Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora

Jianwen Zhang†, Yangqiu Song‡, Changshui Zhang†, Shixia Liu‡

†Tsinghua University, Beijing, China, ‡IBM Research - China
Outline

• Motivation
• Preliminaries
• Evolutionary HDP
• Experiments
  – Synthetic data
  – Finance related online document collections
• Conclusions
Motivation

• Mining cluster evolution patterns in multiple correlated time-varying corpora

News:
- World braces for insurer AIG's crash
- Congress wrestles with Wall Street bailout package
- Obama, McCain debate economic, foreign policy…

Blogs:
- Financial crash: A system in chaos
- Preparing for a Financial Crisis
- Bailout people, not banks

Message boards:
- Do I have to tell my landlord I lost my job?
- Is Obama a US citizen
- Canceling my shopping

An example of online textual data from three corpora: news, blogs, and message boards.
Motivation (Cont’)

• Patterns to discovery
  – Clusters within each corpus at each epoch
  – Shared clusters among different corpora
  – Evolving of clusters within a corpus and across corpora overtime

• Challenges
  – Single integrated model
  – Commonality & diversity
  – Time dependencies
  – Cluster numbers

• Previous works
  – Multiple corpora
  – Time-varying corpus, evolutionary clustering
Preliminaries

- Dirichlet process mixture models (DPM)
  - DPM: (prior) infinite mixture model which can automatically determine the component number by placing a Dirichlet process (DP) prior for a mixture model

A text corpus, e.g., news

\[ \begin{align*}
G & \sim \text{DP}(\alpha_0, G_0) \\
\theta_i & \in \{\text{politics, market, companies, ...}\} \\
\theta_i & \sim G, x_i \sim F(x|\theta_i)
\end{align*} \]
Preliminaries (Cont’)

• HDP mixture model: to model multiple corpora to enable them sharing components
  – Multiple DPMs sharing a same DP prior

HDP mixture model, Teh et al. JASA’06
Evolutionary HDP

- Modeling multiple time varying corpora
Evolutionary HDP (Cont’)

1. Draw an overall measure
   \[ G \sim DP(\gamma, H) \]

2. For each epoch \( t \)
   ① Draw the snapshot global measure
   \[ G_0^t \sim DP(\gamma^t, w^t G_0^{t-1} + (1 - w^t)G') \]
   ② Draw the snapshot local measures
   \[ G_j^t \sim DP(\alpha_0^t, v_j^t G_j^{t-1} + (1 - v_j^t)G_0^t) \]
   ③ Draw observations
   \[ \theta_{ji}^t \overset{i.i.d.}{\sim} G_j^t, \quad x_{ji}^t \sim F(x|\theta_{ji}^t) \]

The time dependency model
The time dependency model

\( G \): plays as a bookkeeper of all the components

(\text{the common taste})

\[ G_0^t \sim DP \left( \gamma^t, w^t G_0^{t-1} + (1 - w^t) G \right) \]

A part of the atoms of \( G_0^t \) are drawn from the previous one \( G_0^{t-1} \) while others are drawn from \( G \)

Some are inherited from the previous, and some are newly from the common taste.

\[ G_j^t \sim DP \left( \alpha^t_j, \nu^t_j G_j^{t-1} + (1 - \nu^t_j) G_0^t \right) \]

Similarly…
More…

- Different perspectives to the model (necessary to lead to the sampling scheme)
- Gibbs sampling to infer the model

(Detailing and boring. If you are interested in, we’re appreciated if you would like to read the paper instead)
Experiments

• Synthetic data
• Real financial related web text collections
Experiments on synthetic data

\[ p_j^t(x) = \sum_{\tau=1}^{3} \frac{1}{3} \text{Multinomial} \left( x; \phi_{k_j^t} \right) \]

### Table 1: Synthetic data set.

<table>
<thead>
<tr>
<th>Global components (dishes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k )</td>
</tr>
<tr>
<td>( \phi_{k,1} )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Local components (tables) and corpora sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k_j^{t_1}, k_j^{t_2}, k_j^{t_3} )</td>
</tr>
<tr>
<td>( j = 1 )</td>
</tr>
<tr>
<td>( t = 1 )</td>
</tr>
<tr>
<td>( t = 2 )</td>
</tr>
<tr>
<td>( t = 3 )</td>
</tr>
<tr>
<td>( t = 4 )</td>
</tr>
</tbody>
</table>

### Three corpora, four time epochs

\[ j = 1 \]
\[ j = 2 \]
\[ j = 3 \]
Evaluation criteria

• Static criteria
  – NMI
  – $\log(\text{perword-perplexity})$ \textit{LogPerp}

$$-\frac{1}{n_{test}} \sum_{t,j,i} \log p\left(x_{ji,test}^t | \text{Model}, X_{train}\right)$$

• Temporal criteria
  – Temporal correlations / divergences overtime

• Compared to HDP without considering time dependencies
Better predictability

Better clustering performance

Stronger correlation overtime

Figure 6: Results on the synthetic data set: static performances, averaged on 10-fold cross validation.

Figure 7: Results on the synthetic data set: temporal correlations, averaged on 10-fold cross validation.

Figure 8: Results on the synthetic data set: temporal divergences, averaged on 10-fold cross validation.
Experiments on Real Data

• 103,986 text articles queried from a search engine Boardreader.

• Financial related. Queries: 20 financial companies’ names, e.g., “AIG insurance”, “Bank of America”, etc.

• Three types. News, blogs, message boards.


• Dictionary size \( W = 77,999 \)
Better predict ability

Stronger correlation overtime
Visualization of clusters utilizing the time-based topic visualization tool TIARA (Liu et al. CIKM’09)
News

#Documents

Bank Related
Financial Crisis
Barclays Premier League
Trade Market Info

#Documents

Conclusions

• An EvoHDP model to mine cluster evolution patterns from multiple correlated time-varying corpora
• Extension of the original HDP
• Gibbs sampling
• Better predicting ability and stronger correlations across corpora overtime
• Cluster evolution patterns in real financial related web data
Thank You!!!