External Evaluation of Topic Models: A Graph Mining Approach

Hau Chan*  Leman Akoglu

ICDM 2013  December 7-10, 2013
Topic Models

T (topics)

Document Corpora

Topic Model (e.g. LDA)

Topic 1: \( w_{11}, \ldots, w_{1N} \)
Topic 2: \( w_{21}, \ldots, w_{2N} \)
Topic 3: \( w_{31}, \ldots, w_{3N} \)
... 
... 
... 
Topic T: \( w_{T1}, \ldots, w_{TN} \)

Example topics T1 (high-quality) and T2 (low-quality) of a topic model.

\begin{itemize}
  \item \textcolor{green}{T1: steam, engine, valve, piston, cylinder, pressure, boiler, air, pump, pipe}
  \item \textcolor{red}{T2: cut, system, capital, pointed, opening, building, character, round, france, paris}
\end{itemize}

Our goal: distinguish \textbf{good} topics from \textbf{poor} topics!
Motivation

- Negative correlation (!) between
  - Human evaluation & Statistical evaluation
    [Chang+ NIPS’09]

- Applications where human-perceived quality essential:
  - doc-doc similarity (via topic distribution)
  - word-sense disambiguation
  - multi-doc summarization
Main idea: exploit Wikipedia

Understand how humans navigate Wikipedia

Get an idea of how people connect concepts

[West-Leskovec, 2012]
Main idea: exploit Wikipedia

Wikipedia is a graph!
node = entity & (directed) edge = pagelink

Intuition: Good topic words are conceptually “coherent” → “close-by” in WikiLinks
Our Approach

1. Topic Model
   \[ M \]
   Topics
   - T1. steam engine valve piston ...
   - T2. cut system capital building ...
   - T3. pain disease cases fever ...
   - T4. furniture chair table wood ...
   - T5. modern view study face ...
   - T6. porcelain pottery ceramic ...

2. Wikipedia Links Graph

3. Projection Graph

4. Evidential Graph-based Features

5. Predictions
   Learning / Prediction Model
   Labels (Good / Poor)

Chan & Akoglu
External Evaluation of Topic Models: A Graph Mining Approach
Projection Graph

Example topics T1 (high-quality) and T2 (low-quality) of a topic model.

T1: steam, engine, valve, piston, cylinder, pressure, boiler, air, pump, pipe

T2: cut, system, capital, pointed, opening, building, character, round, france, paris

match T2 words

match T1 words

PROJ T1

PROJ T2

WikiLinks Graph
Spanning Graph

PROJ T1
WikiLinks induced on topic-words

SPAN T1
Connection subgraph of topic-words
Our Approach

1. Topic Model

M

Topics

T1. steam engine valve piston ...
T2. cut system capital building ...
T3. pain disease cases fever ...
T4. furniture chair table wood ...
T5. modern view study face ...
T6. porcelain pottery ceramic ...

2. Wikipedia Links Graph

3. Projection Graph

4. Evidential Graph-based Features

5. Predictions

Learning / Prediction Model

Labels
(Good / Poor)
All Features: 3 groups

PROJ
- Boiler
- Valve
- Air
- Pipe
- Pressure
- Piston
- Cylinder
- Pump

SPAN
- Air
- Atmosphere_of_Earth
- Water_vapor
- Steam
- Piston
- Piping
- Pipe
- Cylinder

SP
- Valve
- Engine
- Boiler
- Pressure
- Pump
- Cylinder_(engine)

4 features
E.g.
Largest CC Size

9 features
E.g.
#Connector Nodes

9 features
E.g.
Average SP

Air ... Pipe

Air 1 5
Shortest Path
Pipe NP 1
### Labels: Good vs. Poor Topics

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Documents</th>
<th># Topics</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Press</td>
<td>2,246</td>
<td>100</td>
<td>No</td>
</tr>
<tr>
<td>Brain Injury</td>
<td>10,000</td>
<td>200</td>
<td>No</td>
</tr>
<tr>
<td>Books</td>
<td>12,000</td>
<td>120</td>
<td>Yes</td>
</tr>
<tr>
<td>News</td>
<td>55,000</td>
<td>117</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Relative Quality Prediction
*(Labels – Generated)*

T topics; W words; Top 10 vs Top X-Y

\[ w_1, w_2, \ldots, w_{10}, \ldots, w_{11}, \ldots, w_{20}, \ldots, w_{31}, \ldots, w_{40}, \ldots, w_{91}, \ldots, w_{100} \]

### Absolute Quality Prediction
*(Labels – Human Annotators)*

*We thank David Newman and his group for sharing these data*
Our Approach

1. **Topic Model**
   - **M**

2. **Wikipedia Links Graph**

Topic Model Topics

- T1. steam engine valve piston ...
- T2. cut system capital building ...
- T3. pain disease cases fever ...
- T4. furniture chair table wood ...
- T5. modern view study face ...
- T6. porcelain pottery ceramic ...

3. **Projection Graph**

4. **Evidential Graph-based Features**

5. **Predictions**
   - Learning / Prediction Model
   - Labels (Good / Poor)
Relative Quality Prediction

Consistent higher accuracy for easier relative tasks

<table>
<thead>
<tr>
<th>Feature set</th>
<th>top-10 vs.</th>
<th>top-[11-20]</th>
<th>top-[31-40]</th>
<th>top-[91-100]</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE-MAJORITY</td>
<td>PRESS=0.500 BRAIN=0.500</td>
<td>PRESS=0.500 BRAIN=0.500</td>
<td>PRESS=0.500 BRAIN=0.500</td>
<td>PRESS=0.500 BRAIN=0.500</td>
</tr>
<tr>
<td>PROJ</td>
<td>PRESS=0.505 BRAIN=0.622</td>
<td>PRESS=0.715 BRAIN=0.705</td>
<td>PRESS=0.765 BRAIN=0.725</td>
<td></td>
</tr>
<tr>
<td>D-SPAN</td>
<td>PRESS=0.650 BRAIN=0.687</td>
<td>PRESS=0.760 BRAIN=0.740</td>
<td>PRESS=0.805 BRAIN=0.762</td>
<td></td>
</tr>
<tr>
<td>D-SP</td>
<td>PRESS=0.605 BRAIN=0.665</td>
<td>PRESS=0.710 BRAIN=0.760</td>
<td>PRESS=0.750 BRAIN=0.790</td>
<td></td>
</tr>
<tr>
<td>PROJ+D-SPAN</td>
<td>PRESS=0.650 BRAIN=0.687</td>
<td>PRESS=0.745 BRAIN=0.722</td>
<td>PRESS=0.790 BRAIN=0.777</td>
<td></td>
</tr>
<tr>
<td>PROJ+D-SP</td>
<td>PRESS=0.650 BRAIN=0.672</td>
<td>PRESS=0.710 BRAIN=0.752</td>
<td>PRESS=0.815 BRAIN=0.800</td>
<td></td>
</tr>
<tr>
<td>PROJ+D-SPAN+D-SP</td>
<td>PRESS=0.660 BRAIN=0.687</td>
<td>PRESS=0.735 BRAIN=0.752</td>
<td>PRESS=0.810 BRAIN=0.807</td>
<td></td>
</tr>
</tbody>
</table>

>15% improvement over random baseline

>23%

>30%
Absolute Quality Prediction – Baselines

Google Features:
1. allintext
2. intitle
3. inanchor
4. inurl

Baseline (1)

Personalize PageRank:
1. average pairwise score
2. median pairwise score
3. average pairwise rank
4. median pairwise rank

Baseline (2)
### Absolute Quality Prediction

<table>
<thead>
<tr>
<th>Feature set</th>
<th>BOOKS</th>
<th>NEWS</th>
<th>BOOKS + NEWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE-MAJORITY</td>
<td>0.610</td>
<td>0.521</td>
<td>0.549</td>
</tr>
<tr>
<td>BASELINE-GOOGLE</td>
<td>0.642</td>
<td>0.624</td>
<td>0.629</td>
</tr>
<tr>
<td>BASELINE-PPR</td>
<td>0.842</td>
<td>0.735</td>
<td>0.785</td>
</tr>
<tr>
<td>PROJ</td>
<td>0.875</td>
<td>0.812</td>
<td>0.848</td>
</tr>
<tr>
<td>D-SPAN</td>
<td>0.892</td>
<td>0.769</td>
<td>0.844</td>
</tr>
<tr>
<td>D-SP</td>
<td>0.883</td>
<td>0.786</td>
<td><strong>0.852</strong></td>
</tr>
<tr>
<td>PROJ+D-SPAN</td>
<td>0.883</td>
<td>0.795</td>
<td>0.844</td>
</tr>
<tr>
<td>PROJ+D-SP</td>
<td>0.892</td>
<td>0.795</td>
<td>0.848</td>
</tr>
<tr>
<td>PROJ+D-SPAN+D-SP</td>
<td><strong>0.900</strong></td>
<td><strong>0.821</strong></td>
<td>0.831</td>
</tr>
</tbody>
</table>

6% – 9% improvement over ALL baselines
Our graph-centric features are domain-independent (only based on “graph closeness”)

<table>
<thead>
<tr>
<th></th>
<th>Test</th>
<th>BOOKS</th>
<th>NEWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>BOOKS</td>
<td>0.900</td>
<td>0.769</td>
</tr>
<tr>
<td></td>
<td>NEWS</td>
<td>0.867</td>
<td>0.821</td>
</tr>
</tbody>
</table>
Learned Coefficients

<table>
<thead>
<tr>
<th>Selected Feature</th>
<th>Coef: BOOKS</th>
<th>Coef: NEWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_M \text{NumMiss}$</td>
<td>0.0626</td>
<td>0.0918</td>
</tr>
<tr>
<td>$g_S \text{RatioC}$</td>
<td>0.2940</td>
<td>0.5909</td>
</tr>
<tr>
<td>$g_M \text{MaxDeg}$</td>
<td>-0.2921</td>
<td>-0.4541</td>
</tr>
<tr>
<td>$g_M \text{SizeMaxComp}$</td>
<td>-0.8667</td>
<td>--</td>
</tr>
<tr>
<td>$g_S \text{AvgMSTWeight}$</td>
<td>--</td>
<td>0.2598</td>
</tr>
<tr>
<td>NumSP2</td>
<td>-0.9685</td>
<td>--</td>
</tr>
</tbody>
</table>

**Good topics:**

- Fewer missing (matched) words on WikiLinks
- Fewer connector nodes (in spanning graph)
- Higher maximum degree (in projection graph)
Thank you!

- SBU Office of the Vice President for Research
- NSF Graduate Research Fellowship
- ICDM 2013 Travel Grants

hauchan@cs.stonybrook.edu
http://www.cs.sunysb.edu/~hauchan
4. Prediction

- For **relative quality prediction**, we consider topics with 20, 40, and 100 words (ordered by descending probability of describing the topics) where the topics are generated by LDA
  - Positive Class: top 10 words for each topic
  - Negative Class: Bottom 11-20, Bottom 31-40, and Bottom 91-100 words for each topic

- For **absolute quality prediction**
  - Positive Class: Good topics
  - Negative Class: Bad topics
5. Learning Model

- Logistics Regression Classifier
  - L1-Norm Regularization

- We report Leave-One-Out Cross-Validation (LOOCV)