iTopicModel: Information Network-Integrated Topic Modeling (ICDM 09’)

Yizhou Sun, Jiawei Han, Jing Gao and Yintao Yu

Department of Computer Science
University of Illinois at Urbana-Champaign

Presenter: Zhijin Li
networks associated with text information becoming increasingly popular
Traditional topic model vs. iTopicModel

1. papers linked by citations
2. webpages linked by hyper links
3. authors linked by co-authorship
   ....

Each document has a topic distribution

a. Independent Documents
b. Document Network
How can the connections help topic modeling?

Closely related documents should have similar text information

e.g. two researchers may share similar research interests if they co-authored a lot

Webpages…
iTopicModel

utilize both the link and text information among documents to do better topic modeling

\[ p(X, \theta | G, \beta) = p(\theta | G)p(X | \theta, \beta) = p(\theta | G) \prod_{i=1}^{N} p(x_i | \theta_i, \beta) \]
Definitions of Document Networks

Definition 1. **Document.** A document \( x_i \) in a document collection \( X = \{x_1, x_2, \ldots, x_N\} \) is comprised of a bag of words from a vocabulary \( Y = \{y_1, y_2, \ldots, y_M\} \), and is represented with vector \( x_i = (c_{i1}, c_{i2}, \ldots, c_{iM}) \), where \( c_{il} \) denotes the occurrence number of word \( y_l \) in document \( x_i \).

Definition 2. **Document Network.** A document network \( G = \langle X, E, W \rangle \) is a graph defined on a document set \( X \). \( E \) is the link set, and \( e = \langle x_i, x_j \rangle \in E \) if there is a link from document \( x_i \) to \( x_j \). \( W \) is the adjacency matrix denoting the weights of the links, \( w_{ij} > 0 \) if there is a link from node \( x_i \) to \( x_j \), and the value of \( w_{ij} \) is the strength of the link \( e = \langle x_i, x_i \rangle; w_{ii} = 0 \), otherwise.

e.g. In a **Paer Citation Network**, \( X \) is a collection of papers, \( Y \) is a vocabulary, if \( x_i \) cites \( x_j \), a link \( e=<x_i, x_j> \) with the weight \( w_{ij}=1 \) is then added to \( E \).
Markov Random Fields (MRF)

Definition 4. Markov Random Field. Given a graph $G = \langle V, E \rangle$, where $V = \{1, \ldots, n\}$, with each number as the label for each node. Let $F = \{F_i\}_{i=1}^{n}$ be a family of random variables defined on the node set $V$, i.e., each node $i$ is associated with a random variable $F_i$. $F$ is said to be a Markov Random Field on $V$ with respect to graph $G$ if and only if the following two conditions are satisfied:

$$P(f) > 0, \forall f \in F$$

(1)

$$P(f_i|f_{-i}) = P(f_i|f_{N(i)})$$

(2)

According to Hammersley-Clifford theorem, an MRF $P(f)$ can be factorized as

$$p(\theta) = \frac{1}{Z} \exp\left\{-\sum_{c \in C} V_c(\theta)\right\}$$
Structure Modeling

Use MRF to model structure dependencies

Probability of the configuration for a document:

\[
p(\theta_i|\theta_{N_{out}(i)}) \sim \text{Dirichlet}(\alpha_i) = \frac{1}{B(\alpha_i)} \prod_{k=1}^{T} \theta_{ik}^{\alpha_{ik}-1}
\]

where \( \alpha_i = \alpha_{i|N_{out}(i)} = \alpha_i^0 + \sum_{j \in N_{out}(i)} w_{ij} \theta_j \)

To prove Theta is a MRF, we need to prove:

\[
p(\theta) = \frac{1}{Z} \exp\{- \sum_{c \in C} V_c(\theta)\}
\]

\[
V_i(\theta_i) = -(\alpha_i^0 - \bar{1})^T \log(\theta_i)
\]

\[
V_{i \rightarrow j}(\theta_i, \theta_j) = \begin{cases} 
-(w_{ij} \theta_j)^T \log(\theta_i), & \text{if } (x_i, x_j) \in E; \\
0, & \text{otherwise.}
\end{cases}
\]

\[
p(\theta|G) = \frac{1}{Z} \exp\{\sum_{i} [(\alpha_i^0 + \sum_{j \in N(i)} w_{ij} \theta_j - \bar{1})^T \log(\theta_i)]\}
\]
Text Modeling

Use traditional topic modeling technique to model the generation of text for each document.

Each document is a mixture model over topics:

$$p(y_i|x_i) = \sum_{k=1}^{T} p(z = k|x_i) p(y_i|z = k) = \sum_{k=1}^{T} \theta_{ik}\beta_{kl}$$

Probability to generate a document:

$$p(x_i|\theta,\beta) = \prod_{l=1}^{M} p(y_l|x_i, \theta, \beta)^{c_{il}} = \prod_{l=1}^{M} \left[ \sum_{k=1}^{T} \theta_{ik}\beta_{kj} \right]^{c_{il}}$$

$$p(X, \theta|G, \beta) = p(\theta|G)p(X|\theta, \beta) = p(\theta|G) \prod_{i=1}^{N} p(x_i|\theta_i, \beta)$$
Parameter Estimation

Log likelihood objective function:

\[ \log L = \log p(X, \theta | G, \beta) \]

\[ = \sum_{i=1}^{N} \sum_{k=1}^{T} \left( (\alpha_{i,k}^{0} - 1) \log \theta_{ik} + \sum_{j=1}^{N} w_{ij} \theta_{jk} \log \theta_{ik} \right) + \sum_{i=1}^{N} \sum_{l=1}^{M} c_{il} \log \left( \sum_{k=1}^{T} \beta_{kl} \theta_{ik} \right) - \log Z \]

E-step:

\[ Q(\Psi | \Psi^{(t)}) = E_{z | X, \Psi^{(t)}} (\log L) \]

\[ = \sum_{i=1}^{N} \sum_{k=1}^{T} \left( (\alpha_{i,k}^{0} - 1) \log \theta_{ik} + \sum_{j=1}^{N} w_{ij} \theta_{jk}^{(t)} \log \theta_{ik} \right) \]

\[ + \sum_{i=1}^{N} \sum_{l=1}^{M} c_{il} \sum_{k=1}^{T} p(z = k | x_i, y_l, \Psi^{(t)}) \log(\beta_{kl} \theta_{ik}) \]

M-step:

\[ \theta_{ik}^{(t+1)} = \frac{\alpha_{i,k}^{0} - 1 + \sum_{j=1}^{N} w_{ij} \theta_{jk}^{(t)}}{\sum_{k=1}^{T} \alpha_{i,k}^{0} + \sum_{j=1}^{N} w_{ij} + \sum_{l=1}^{M} c_{il}} \]

\[ \beta_{kl}^{(t+1)} = \frac{\sum_{i=1}^{N} c_{il} p(z = k | x_i, y_l, \Psi^{t})}{\sum_{l' = 1}^{M} \sum_{i=1}^{N} c_{il'} p(z = k | x_i, y'_l, \Psi^{(t)})} \]

\[ \Psi = (\theta_{\{N \times T\}}, \beta_{\{T \times M\}}) \]
Experiments – Datasets

1. DBLP dataset

   (1) “all-area” dataset: top 1000 conferences and top 50000 authors, and all the publications of these authors.

   (2) “four-area” dataset: 28k authors and their publications in the conferences about DB, DM, ML and IR.

2. Cora Research Paper Classification data set

   19k papers with their citation lists, author lists and title information. Each paper has a classification label from total 70 classes.
Experiments – Document Networks Formation

For conferences:

All the titles of papers in KDD

All the titles of papers in CIKM

Weight = # authors shared

For authors:

Author 1

Author 2

Weight = # papers co-authored

For papers:

Title of paper 1

Title of paper 2

Weight = # authors shared
### Experiments – Topics Found

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 6</th>
<th>Topic 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>network</td>
<td>network</td>
<td>parallel</td>
<td>system</td>
<td>system</td>
<td>graph</td>
<td>image</td>
</tr>
<tr>
<td>system</td>
<td>model</td>
<td>design</td>
<td>data</td>
<td>software</td>
<td>problem</td>
<td>recognition</td>
</tr>
<tr>
<td>performance</td>
<td>algorithm</td>
<td>test</td>
<td>model</td>
<td>design</td>
<td>algorithm</td>
<td>base</td>
</tr>
<tr>
<td>distribute</td>
<td>neural</td>
<td>simulation</td>
<td>database</td>
<td>web</td>
<td>set</td>
<td>model</td>
</tr>
<tr>
<td>wireless</td>
<td>fuzzy</td>
<td>high</td>
<td>language</td>
<td>information</td>
<td>bound</td>
<td>3d</td>
</tr>
<tr>
<td>protocol</td>
<td>method</td>
<td>architecture</td>
<td>base</td>
<td>computer</td>
<td>tree</td>
<td>video</td>
</tr>
<tr>
<td>control</td>
<td>analysis</td>
<td>power</td>
<td>program</td>
<td>model</td>
<td>complexity</td>
<td>detection</td>
</tr>
<tr>
<td>scheme</td>
<td>learn</td>
<td>circuit</td>
<td>approach</td>
<td>environment</td>
<td>linear</td>
<td>robot</td>
</tr>
<tr>
<td>channel</td>
<td>problem</td>
<td>memory</td>
<td>query</td>
<td>management</td>
<td>order</td>
<td>motion</td>
</tr>
<tr>
<td>mobile</td>
<td>genetic</td>
<td>fault</td>
<td>knowledge</td>
<td>development</td>
<td>function</td>
<td>segmentation</td>
</tr>
</tbody>
</table>

#### Table II

**First Level Topics in DBLP (Best T = 7, top-10 words for each topic)**

#### Table III

**Sub Topics of Topic 4 in Level 1 (Best T = 6, top-10 words for each topic)**
Experiments – Settings

1. Besides finding topics among documents, apply iTopicModel to do clustering.

2. Use three different networks, conference net and author net from DBLP, and paper citation net from Cora.

3. Label the 20 conferences and 200 authors sampled from the 1000 authors to the four areas for conference net and author net, and use the classification labels for papers from Cora.
Experiments – Clustering Accuracy

Compare iTopicModel with PLSA, LDA and NetPLSA:

<table>
<thead>
<tr>
<th></th>
<th>PLSA</th>
<th>LDA</th>
<th>NetPLSA</th>
<th>iTopicModel</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConfNet</td>
<td>0.7959</td>
<td>0.7469</td>
<td>0.7291</td>
<td>0.8255</td>
</tr>
<tr>
<td>AuthorNet</td>
<td>0.4059</td>
<td><strong>0.5639</strong></td>
<td>0.4761</td>
<td>0.5360</td>
</tr>
<tr>
<td>PaperNet-Cite</td>
<td>0.1287</td>
<td>0.0674</td>
<td>0.4291</td>
<td><strong>0.4424</strong></td>
</tr>
</tbody>
</table>

Table IV

DOCUMENT CLUSTERING ACCURACY: NMI

<table>
<thead>
<tr>
<th></th>
<th>PLSA</th>
<th>LDA</th>
<th>NetPLSA</th>
<th>iTopicModel</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConfNet</td>
<td>0.4364</td>
<td>0.4370</td>
<td>0.4368</td>
<td><strong>0.4440</strong></td>
</tr>
<tr>
<td>AuthorNet</td>
<td>0.4059</td>
<td>0.4910</td>
<td>0.4760</td>
<td><strong>0.4938</strong></td>
</tr>
<tr>
<td>PaperNet-Cite</td>
<td>0.1703</td>
<td>0.0984</td>
<td>0.4760</td>
<td><strong>0.5783</strong></td>
</tr>
</tbody>
</table>

Table V

CLUSTERING CONSISTENCY: Q-FUNCTION
Thank you

Any question?