Linked Latent Dirichlet Allocation in Web Spam Filtering

István Bíró¹  Dávid Siklósi  Jácint Szabó¹
András A. Benczúr¹

¹Data Mining and Web Search Group
Computer and Automation Institute
Hungarian Academy of Sciences

AIRWeb Workshop, April 21, 2009, Madrid, Spain.
Latent Dirichlet Allocation

- Blei, Ng, Jordan, 2003
- fully generative statistical natural language model
- extension of latent semantic indexing (LSI)
- has better perplexity than LSI
- a document is represented as a bag-of-words (no bigrams/trigrams are taken into account)
- a lot of extensions and variations of LDA were developed and successfully applied
Latent Dirichlet Allocation

Model

- topic: distribution over the words
- document: distribution over the topics
- for every word-position of the corpus, draw a topic for that document, and then draw a word for that topic

\[
\begin{align*}
\alpha &\rightarrow \varphi_m \\
\beta &\rightarrow \varphi_k \\
\varphi_k &\leftarrow \varphi_k \\
z_{m,n} &\leftarrow \varphi_k \\
w_{m,n} &\leftarrow z_{m,n} \\
\end{align*}
\]
In practice

- given a collection of documents
- keep only semantic words, delete stopwords, stem
- create vocabulary
- choose an appropriate topic-number (about 100)
- make model inference to create the model
- for a topic, the word distribution gives a semantic theme
- for a document, the topic distribution describes to which themes it belongs
In practice

- given a collection of documents
- keep only semantic words, delete stopwords, stem
- create vocabulary
- choose an appropriate topic-number (about 100)
- make model inference to create the model
- for a topic, the word distribution gives a semantic theme
- for a document, the topic distribution describes to which themes it belongs
Related link based models

- copycat and the citation influence models (Dietz, Bickel, Scheffer 2007)
- link-PLSA-LDA and pairwise-link-LDA (Nallapati, Ahmed, Xing, Cohen 2008)

They extend LDA over a bipartition of the corpus into citing and cited documents such that influence flows along links from cited to citing documents.

- Linked LDA is similar to the citation influence model.
- The main difference: in linked LDA there is no need for a citing and a cited copy of each document.
- In linked LDA influence may flow along paths of length more than one.
Linked LDA

- extended LDA model to exploit links between documents
- beside LDA’s words and topic distribution it involves an additional distribution over the outneighbors
The smoothing parameter vector $\gamma_d$:

$$\gamma_d(c) \propto \text{the multiplicity of the } d \rightarrow c \text{ link}$$

$$\sum \gamma_d(c) = \text{document length}/p$$

- $p$ is a normalization parameter
Experiments

Linked LDA on UK2007-WEbspam, apparently primarily content spammed

- \( \sim 115000 \) sites (\( \sim 6000 \) labeled)
- \( \sim 4000 \) for train set
- \( \sim 2000 \) for test set

- document: concatenation of all pages of a site
- weight directed links by their multiplicity (max weight: \( \sim 10 \))
- use the topic distribution of a site as features
- C4.5 on the public content and link features
- SVM on tf.idf
- BayesNet on linked LDA features
- combination by log-odds averaging (Lynam and Cormack)
LDA parameters

- $k$ - number of topics
- $p$ - normalization parameter
- The Dirichlet parameter vector $\beta$ is constant $\frac{200}{|V|}$, and $\alpha$ is constant $\frac{50}{k}$

<table>
<thead>
<tr>
<th>$k$</th>
<th>$p = 1$</th>
<th>$p = 4$</th>
<th>$p = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>0.768</td>
<td>0.784</td>
<td>0.783</td>
</tr>
<tr>
<td>90</td>
<td>0.764</td>
<td>0.777</td>
<td>0.773</td>
</tr>
</tbody>
</table>

*Table:* Classification accuracy for linked LDA with various parameters, classified by BayesNet.
Baseline methods

<table>
<thead>
<tr>
<th>features</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linked LDA with BayesNet</td>
<td>0.784</td>
</tr>
<tr>
<td>LDA with BayesNet</td>
<td>0.766</td>
</tr>
<tr>
<td>tf.idf with SVM</td>
<td>0.795</td>
</tr>
<tr>
<td>public (link) with C4.5</td>
<td>0.724</td>
</tr>
<tr>
<td>public (content) with C4.5</td>
<td>0.782</td>
</tr>
</tbody>
</table>

Table: Classification accuracy for the baseline methods.
Combination

<table>
<thead>
<tr>
<th>features</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>tf.idf &amp; LDA</td>
<td>0.827</td>
</tr>
<tr>
<td>tf.idf &amp; linked LDA</td>
<td>0.831</td>
</tr>
<tr>
<td>public &amp; LDA</td>
<td>0.820</td>
</tr>
<tr>
<td>public &amp; linked LDA</td>
<td>0.829</td>
</tr>
<tr>
<td>public &amp; tf.idf</td>
<td>0.827</td>
</tr>
<tr>
<td>public &amp; tf.idf &amp; LDA</td>
<td>0.845</td>
</tr>
<tr>
<td>public &amp; tf.idf &amp; linked LDA</td>
<td>0.854</td>
</tr>
<tr>
<td>public &amp; tf.idf &amp; LDA &amp; linked LDA</td>
<td>0.854</td>
</tr>
</tbody>
</table>

Table: Classification accuracy by combining the classifications with a log-odds based random forest. For linked LDA the parameters are chosen to be $p = 4, k = 30$. 
Linked LDA slightly outperforms LDA.

Combining tf.idf, the public and the linked LDA features with a log-odds based random forest we achieved an AUC of 0.854, beating the Web Spam Challenge 2008 winner (0.848).

Measuring the inferred linked LDA edge weights by using them in a stacked graphical classification.
Questions?

jacint@ilab.sztaki.hu, ibiro@ilab.sztaki.hu, sdavid@ilab.sztaki.hu