Linked Latent Dirichlet Allocation in Web Spam Filtering

István Bíró[1] Dávid Siklósi Jácint Szabó[1]
András A. Benczúr[1]

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

AIRWeb Workshop, April 21, 2009, Madrid, Spain.

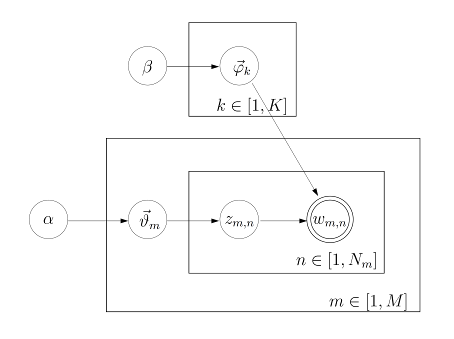
Latent Dirichlet Allocation

- 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



Latent Dirichlet Allocation

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

Latent Dirichlet Allocation

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

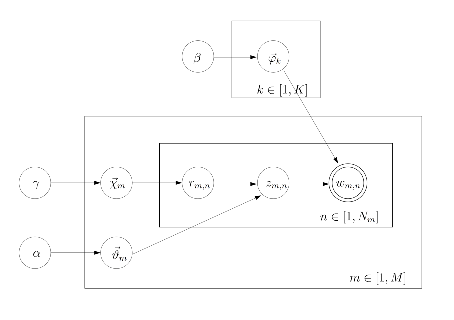
- copycat and the citation influence models (Dietz, Bickel, Scheffer 2007)
- link-PLSA-LDA and pairwise-link-LDA (Nallapati, Ahmed, Xing, Cohen 2008)

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



Linked LDA

The smoothing parameter vector γ_d:

$$\gamma_d(c) \propto \text{ the multiplicity of the } d \to c \text{ link}$$
$$\sum \gamma_d(c) = \text{ document length}/p$$

- 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 β is constant $200/|V|$, and α is constant $50/k$

	$p = 1$	$p = 4$	$p = 10$
$k = 30$	0.768	**0.784**	0.783
$k = 90$	0.764	0.777	0.773

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

Baseline methods

features	AUC
Linked LDA with BayesNet	0.784
LDA with BayesNet	0.766
tf.idf with SVM	0.795
public (link) with C4.5	0.724
public (content) with C4.5	0.782

Table: Classification accuracy for the baseline methods.

Combination

features	AUC
tf.idf & LDA	0.827
tf.idf & linked LDA	0.831
public & LDA	0.820
public & linked LDA	0.829
public & tf.idf	0.827
public & tf.idf & LDA	0.845
public & tf.idf & linked LDA	**0.854**
public & tf.idf & LDA & linked LDA	**0.854**

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

Conclusion and Future Work

- 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