The Topic–Perspective Model for Social Tagging Systems

Caimei Lu et al. (KDD 2010)

Presented by Anson Liang
Movitation

Users:

Documents:
articles, web pages, etc.

Add

Annotate

Tags:
e.g. economy, science, facebook, twitter ...
Movitation

E.g.

- Delicious.com
  - Social bookmarking
Movitation
Movitation

Different users have different perspectives!
Movitation

- Objectives
  - Simulate the generation process of social annotations
  - Automatically generate tags for documents based on user perspective
Related Work

- Topic Analysis using Generative Models
  - Latent Dirichlet Allocation (LDA) model
    - Discover topics for a given document

- A document has many words
- A document may belong to many topics
- A word may belong to many topics
- A topic may be associated with many words
Related Work

LDA – Plate notation

- $\alpha$ is the parameter of the uniform Dirichlet prior on the per-document topic distributions.
- $\beta$ is the parameter of the uniform Dirichlet prior on the per-topic word distribution.
- $\theta_i$ is the topic distribution for document $i$.
- $z_{ij}$ is the topic for the $j$th word in document $i$, and
- $w_{ij}$ is the specific word.
- The $w_{ij}$ are the only observable variables, and the other variables are latent variables.
- $\phi$ is a $K \times V$ (V is the dimension of the vocabulary) Markov matrix each row of which denotes the word distribution of a topic.

Related Work

- Conditionally-independent LDA (Ramage et al.)
  - Tags are generated from the topics (the same source as the words)

(a) CI-LDA
Related Work

- Correlated LDA
  (Bundschus et al.)
  1. Generates topics associated with the words for a document
  2. Generates tags from the topics associated with the words
Related Work

Problem

- Assume words and tags are generated from the same source

- However, words are generated by the authors, and tags are generated by users (readers)

- User information is missed in the tag generation process!
Related Work

- Three classification schemas of social tags

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>What or who it is about</td>
<td>Topic</td>
<td></td>
</tr>
<tr>
<td>Refining categories</td>
<td>Time</td>
<td>Factual</td>
</tr>
<tr>
<td>What it is</td>
<td>Type</td>
<td></td>
</tr>
<tr>
<td>Who owns it</td>
<td>Author/Owner</td>
<td></td>
</tr>
<tr>
<td>Qualities and characteristics</td>
<td>Opinions/Qualities</td>
<td>Subjective</td>
</tr>
<tr>
<td>Task organization</td>
<td>Usage context</td>
<td>Personal</td>
</tr>
<tr>
<td>Self reference</td>
<td>Self reference</td>
<td></td>
</tr>
</tbody>
</table>
Overview

Two factors act on the tag generation process:
1. The topics of the document
2. The perspective adopted by the user
**Topic-Perspective Model**

- **Model Formulation**
  - $X=0$: Tags are generated from user perspectives
  - $X=1$: Tags are generated from the topics learned from the words

**Standard LDA**
- User-perspective distribution
- Document-topic distribution
- Probability that each tag is generated from topics
- Perspective-tag distribution
- Topic-word distribution
- Switch 0/1
- Topic-tag distribution
Topic–Perspective Model

1) For each of the $D$ documents $d$, sample $\theta^{(d)}_d \sim \text{Dirichlet}(\alpha_d)$;
2) For each of the $U$ users $u$, sample $\theta^{(u)}_u \sim \text{Dirichlet}(\alpha_u)$;
3) For each of the $K$ topics $k$, sample $\phi^{(w)}_k \sim \text{Dirichlet}(\beta_w)$, and sample $\phi^{(t)}_k \sim \text{Dirichlet}(\beta_t)$;
4) For each of the $L$ user perspectives $l$, sample $\psi_l \sim \text{Dirichlet}(\eta)$;
5) For each of the $N_d$ word tokens $w_i$ in document $d$:
   a) sample a topic $z_i \sim \text{Multinomial}(\theta^{(d)}_d)$;
   b) sample a word $w_i \sim \text{Multinomial}(\phi^{(w)}_{z_i})$;
6) For each of the T tags $t$ in the collection $D$, sample $\lambda_t \sim \text{Beta}(\gamma)$;
7) For each of the $M_d$ tag tokens $t_j$ in document $d$ created by user $u$;
   a) sample a flag $X \sim \text{Binomial}(\lambda_{t_j})$;
   b) if ($X = 1$):
      i) Sample a topic $z_{t_j} \sim \text{Uniform}(z_{w_1}, \ldots, z_{w_{N_d}})$;
      ii) Sample a tag $t_j \sim \text{Multinomial}(\phi^{(b)}_{z_j})$;
   c) if ($X = 0$):
      i) Sample a perspective $p_j \sim \text{Multinomial}(\theta_u)$;
      ii) Sample a tag $t_j \sim \text{Multinomial}(\psi_{p_j})$;
Topic–Perspective Model

- Parameter Estimation
  1. document–topic distribution $\theta(d)$
  2. topic–word distribution $\phi(w)$
  3. topic–tag distribution $\phi(t)$
  4. user–perspective distribution $\theta(u)$
  5. perspective–tag distribution $\psi$
  6. binomial distribution $\lambda$

- Gibbs Sampling!
Topic–Perspective Model

- Parameter Estimation

Sampling equation of the word topic variables for each content word \( w_i \). (The same as standard I.DA model):

\[
p(z_i = k \mid w_i = v, z_{-i}, w_{-i}, \alpha_d, \beta_w) \propto \frac{C_{kd,-i}^{KD} + \alpha_d}{\sum_{k'} C_{k'd,-i}^{KD} + K\alpha_d} \cdot \frac{C_{vk,-i}^{WK} + \beta_w}{\sum_{v'} C_{v'k,-i}^{WK} + V\beta_w}
\] (1)
Topic–Perspective Model

- **Parameter Estimation**

  - Sampling equation of the tag topic variables when the switch variable $X = 1$:

  $$
p(x_j = 1, z_j^{(t)} = \tilde{z} \mid t_j = q, z_{-j}, t_{-j}, \beta_w, \beta_t, \gamma) \propto
  \frac{\tilde{n}_{q,-j} + \gamma}{n_q + \tilde{n}_{q,-j} + 2\gamma} \cdot \frac{C_{\tilde{z}d}^{KD}}{N_{wd}} \cdot \frac{C_{q\tilde{z},-j}^{TK} + \beta_t}{\sum_{q'} C_{q'\tilde{z},-j}^{TK} + T\beta_t}
  $$

  (2)
Topic-Perspective Model

- **Parameter Estimation**

  - Sampling equation of the tag perspective variables when the switch variable $X = 0$:

    $$
    p(x_j = 0, p_j = l | t_j = q, p_{-j}, t_{-j}, \alpha_u, \beta_t, \gamma) \propto \frac{n_{q,-j} + \gamma}{n_q + \tilde{n}_{q,-j} + 2\gamma} \cdot \frac{C_{lu,-j}^{LU} + \alpha_u}{\sum_{l'} C_{lu,-j}^{LU} + L\alpha_u} \cdot \frac{C_{ql,-j}^{TL} + \beta_t}{\sum_{q'} C_{ql,-j}^{TL} + T\beta_t}
    $$

    (3)
Topic–Perspective Model

- Parameter Estimation

\[ \theta_{kd}^{(d)} = \frac{C_{kd,-i}^{K} + \alpha_d}{\sum_{k'} C_{k'd,-i}^{K} + K \alpha_d} \]

\[ \phi_{vk}^{(w)} = \frac{C_{vk}^{W} + \beta_w}{\sum_{v'} C_{v'k}^{W} + V \beta_w} \]

\[ \phi_{qz}^{(t)} = \frac{C_{qk,-j}^{T} + \beta_t}{\sum_{q'} C_{q'k,-j}^{T} + T \beta_t} \]

\[ \phi_{lu}^{(m)} = \frac{C_{lu,-j}^{L} + \alpha_u}{\sum_{l'} C_{l'u,-j}^{L} + L \alpha_u} \]

\[ \psi_{ql} = \frac{C_{ql,-j}^{T} + \beta_t}{\sum_{q'} C_{q'l,-j}^{T} + T \beta_t} \]

\[ \lambda_q = \frac{\tilde{n}_{q,-j} + \gamma}{n_q + \tilde{n}_{q,-j} + 2\gamma} \]
Datasets & Setup
- A social tagging dataset from delicious.com after refining
  - 41190 documents
  - 4414 users
  - 28740 unique tags
  - 129908 unique words
- 90% for training, 10% for testing
Experiments & Results

- Evaluation Criterion
  - Perplexity
    - Reflects the ability of the model to predict tags for new unseen documents
    - A lower perplexity score $\rightarrow$ better generalization performance

\[
\text{perplexity} \ (D_{\text{test}}) = \exp\left\{ \frac{\sum_{d=1}^{D_{\text{test}}} \log(p(t_d))}{\sum_{d=1}^{D_{\text{test}}} M_d} \right\}.
\]

\[
p(t_d) = p(x_{t,d} = 1) \sum_{k=1}^{K} p(t_d \mid z_k) p_{\text{test}}(z_k \mid d) + p(x_{t,d} = 0) \sum_{i=1}^{L} p(t_d \mid p_i) p_{\text{test}}(p_i \mid u)
\]
Experiments & Results

- Number of topics vs. number of perspectives

Figure 3. The perplexities over the iterations for five settings of topic number when perspective number $L=80$

Figure 4. The perplexities over the iterations for five settings of perspective number when topic number $K=80$
Experiments & Results

- Tag Perplexity

Figure 5. The perplexity results of CorrLDA, CI-LDA and TP-LDA (Topic-Perspective Model) for topic number $K=10$, 20, 40, 80, 160.
Experiments & Results

- Discovered Topics
  - Observed high semantic correlations among the top words and tags for each topic

<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Top words</th>
<th>Top tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>war world militaries nation state force govern unite Iraq countries international israel american armies peace</td>
<td>politics history world international war military activism poverty information africa islam government middle east europe human right</td>
</tr>
<tr>
<td>13</td>
<td>mountain fish camp boat adventure sea river park trail climb ski new lake gear sail</td>
<td>travel camp backpack hike photography climb sail photo knot boat gear nature adventure ski kayak</td>
</tr>
<tr>
<td>15</td>
<td>movie film star video dvd man episode new release trailer love review girl fan season</td>
<td>movy video entertainment film music review movie humor television medium fun funny cinema stream comic strip</td>
</tr>
</tbody>
</table>
Experiments & Results

- Discovered Perspectives
  - Perspectives are more complicated than topics
  - The correlation among the tags assigned to each perspective is not as obvious as those for topics

<table>
<thead>
<tr>
<th>Perspective ID</th>
<th>Top tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>reference guide multimedia list help codec portal emulator comparison boot upload anonymous virtual organize proxy</td>
</tr>
<tr>
<td>24</td>
<td>competition switch event blogroll likeddesign inspire wysiwyg creative artistresource ria tagthese domainname affiliate editorial cooky</td>
</tr>
<tr>
<td>32</td>
<td>conference metadata openacce tag association folksonomy censorship preservation digitallibrary sheetmusic librarian secondlife rfid directory digitalgame</td>
</tr>
<tr>
<td>36</td>
<td>search link portal directory list system indie tag rock about customize current label usenet ezine kaizen synchronization</td>
</tr>
</tbody>
</table>
Experiments & Results

The generation sources of tags

Tags are generated from the topics learned from the words

<table>
<thead>
<tr>
<th>( \lambda = 1 )</th>
<th>( \lambda = 0.5 )</th>
<th>( \lambda = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>library shop</td>
<td>palmpre audiomagazine</td>
<td>tag app interest</td>
</tr>
<tr>
<td>internet research</td>
<td>mathswarem postapocalyptic</td>
<td>archive toread</td>
</tr>
<tr>
<td>social network</td>
<td>educause vomit nwiqpartn</td>
<td>datum code todo</td>
</tr>
<tr>
<td>statistic ruby ajax</td>
<td>singlespe masterproef</td>
<td>webservice</td>
</tr>
<tr>
<td>javascript webdev</td>
<td>richmullin sundial selenium</td>
<td>directory list guide</td>
</tr>
<tr>
<td>culture music</td>
<td>showstep webmath</td>
<td>link portal training</td>
</tr>
<tr>
<td>health graphic math</td>
<td>randynewman immortalism</td>
<td>site track article</td>
</tr>
<tr>
<td>security firefox cs</td>
<td>malazan architecturalproduct</td>
<td>reference web20</td>
</tr>
<tr>
<td>politics recipe</td>
<td>fotologsrevista biblioteque</td>
<td>online search tool</td>
</tr>
<tr>
<td>photography</td>
<td>caribbean europeana</td>
<td>free cool</td>
</tr>
</tbody>
</table>
Future Work

- Apply the results of the model for
  1. tag recommendation
  2. Personalized information retrieval
Summary

- Topic-perspective LDA model
  - Simulate the tag generation process in a more meaningful way
  - Tags are generated from both the topics in the document and the user perspective
  - May be applied in computer vision tagging problems
Question?

- Thank you!