma_j\right) \right)$$

• Setting this equation to zero yields a maximum at:

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$$
Solve the Optimization Problem

• Derivate the KL divergence and setting them equal to zero, we obtain the following update equations:

\[
\phi_{ni} \propto \beta_{iw_n} \exp\{E_q[\log(\theta_i) | \gamma]\}
\]

\[
\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}
\]

where the expectation in the multinomial update can be computed as follows:

\[
E_q[\log(\theta_i) | \gamma] = \Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)
\]

where \(\psi\) is the first derivative of the log\(\Gamma\) function which is computable via Taylor approximations.
Computing $E[\log(\theta_i|\alpha)]$

- Recall that a distribution is in the exponential family if it can be written in the form:
  \[ p(x | \eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\} \]
  
  where $\eta$ is the natural parameter, $T(x)$ is the sufficient statistic, and $A(\eta)$ is the log of the normalization factor.

- As we can write the Dirichlet in this form by exponentiating the log of Eq.: $p(\theta|\alpha)$
  \[ p(\theta | \alpha) = \exp\{\left(\sum_{i=1}^{k} (\alpha_i - 1) \log \theta_i\right) + \log \Gamma(\sum_{i=1}^{k} \alpha_i) - \sum_{i=1}^{k} \log \Gamma(\alpha_i)\} \]
Computing $E[\log(\theta_i|\alpha)]$ (cont.)

- From this form we see that the natural parameter of the Dirichlet is $\eta_i = \alpha_i - 1$ and the sufficient statistic is $T(\theta_i) = \log \theta_i$. Moreover, based on the general fact that the derivative of the log normalization factor w.r.t. the natural parameter is equal to the expectation of the sufficient statistic, we obtain:

$$E[\log(\theta_i) \mid \alpha] = \Psi(\alpha_i) - \Psi(\sum_{j=1}^{k} \alpha_j)$$

where $\psi$ is the digamma function, the first derivative of the log Gamma function.
Variational Inference Algorithm

(1) initialize $\phi_{ni}^0 = \frac{1}{k}$ for all $i$ and $n$

(2) initialize $\gamma_i = \alpha_i + \frac{N}{k}$ for all $i$

(3) repeat
(4) for $n=1$ to $N$
(5) for $i=1$ to $k$
(6) $\phi_{ni}^{t+1} = \beta_{iw_n} \exp(\Psi(\gamma_i^t))$
(7) normalize $\phi_{ni}^{t+1}$ to sum to 1
(8) $\gamma_i^t = \alpha + \sum_{n=1}^{N} \phi_{ni}^{t+1}$
(9) until convergence

Each iteration requires $O((N+1)k)$ operations

For a single document
the iteration number
is on the order of the
number of words in it

Thus, the total number
of operations roughly
on the order of $N^2k$
Parameter Estimation

• We can use an empirical Bayes method for parameter estimation. In particular, we wish to find parameters $\alpha$ and $\beta$ that maximize the marginal log likelihood:

$$
\log(\alpha, \beta) = \sum_{d=1}^{M} \log p(w_d | \alpha, \beta)
$$

• The quantity $p(w|\alpha, \beta)$ can be computed by the variational inference as described above. Thus, we can find approximate empirical Bayes estimates for the LDA model via an alternating variational EM procedure that maximizes a lower bound w.r.t. the variational parameters $\gamma$ and $\Phi$, and then fixed values of the variational parameters, maximizes the lower bound w.r.t. the model parameter $\alpha$ and $\beta$. 
Variational EM

1. (E-step) For each document, find the optimizing values of the variational parameters \( \{\gamma_d^*, \phi_d^* : d \in D\} \). This is done as described in the previous section.

2. (M-step) Maximize the resulting lower bound on the log likelihood w.r.t. the model parameters \( \alpha \) and \( \beta \). This corresponds to finding maximum likelihood estimates with expected sufficient statistics for each document under the approximate posterior which is computed in the E-step. Actually, the update for the conditional multinomial parameter \( \beta \) can be written out analytically:

\[
\beta_{ij} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dni}^* W_{dn}^j
\]  

(9)

The update for \( \alpha \) can be implemented using an efficient Newton-Raphson method. These two steps are repeated until converges.
Parameter Estimation

• We now consider how to obtain empirical Bayes estimates of the model parameters $\alpha$ and $\beta$.
• We solve this problem by using the variational lower bound as a surrogate for the marginal log likelihood, with the variational parameters $\Phi$ and $\gamma$ fixed to the values found by variational inference.
• We then obtain empirical Bayes estimates by maximizing this lower bound w.r.t. the model parameters.
Parameter Estimation (cont.)

- Recall our approach for finding the empirical Bayes estimates is based on a variational EM procedure.
- In the variational E-step, we maximize the bound $L(\phi, \gamma; \alpha, \beta)$ w.r.t. the variational parameter $\gamma$ and $\Phi$. In the M-step, which we describe in this section, we maximize the bound w.r.t. the model parameters $\alpha$ and $\beta$. The overall procedure can thus be viewed as coordinate ascent in $L$. 
Conditional Multinomials

- To maximize w.r.t. $\beta$, we isolate terms and add Lagrange multipliers:

$$L_{[\beta]} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{z=1}^{K} \sum_{v=1}^{V} \phi_{dnz} w_{dn}^v \log \beta_{zv} + \sum_{z=1}^{K} \lambda_z \left( \sum_{v=1}^{V} \beta_{zv} - 1 \right)$$

- Taking the derivative w.r.t. $\beta_{ij}$ and set it to zero, we have

$$\beta_{ij} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^j$$
Dirichlet

- First, we have

\[
L_{[\alpha]} = \sum_{d=1}^{D} \left( \log \Gamma \left( \sum_{z=1}^{K} \alpha_z \right) - \sum_{z=1}^{K} \log \Gamma (\alpha_z) + \sum_{z=1}^{K} (\alpha_z - 1)(\Psi(\gamma_{dz}) - \Psi(\sum_{z=1}^{K} \gamma_{dz})) \right)
\]

- Taking derivative w.r.t. \( \alpha_i \), we obtain:

\[
\frac{\partial L}{\partial \alpha_z} = M \left( \Psi \left( \sum_{z=1}^{K} \alpha_z \right) - \Psi (\alpha_z) \right) + \sum_{d=1}^{D} \left( \Psi (\gamma_{dz}) - \Psi(\sum_{z=1}^{K} \gamma_{dz}) \right)
\]

- This derivative depends on \( \alpha_i \), where \( j \leftrightarrow i \), and we therefore must use an iterative method to find the maximal \( \alpha \). In particular, the Hessian is in the form found in equation:

\[
\frac{\partial L}{\partial \alpha_i \alpha_j} = \delta(i, j)M\Psi'(\alpha_i) - \Psi'(\sum_{z=1}^{K} \alpha_z)
\]
Dirichlet (cont.)

• Finally, we can use the same algorithm to find an empirical Bayes point estimate of $\eta$, the scalar parameter for the exchangeable Dirichlet in the smoothed LDA model as mentioned above.
• Simple Laplace smoothing is no longer justified as a maximum a posteriori method in LDA setting.
• We can then assume that each row in $\mathbf{\beta}_{kxV}$ is independently drawn from an exchangeable Dirichlet distribution. That is to treat $\mathbf{\beta}_i$ as random variables that are endowed with a posterior distribution, conditioned on the data.
Smoothing Model

• Thus we obtain a variational approach to Bayesian inference:

\[
q(\beta_{1:K}, \theta_{1:D}, z_{1:D} \mid \eta, \gamma, \phi) = \prod_{z=1}^{K} \text{Dir}(\beta_z \mid \eta_z) \prod_{i=1}^{N_d} q_d(\theta_d, z_i \mid \phi_i, \gamma_d)
\]

where \( q_d(\theta_d, z_n \mid \phi_n, \gamma_d) \) is the variational distribution defined for LDA as above and the update for the new variational parameter \( \eta \) is as follow:

\[
\eta_{zv} \propto \eta + \sum_{d=1}^{D} \sum_{i=1}^{N_d} \phi_{dnz}^* w_{di}^y
\]
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.

\[
\text{perplexity}(D_{test}) = \exp\left\{- \frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right\}
\]

• 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:

$$predictive \ - \ perplexity(D_{test}) = \exp\left\{- \frac{1}{M} \sum_{d=1}^{M} \log p(w_{d,N_d} \mid w_{d,1:N_d-1})\right\}$$
Our work related to topic model
200 topics have been discovered automatically from the academic network.
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
Shafie and Milios

ICDM 0.23
KDD 0.19
ICDM 0.23
KDD 0.19
...
...
...
...
...
...
...
ICML 0.23
NIPS 0.19
...
...
...
...
...
...
...
NIPS 0.19
ICML 0.23
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
Modeling Multiple Objects in Academic network (KDD’08)
ACT Model 1

Generative process:

1. For each topic $z$, draw $\phi_z$ and $\psi_z$ respectively from Dirichlet priors $\beta_z$ and $\mu_z$;

2. For each word $w_{di}$ in document $d$:
   - draw an author $x_{di}$ from $a_d$ uniformly;
   - draw a topic $z_{di}$ from a multinomial distribution $\theta_{x_{di}}$ specific to author $x_{di}$, where $\theta$ is generated from a Dirichlet prior $\alpha$;
   - draw a word $w_{di}$ from multinomial $\phi_{z_{di}}$;
   - draw a conference stamp $c_{di}$ from multinomial $\psi_{z_{di}}$.

\[
P(z_{di}, x_{di}|z_{-di}, x_{-di}, w, c, \alpha, \beta, \mu) \propto \frac{m_{x_{di}z_{di}}^{-d} + \alpha_{z_{di}}}{\sum_z (m_{x_{di}z_{-di}} + \alpha_z)} \frac{n_{z_{di}w_{di}}^{-d} + \beta_{w_{di}}}{\sum_v (n_{z_{di}v}^{-d} + \beta_v)} \frac{n_{z_{di}c_{di}}^{-d} + \mu_{c_{di}}}{\sum_c (n_{z_{di}c_{-di}}^{-d} + \mu_c)}\]
## Expertise Search Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Object</th>
<th>P@5</th>
<th>P@10</th>
<th>P@20</th>
<th>R-pre</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>Paper</td>
<td>40.0</td>
<td>38.6</td>
<td>37.1</td>
<td>10.0</td>
<td>46.4</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>65.7</td>
<td>44.3</td>
<td>25.0</td>
<td>58.8</td>
<td>73.4</td>
</tr>
<tr>
<td></td>
<td>Conference</td>
<td>51.4</td>
<td>32.9</td>
<td>21.4</td>
<td>47.6</td>
<td>63.1</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>52.4</td>
<td>38.6</td>
<td>27.9</td>
<td>38.8</td>
<td>61.0</td>
</tr>
<tr>
<td>LDA</td>
<td>Paper</td>
<td>31.4</td>
<td>48.6</td>
<td>42.9</td>
<td>13.5</td>
<td>45.8</td>
</tr>
<tr>
<td>AT</td>
<td>Paper</td>
<td>42.9</td>
<td>48.6</td>
<td>42.9</td>
<td>13.1</td>
<td>49.3</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>82.9</td>
<td>45.7</td>
<td>25.7</td>
<td>73.5</td>
<td>78.1</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>62.9</td>
<td>47.1</td>
<td>34.3</td>
<td>43.3</td>
<td>63.7</td>
</tr>
<tr>
<td>ACT1</td>
<td>Paper</td>
<td>42.9</td>
<td>45.7</td>
<td>43.6</td>
<td>16.6</td>
<td>51.0</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>91.4</td>
<td>50.0</td>
<td>26.4</td>
<td>80.0</td>
<td>89.6</td>
</tr>
<tr>
<td></td>
<td>Conference</td>
<td>62.9</td>
<td>41.4</td>
<td>23.6</td>
<td>60.7</td>
<td>72.3</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>65.7</td>
<td>45.7</td>
<td>31.2</td>
<td>52.4</td>
<td>71.0</td>
</tr>
<tr>
<td>ACT2</td>
<td>Paper</td>
<td>42.9</td>
<td>47.1</td>
<td>39.3</td>
<td>15.0</td>
<td>47.7</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>74.3</td>
<td>50.0</td>
<td>25.7</td>
<td>69.4</td>
<td>80.1</td>
</tr>
<tr>
<td></td>
<td>Conference</td>
<td>54.3</td>
<td>41.4</td>
<td>22.1</td>
<td>54.2</td>
<td>63.9</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>57.1</td>
<td>46.2</td>
<td>29.1</td>
<td>46.2</td>
<td>63.9</td>
</tr>
<tr>
<td>ACT3</td>
<td>Paper</td>
<td>42.9</td>
<td>38.6</td>
<td>41.4</td>
<td>17.1</td>
<td>47.0</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>71.4</td>
<td>47.1</td>
<td>25.7</td>
<td>70.0</td>
<td>78.7</td>
</tr>
<tr>
<td></td>
<td>Conference</td>
<td>57.1</td>
<td>38.6</td>
<td>23.6</td>
<td>58.3</td>
<td>65.7</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>57.1</td>
<td>41.4</td>
<td>30.2</td>
<td>48.5</td>
<td>63.8</td>
</tr>
</tbody>
</table>
Work2: Integration of Links
—A Unified Probabilistic Ranking Model (ICDM’08)

Unified Probabilistic Model Combining Random Walk over Heterogeneous Network

- The unified model integrates different kinds of objects, called heterogeneous graph.
- The weights $p(c_i|d_j), p(e_i|c_j)$, $p(d_j|c_i)$, and $p(d_j|e_i)$ 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, $p(\theta_i|d_j), p(\theta_i|q), p(d_j|\theta_i)$, 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.

$$ r(d_j) = \lambda_{dd} \sum_{(d_i,d_j) \in E_{dd}} p(d_i|d_j) \cdot r(d_j) + \lambda_{cd} \sum_{(e_i,d_j) \in E_{de}} P(d_i|e_j) \cdot r(e_j) $$

$$ + \lambda_{ed} \sum_{(e_i,d_j) \in E_{ed}} P(d_i|e_j) \cdot r(c_i) + \lambda_{cd} \sum_{(\theta_i,d_j) \in E_{dt}} P(d_i|\theta_j) \cdot r(\theta_j) $$

$$ r(e_j) = \lambda_{de} \sum_{(d_i,e_j) \in E_{de}} p(e_i|d_j) \cdot r(d_j) $$

$$ r(c_i) = \lambda_{dc} \sum_{(d_j,c_i) \in E_{dc}} p(c_i|d_j) \cdot r(d_j) $$
# Expertise Search Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Object</th>
<th>P@5</th>
<th>P@10</th>
<th>P@20</th>
<th>R-pre</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>Paper</td>
<td>42.9</td>
<td>45.7</td>
<td>41.4</td>
<td>12.0</td>
<td>47.2</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>77.1</td>
<td>47.1</td>
<td>26.4</td>
<td>67.5</td>
<td>85.5</td>
</tr>
<tr>
<td></td>
<td>Conference</td>
<td>51.4</td>
<td>38.6</td>
<td>22.9</td>
<td>48.8</td>
<td>66.0</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>57.1</td>
<td>43.8</td>
<td>30.2</td>
<td>42.8</td>
<td>66.2</td>
</tr>
<tr>
<td>LM</td>
<td>Paper</td>
<td>40.0</td>
<td>38.6</td>
<td>37.1</td>
<td>10.0</td>
<td>46.4</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>65.7</td>
<td>44.3</td>
<td>25.0</td>
<td>58.8</td>
<td>73.4</td>
</tr>
<tr>
<td></td>
<td>Conference</td>
<td>51.4</td>
<td>32.9</td>
<td>21.4</td>
<td>47.6</td>
<td>63.1</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>52.4</td>
<td>38.6</td>
<td>27.9</td>
<td>38.8</td>
<td>61.0</td>
</tr>
<tr>
<td>LM+RW1</td>
<td>Paper</td>
<td>62.9</td>
<td>55.7</td>
<td>44.3</td>
<td>12.9</td>
<td>65.3</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>71.4</td>
<td>48.6</td>
<td>25.7</td>
<td>64.6</td>
<td>83.8</td>
</tr>
<tr>
<td></td>
<td>Conference</td>
<td>60.0</td>
<td>35.7</td>
<td>22.1</td>
<td>53.6</td>
<td>64.6</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>64.8</td>
<td>46.7</td>
<td>30.7</td>
<td>43.7</td>
<td>71.2</td>
</tr>
<tr>
<td>pLSI</td>
<td>Paper</td>
<td>32.5</td>
<td>33.8</td>
<td>30</td>
<td>9.7</td>
<td>40.4</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>65.0</td>
<td>40.0</td>
<td>22.5</td>
<td>60.4</td>
<td>75.5</td>
</tr>
<tr>
<td></td>
<td>Conference</td>
<td>47.5</td>
<td>36.3</td>
<td>21.3</td>
<td>45.1</td>
<td>54.1</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>48.3</td>
<td>36.7</td>
<td>24.6</td>
<td>38.4</td>
<td>56.7</td>
</tr>
<tr>
<td>LDA</td>
<td>Paper</td>
<td>31.4</td>
<td>48.6</td>
<td>42.9</td>
<td>13.5</td>
<td>45.8</td>
</tr>
<tr>
<td>AT</td>
<td>Paper</td>
<td>42.9</td>
<td>48.6</td>
<td>42.9</td>
<td>13.1</td>
<td>49.3</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>82.9</td>
<td>45.7</td>
<td>25.7</td>
<td>73.5</td>
<td>78.1</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>62.9</td>
<td>47.1</td>
<td>34.3</td>
<td>43.3</td>
<td>63.7</td>
</tr>
<tr>
<td>ACT1</td>
<td>Paper</td>
<td>42.9</td>
<td>45.7</td>
<td>43.6</td>
<td>16.6</td>
<td>51.0</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>91.4</td>
<td>50.0</td>
<td>26.4</td>
<td>80.0</td>
<td>89.6</td>
</tr>
<tr>
<td></td>
<td>Conference</td>
<td>62.9</td>
<td>41.4</td>
<td>23.6</td>
<td>60.7</td>
<td>72.3</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>65.7</td>
<td>45.7</td>
<td>31.2</td>
<td>52.4</td>
<td>71.0</td>
</tr>
<tr>
<td>ACT1+RW1</td>
<td>Paper</td>
<td>68.6</td>
<td>61.4</td>
<td>50.7</td>
<td>17.1</td>
<td>66.6</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>80.0</td>
<td>51.4</td>
<td>27.1</td>
<td>77.6</td>
<td>87.4</td>
</tr>
<tr>
<td></td>
<td>Conference</td>
<td>62.9</td>
<td>42.9</td>
<td>23.6</td>
<td>59.5</td>
<td>72.0</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td><strong>70.5</strong></td>
<td><strong>51.9</strong></td>
<td><strong>33.8</strong></td>
<td><strong>51.4</strong></td>
<td><strong>75.4</strong></td>
</tr>
</tbody>
</table>
Work3: Multi-topic based Query-oriented Summarization (SDM’09)

Challenging questions:
- How to identify the topics?
- How to extract the summary for each topic?

Topic 1: Generative probabilistic model
LDA is a three-level hierarchical Bayesian model, in which each item of a collection is
obtain a generative model for classification...
qLDA – Query Latent Dirichlet Allocation
qLDA

1. For each topic \( t \), draw the word distribution \( \Phi_t \) from Dirichlet(\( \beta \))

2. For each query \( q \)
   a) Draw a topic distribution \( \theta_q \) from Dirichlet(\( \alpha_q \))
   b) For each word \( w_{qi} \) in query,
      i) Draw a topic \( t_{qi} \) from \( \theta_q \)
      ii) Draw a word \( w_{qi} \) from \( \Phi_{t_{qi}} \)

3. For each document \( d \)
   a) Draw a topic mixture \( \theta_d \) from Dirichlet(\( \alpha \))
   b) Draw the proportion between words associated with a specific query and those associated with its general topic \( \lambda=p(x=0|d)\sim\text{beta}(\gamma_q, \gamma) \).
   c) For each word \( w_{di} \),
      i) Toss a coin \( x_{di} \) from \( \lambda \) with \( x=\text{bernoulli}(\lambda) \)
      ii) if \( x=0 \),
         a) Draw a topic \( z_{di} \) from \( \theta_q \)
         b) Draw a word \( w_{di} \) from \( \Phi_{z_{di}} \)
      iii) else
         a) Draw a topic \( z_{di} \) from \( \theta_d \)
         b) Draw a word \( w_{di} \) from \( \Phi_{z_{di}} \)
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 ($\theta$, $\theta_q$, $\phi$) randomly;
2. run 200 burn-in sampling iterations using LDA [5];
3. run a two-stage training process iteratively:

   (a) train the model parameters ($\theta$, $\theta_q$, $\phi$) 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 $\theta_{qz}$ is the $z$-th dimension of the multinomial of query $q$ with $1 \leq z \leq T$; $\alpha_{dz}$ is a parameter specific to the topic of document $d$; and $\eta$ is a parameter.

   (b) fix $\phi$, and re-estimate the multinomial $\theta$ to minimize $O_{\xi}$; we employ the algorithm proposed in [20] to optimize $O_{\xi}$, (i.e. run an iterative process to obtain the new $\theta$ 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 $\mu$ 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.

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

DUC 05 (document understanding conference)
Work4: Topic Distributions over Links on Web (ICDM’09)

Introduction

... Our method increases query throughput by 15% over the method proposed by Anh and Moffat [3] while still remaining rank safe results...

... We show that storing inverted lists in memory can significantly improve performance, adding to previous results from Butcher and Clarke [7]...

Related Work

... Impact-sorted indexes are described in a series of papers by Anh and Moffat [1, 2, 3]...

... Fagin et al. proposed a class of algorithms known as threshold algorithms [12]. These algorithms, like the ones shown in this paper, ...

... In our work, we process less index data by organizing the index for easy skipping and query termination. Another way to process less data is to store less data in the index. Static pruning methods remove information from the index that is unlikely to affect query effectiveness. Carmel et al. considered this process [9]. More recently...

Approach

... A similar technique has been used previously by Butcher and Clarke [7], ...

... Moffat and Zobel [14] suggest a method for determining this parameter. We use a slightly different formulation in this paper...

Citation context words

Citation context for the target paper [3]

Citation position/link position

Link semantics

Cited/Target paper


[14] Self-Indexing Inverted Files for Fast Text Retrieval

[9] Static Index Pruning for Information Retrieval Systems

Efficient Document Retrieval in Main Memory

Topics distribution:

55% 20% 13% 12%

Citation Relationship Type

Basic
Comparable work
Other

Influence Strength

Weak
Middle
Strong

Topic modeling over links

Topics

Topic 31: Ranking and Inverted Index

Topic 27: Information retrieval

Topic 23: Index method

Other
Pairwise Restricted Boltzmann Machines (PRBMs)

Example

- **Introduction**
  - Our method increases query throughput by 15% over the method proposed by Ash and Moffett [1], achieving similar results.

- **Related Work**
  - Impact-indexed indices are described in a series of papers by Ash and Moffett [1, 2, 3].
  - Page et al. proposed a class of algorithms known as tandem algorithms [4].

- **Approach**
  - A similar technique has been used previously by Shreftcher and Chiriz [5].

- **Methods**
  - Moffitt and Zobel [6] suggest a method for determining this parameter. We use a slightly different formulation in our paper.

- **Link category**
  - Defined over the link to bridge the two pages.

- **Latent variables**
  - Defined over the link to bridge the two pages.

- **PRBM model**

- **Citation Relationship Type**
  - Basic, comparable work, OTHER.

- **Influence Strength**
  - Weak, Middle, Strong.

- **Efficient Document Retrieval in Main Memory**
  - Topics: 15%, 25%, 15%.

- **Topic distribution**
  - Topic 1, Topic 2, Topic 3.

- **Citation Counting**
  - Source page $a$, Target page $a'$.

- **Pairwise Restricted Boltzmann Machines (PRBMs)**

### Results: Accuracy of Link Categorization

<table>
<thead>
<tr>
<th>Approach</th>
<th>Type</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>drill down</td>
<td>64.09</td>
<td>53.60</td>
<td>58.24</td>
<td>88.50</td>
</tr>
<tr>
<td>SVM+LDA</td>
<td>similar</td>
<td>77.66</td>
<td>83.34</td>
<td>80.33</td>
<td>76.59</td>
</tr>
<tr>
<td>SVM+RBM</td>
<td>other</td>
<td>64.63</td>
<td>50.58</td>
<td>61.77</td>
<td>70.71</td>
</tr>
<tr>
<td>gPRBM</td>
<td>Topics ((z^5)) of Citing Papers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Topic 37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Probabilistic model)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dPRBM</td>
<td>Topics ((z^4)) of Cited Papers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Topic 51</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Maximum entropy)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hPRBM</td>
<td>Influential Strength</td>
<td>5.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Source Paper</td>
<td></td>
<td></td>
<td>Target Paper</td>
<td>Influence</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Drill Down (Basic theory)</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>Latent dirichlet allocation</td>
<td></td>
<td></td>
<td>A variational bayesian framework for graphical models</td>
<td>2.36</td>
</tr>
<tr>
<td>SVM+LDA</td>
<td>A sentiment-aware model for predicting sales</td>
<td></td>
<td></td>
<td>Probabilistic latent semantic analysis</td>
<td>1.93</td>
</tr>
<tr>
<td>SVM+RBM</td>
<td>Mining and summarizing customer reviews</td>
<td></td>
<td></td>
<td>Foundations of statistical natural language processing</td>
<td>3.25</td>
</tr>
<tr>
<td>gPRBM</td>
<td>Similar (Comparable work)</td>
<td></td>
<td></td>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>dPRBM</td>
<td>Constraint-driven clustering</td>
<td></td>
<td></td>
<td>Max-min d-cluster formation in wireless ad hoc networks</td>
<td>2.00</td>
</tr>
<tr>
<td>hPRBM</td>
<td>Hmms and coupled hmms for multi-channel eeg classification</td>
<td></td>
<td></td>
<td>Gaussian observation hidden markov models for eeg analysis</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>Scalable collaborative filtering using cluster-based smoothing</td>
<td></td>
<td></td>
<td>Evaluating collaborative filtering recommender systems</td>
<td>1.51</td>
</tr>
<tr>
<td></td>
<td>Latent dirichlet allocation</td>
<td></td>
<td></td>
<td>Overview of the first text retrieval conference</td>
<td>3.93</td>
</tr>
<tr>
<td></td>
<td>Utility scoring of product reviews</td>
<td></td>
<td></td>
<td>Introduction to modern information retrieval</td>
<td>3.16</td>
</tr>
<tr>
<td></td>
<td>Sentiment analyzer: extracting sentiments about a given topic</td>
<td></td>
<td></td>
<td>Finding parts in very large corpora</td>
<td>2.71</td>
</tr>
</tbody>
</table>
Work5: Towards Ontology Learning from Folksonomies (IJCAI’09)
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?
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

Generative process:

1. For each topic $z$, draw a multinomial distribution $\phi_z$ from a Dirichlet prior $\beta_z$;
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 $\theta$ is generated from a Dirichlet prior $\alpha$;
   - (c) draw a word $w_{di}$ from the multinomial distribution $\phi_{z_{di}}$.

$$P(z_{di}, u_{di}| z_{-di}, u_{-di}, w, t, \alpha, \beta) = \frac{P(z, u, w, t|\alpha, \beta)}{P(z_{-di}, u_{-di}, w, t|\alpha, \beta)} \times \frac{m^{-di}_u + \alpha_z}{\sum_z (m^{-di}_u + \alpha_z)} \times \frac{n^{-di}_z + \beta_v}{\sum_v (n^{-di}_z + \beta_v)}$$
The new objective function:

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

with

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

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

**Algorithm:** Parameter estimation

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

   (a) train the model parameters ($\theta, \phi$) 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

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

   where $\theta_{t_i, k}$ is the $k$-th dimension of the multinomial of tag $t$ with $1 \leq k \leq K$, $|G_{t_i}|$ is the number of neighbors of tag $t_i$ in the tag network $G$;

   (b) fix $\phi$, and re-estimate the multinomial $\theta$ to minimize $O_\varepsilon$; we employ the algorithm proposed in [Mei et al., 2008] to optimize $O_\varepsilon$, (i.e. run an iterative process to obtain the new $\theta$ 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 $\gamma$ is a coefficient to smooth the topic distribution among the neighbors.
Divergence Estimation

• Tag divergence

\[ \text{diverg}(t_i, t_j) = D_{KL}(\theta_{t_i} \mid \mid \theta_{t_j}) = \sum_{z=1}^{K} \theta_{t_i,z} \log \frac{\theta_{t_i,z}}{\theta_{t_j,z}} \]

• Hypernym-divergence

\[ \text{hyper-diverg}(t_i, t_j) = \sum_{z=1}^{K} \frac{P(t_i \mid z_k) - P(t_j \mid z_k)}{P(t_i \mid z_k)} \]

• Merging-divergence

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

• Keep-divergence

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

Estimated topic distribution

Posterior probability derived from the topic modeling results
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
   
   (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);
   
   (b) execute the operation:
      
      i. for Subordinate, assign \( t_i \) as the hypernym of \( t_j \), remove \( t_j \) from \( A \), and move \( t_i \) as the parent node of \( t_j \) in \( O \);
      
      ii. for Merge, create a virtual concept \( c_{i,j} \) by combining \( t_i \) and \( t_j \); add it to \( A \) as well as 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 \( c_{i,j} \) with all others tags in \( A \).
      
      iii. for Keep, create a virtual concept \( c_{i,j} \) as the hypernym of both \( t_i \) and \( t_j \); add \( c_{i,j} \) as the parent node of \( t_i \) and \( t_j \) in \( O \), add \( c_{i,j} \) to \( A \), remove \( t_i \) and \( t_j \) from \( A \), and calculate the divergences of the new concept \( c_{i,j} \) with all others tags in \( A \).

4. } until \(|A| \leq 1\); //\(|A|\) is the number of elements in \( A \).
5. return the hierarchical tree \( O \).
Case Study—Movie

By clustering

By TT

By TT with smoothing
Case Study—Paper
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}   }
@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},
}

@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",
  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},
}
@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 = "Hierarchical 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},
}
@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 joint conference on Artificial intelligence (IJCAI'09)},
    year = {2009},
    pages = {2089--2094},
}
Thanks!

Q&A

HP: http://keg.cs.tsinghua.edu.cn/persons/tj/
http://arnetminer.org