Research Topics Evolution Tool: NE Viewer

Penghui Lyu

Wuhan University
Outline

Introduction

Processing

Functions

Conclusion
Usually, the authors of paper are required to provide several keywords to represent the paper’s topic.

Usually, network has hierarchical structure.

By adding the temporal information, we will be able to investigate the evolution of co-word networks.
Kinds of Network Evolution

- Subject Networks;
- Cooperation Networks;
- Co-words Networks;
- Co-citation Networks;
- And so on……

We want to develop a tool to detect and visualize the evolution of network, especially co-word network.
If each community was assumed to be a topic, then the evolution of topics can be investigated by investigating the evolution of network.
Framework of NEViewer’s methods

CORPUS → Preprocess → Network representation → Temporal data process

Community detection → Evolution analysis → Visualization

Cluster Naming

Topic representation | Topic detection | Evolution analysis | Visualization
Community Detection

- the network topology based approach
- the content-based topical analysis approach

NEViewer implements three algorithms now:

- Blondel’s approach *(Blondel, Guillaume, Lambiotte, & Lefebvre, 2008)*
- Ball’s Approach *(Ball, Karrer and Newman 2011)*
- Modularity Maximization Algo *(Newman, 2004)*

- Blondel’s approach is the default option
Cluster Naming

- Give each community a title.
  to find the representative topics for each community, i.e. the representative node finding problem.

- Our solution:
  Z-Value

\[ Z_i = \frac{1}{\sqrt{\langle (k_{s_i})^2 \rangle - \langle k_{s_i} \rangle^2}} \]

Cluster Naming

- Give each community a title.

To find the representative topics for each community, i.e. the representative node finding problem.

Each detected community can be represented by one or more nodes whose

\[ Z\text{-value} \geq 2.5. \]

Community Evolution Analysis

- Six different forms of evolution
  - Birth
  - Growth
  - Merging
  - Contraction
  - Splitting
  - Death

In fact, we simplify the evolution analysis as finding the appropriate successors and predecessors except the *Death* of one community.

\[
Pre(M_{(t+1)j}) = (M_{ti} | M_{ti} \in G_t, d(M_{ti}, M_{(t+1)j}) < \delta) \cup \arg\max_{M_{ti} \in G_t}(d(LM_{ti}, LM_{(t+1)j}))
\]

\[
FS(M_x, M_y) = EJ(N_x, N_y) \ast HS(M_x, M_y) \ast ES(M_x, M_y)
\]
Evolution Visualization

- Alluvial diagram
- Forward and backward coloring
Alluvial Diagram

1. colored rectangle areas represent each community;
2. the colored curve areas between two time stamps denote the evolution process:
3. split, merge, death …

Micro view

- Backward Coloring
- Forward Coloring
The Tool: NEViewer
The Tool: NEViewer

- NEViewer (Network Evolution Viewer).
  - implemented in Java
  - support the NWB file format (.nwb)

- implemented all of the algorithms we mentioned above

- community detection
  - Blondel algorithm (Blondel et al., 2008)
  - Ball’s algorithm on overlap community (Ball, Karrer, & Newman, 2011).
The Tool: NEViewer

- **NEViewer (Network Evolution Viewer).**
  - implemented in Java
  - support the NWB file format (.nwb)
- implemented all of the algorithms we mentioned above
- **community similarity measures**
  - Jaccard similarity
  - Tanimoto similarity
  - cosine similarity
  - overlap of core nodes
  - FS measure
The Tool: NEViewer

- **NEViewer (Network Evolution Viewer)**.
  - implemented in Java
  - support the NWB file format (.nwb)
- implemented all of the algorithms we mentioned above
- **Visualization**
  - alluvial diagram
  - the colored network diagram.
  - Users can choose the network layout.
- Some basic metrics are also supported in this software such as the PageRank value and the centrality measures.
Outline

Introduction

Processing

Functions

Conclusion
A FIVE-CONF dataset was collected
KDD, SIGIR, CIKM, CSCW and JCDL.
The FIVE-CONF dataset contains 7,234 papers that were published between 2000 and 2011
Construct Sequential data

- $T_1 = [t_2000, t_2003]$, $T_2 = [t_2004, t_2007]$ and $T_3 = [t_2008, t_2011]$. There are 2480 papers in $T_1$, 4283 papers in $T_2$ and 5517 papers in $T_3$.

<table>
<thead>
<tr>
<th>Measurements</th>
<th>Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T_1$</td>
</tr>
<tr>
<td>Nodes</td>
<td>1217</td>
</tr>
<tr>
<td>Isolated Nodes</td>
<td>9</td>
</tr>
<tr>
<td>Edges</td>
<td>12076</td>
</tr>
<tr>
<td>Mean degree</td>
<td>19.846</td>
</tr>
<tr>
<td>largest connected component</td>
<td>1202</td>
</tr>
<tr>
<td>Density</td>
<td>0.00816</td>
</tr>
</tbody>
</table>
Community Detection

- the Blondel’s community detection approach was adopted (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Communities</th>
<th>Average number of nodes in each community</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>23</td>
<td>50.71</td>
</tr>
<tr>
<td>T2</td>
<td>34</td>
<td>65.57</td>
</tr>
<tr>
<td>T3</td>
<td>41</td>
<td>69.12</td>
</tr>
</tbody>
</table>

Case Study- Macro View
IR as an Example

<table>
<thead>
<tr>
<th></th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Web Search</strong></td>
<td><strong>Classification and clustering</strong></td>
<td><strong>User Modeling</strong></td>
<td><strong>Learning to Rank</strong></td>
<td><strong>Social Media</strong></td>
</tr>
<tr>
<td>Evaluation</td>
<td>Expansion and feedback</td>
<td>Applications</td>
<td>Retrieval models</td>
<td></td>
</tr>
<tr>
<td>Collaborative Filtering</td>
<td>Web 2.0</td>
<td>Search Engine Architectures and Scalability</td>
<td></td>
<td>Social Media</td>
</tr>
<tr>
<td>Learning to Rank</td>
<td>Retrieval models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-Performance &amp; High Dimensional Indexing</td>
<td>Speech and linguistic processing</td>
<td>Learning to Rank</td>
<td>Collaborative filtering</td>
<td></td>
</tr>
<tr>
<td>User Adaptation &amp; Personalization</td>
<td>Recommenders</td>
<td>Filtering and Recommendation</td>
<td>Query Analysis</td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td>Question answering</td>
<td>Information Retrieval Theory</td>
<td>Communities</td>
<td></td>
</tr>
<tr>
<td>Multilingual &amp; Crosslingual Retrieval</td>
<td>Efficiency</td>
<td>Language Models &amp; IR Theory</td>
<td>Image Search</td>
<td></td>
</tr>
<tr>
<td>Relevance Feedback</td>
<td><strong>Web retrieval</strong></td>
<td>Query Representations &amp; Reformulations</td>
<td>Web Queries</td>
<td></td>
</tr>
<tr>
<td>Summarization</td>
<td>Learning to Rank</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exploatory Search &amp; Filtering</td>
<td>Information extraction</td>
<td>Retrieval Models and Ranking</td>
<td>Multimedia IR</td>
<td></td>
</tr>
<tr>
<td>Multimedia Retrieval</td>
<td>Citation models</td>
<td>User Feedback &amp; User Models</td>
<td>Summarization</td>
<td></td>
</tr>
<tr>
<td>Query Analysis &amp; Models</td>
<td>Vertical search</td>
<td>Web IR and <strong>Social Media Search</strong></td>
<td>Query suggestions</td>
<td></td>
</tr>
<tr>
<td>Non-Topicality</td>
<td>Interactive search</td>
<td>Document Structure &amp; Adversarial Information Retrieval</td>
<td>Linguistic Analysis</td>
<td></td>
</tr>
<tr>
<td>Probabilistic Models</td>
<td>Multimedia</td>
<td>Users and Interactive IR</td>
<td>Effectiveness</td>
<td></td>
</tr>
<tr>
<td>Analysis of Social Networks</td>
<td>Federated, distributed search</td>
<td>Document Representation and Content Analysis</td>
<td>Multilingual IR</td>
<td></td>
</tr>
<tr>
<td>Question-Answering</td>
<td>Industry Track speakers</td>
<td>Test Collections</td>
<td>Recommender systems</td>
<td></td>
</tr>
<tr>
<td>Social Tagging</td>
<td>Evaluation and measurements</td>
<td>Query Log Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content Analysis</td>
<td>Query formulation</td>
<td>Summarization &amp; User Feedback</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning Models for IR</td>
<td>Spanning</td>
<td>Query Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Text Classification</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Backward Coloring View

(a) backward colored network for “Classification”  (b) backward colored network for “Search”

(c) backward colored network for “Classification”  (d) backward colored network for “Search”

(e) backward colored network for “Search”
Forward and Backward Coloring
Forward and Backward Coloring
Outline

Introduction

Processing

Functions

Conclusion
Conclusion

- NEViewer: community detection, community mapping, representative node finding, and evolution visualization.

- The novelty of NEViewer are:
  1. Develop four steps in analyzing network evolution;
  2. Implement the alluvial diagram;
  3. Design and implement the Backward and Forward colored network diagram.
Conclusion

- Limitations:
- Noise removal
- Clustering performance is not so good.
- Lack a firm threshold and principles of predefining the thresholds
Scientific foundation application

- Research on Formation Mechanism and Evolution Laws of Knowledge Networks, Natural Science Foundation of China (NSFC: 71173249)
- Smart City Digital Information Resources Security (NSFC: 71473182)
- Research on Knowledge Organization and Service Innovation in the Big Data Environments (NSFC: 71420107026).
Thanks All!