Semantic Visualization for Spherical Representation

Tuan M. V. Le and Hady W. Lauw

KDD 2014
Visualize Document Collections

• To present the contents/semantics/themes/etc of the documents.

• To show the similarities among documents in a collection.
Each point represents a document.
### Topic Model

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>
Topic Model Not Intended for Visualization
Semantic Visualization Problem

Given a set of documents (bag of words)

Find 2-dimensional visualization coordinates + Find K-dimensional topic distributions
Our Approach for Semantic Visualization

**Spherical**

$L^2$-normalized vector

+ Document: $v_n$
+ Topic: $\tau_z$

**Semantic**

Topic distribution

**Embedding**

Visualization coordinates

+ Document: $x_n$
+ Topic: $\phi_z$
Data Representation – Word Space

Spherical

- Richer feature representations
  - $tf$, $tf$-$idf$, …
- Model directly absences of word.
- Similarity is based on cosine distance:
  - Not sensitive to document length.

Von Mises–Fisher distribution (vMF):

$$ f_p (x; \mu, \kappa) = C_p (\kappa) \exp(\kappa \mu^T x) $$
Data Representation – Topic Space

Each document is represented as a point on the topic simplex

Semantic

Topic distribution
Topic Representation

• Each topic is represented as a point on the sphere (word space)

<table>
<thead>
<tr>
<th>Word</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word 1</td>
<td>0.7</td>
</tr>
<tr>
<td>Word 2</td>
<td>0.2</td>
</tr>
<tr>
<td>Word 3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Multinomial (Sum up to 1)

<table>
<thead>
<tr>
<th>Word</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word 1</td>
<td>0.95</td>
</tr>
<tr>
<td>Word 2</td>
<td>0.27</td>
</tr>
<tr>
<td>Word 3</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Spherical (Unit length)
Relationship between Topic Distributions and Visualization Coordinates

Shorter visualization distance means greater topic probability.

\[
P(z | d_n) = P(z | x_n, \Phi) = \frac{\exp(-\frac{1}{2}||x_n - \phi_z||^2)}{\sum_{z' = 1}^{Z} \exp(-\frac{1}{2}||x_n - \phi_{z'}||^2)}
\]

This formula is also used in PLSV, Iwata et al., KDD 2008
Spherical Semantic Embedding

The observed L2-normalized word vector of $d_n$

Visualization coordinate of document $d_n$

L2-normalized word vector of topic $z$

Visualization coordinate of topic $z$
Generative Process

1. Draw the corpus mean direction: $\mu \sim \text{vMF}(m, \kappa_0)$

2. For each topic $z = 1, \ldots, Z$:
   - Draw $z$’s coordinate: $\phi_z \sim \text{Normal}(0, \beta^{-1}I)$
   - Draw $z$’s spherical direction: $\tau_z \sim \text{vMF}(\mu, \xi)$

3. For each document $d_n$, where $n = 1, \ldots, N$:
   - Draw $d_n$’s coordinate: $x_n \sim \text{Normal}(0, \gamma^{-1}I)$
   - Derive $d_n$’s topic distribution:
     $$\theta_{n,z} = P(z|x_n, \Phi) = \frac{\exp\left(-\frac{1}{2}||x_n - \phi_z||^2\right)}{\sum_{z'=1}^{Z} \exp\left(-\frac{1}{2}||x_n - \phi_{z'}||^2\right)}$$
   - Derive $d_n$’s spherical average: $\tau_n = \frac{\sum_{z=1}^{Z} \theta_{n,z} \cdot \tau_z}{||\sum_{z=1}^{Z} \theta_{n,z} \cdot \tau_z||}$
   - Draw $d_n$’s spherical direction: $\nu_n \sim \text{vMF}(\tau_n, \kappa)$
Parameter Estimation

- Variational EM with MAP estimation.
Comparative Methods

Visualization
- t-SNE

Semantic Visualization
- Joint
  - Spherical: SSE (Ours)
  - Multinomial: PLSV

Pipeline
- Spherical: PE (SAM)
- Multinomial: PE (LDA)
20News Dataset (20 Categories)

![Graph showing accuracy vs number of topics for 20News dataset]

- SSE
- PLSV
- t-SNE
- PE (SAM)
- PE (LDA)
Visualization Comparison

SSE (Spherical)  PLSV (Multinomial)
Conclusion

• Spherical Semantic Embedding (SSE) is designed for data with spherical representation.

• Promising applications for integrated modeling:
  – semantic-rich visualizations
  – assigning categories to documents