Evolutionary Clustering and Analysis of Heterogeneous Bibliographic Networks

Manish Gupta
University of Illinois at Urbana Champaign
gupta58@illinois.edu

Urbana-Champaign, May, 2010
Contributions

- First to present an evolutionary clustering algorithm for heterogeneous information networks
- Define metrics to characterize clustering behavior: Consistency, Quality, Cluster appearance/disappearance, Continue/Merge/Split rates, Stability of objects, Sociability of objects, Social influence
- Evolutionary study of DBLP
- Uses NetClus framework
Problem Formulation

- Net-Cluster, Net-Cluster tree, Net-Cluster tree sequence
- Problem: Given a graph sequence GS, generate a net-cluster tree sequence CTS such that the trees are consistent and represent high-quality clusters.
- Army faces evolutionary adhoc networks on battlefields composed of soldiers, tanks, bunkers, local reporting stations, aircrafts. It would be informative to see how each of the battalions are performing as a group.
- Task 4.1: Link-based and density-based clustering of heterogeneous information networks
Algorithm 1 NetClus with Evolution-Aware Priors

1: Priors: Initialize prior probabilities \( \{P(o|c_k^t)\}_{k=1}^K \).
2: Initialize: Generate initial net-clusters. \( \{c_k\}_{k=1}^K \).
3: Rank: Build probabilistic generative model for each net-cluster, i.e., \( \{P(o|c_k^t)\}_{k=1}^K \).
4: Cluster-target: Compute \( p(c_k^t|o) \) for target objects and adjust their cluster assignments.
5: Iterate: Repeat steps 3 and 4 until the clusters don’t change significantly.
6: Cluster-attribute: Calculate \( p(c_k^*|o) \) for each attribute object in each net-cluster.
7: return \( p(c_k^*|o) \)
Algorithm …

- Simple clustering of each snapshots independently does not offer consistent clusters.
- For the first time instant, initialization of priors and net clusters is similar to NetClus.
- For other time instants
  - Use representativeness to define priors
  - Choose initial clusters intelligently.
- Ranking: Prior weight controls the effect of priors and hence the temporal smoothness.
Evaluation Metrics

- **Consistency**
  - Membership probability of object $o$ of type $t$ to cluster $c_i$ is denoted by $\{b_i(o)\}_{i=1}^{K}$
  - Consistency($\text{clustering c, y1, y2}$)

$$
\frac{1}{|O|} \sum_{o \in O} \frac{\sum_{k=1}^{K} b_k(o)_{y1} \times b_k(o)_{y2}}{\sqrt{\sum_{k=1}^{K} b_k(o)_{y1}^2 \sqrt{\sum_{k=1}^{K} b_k(o)_{y2}^2}}}
$$

- Chained path consistency

- **Snapshot Quality**
  - Compactness $C = \frac{1}{|O|} \sum_{k=1}^{K} \sum_{i=1}^{\left|O_k\right|} \frac{s(o_{ki}, c_k)}{\sum_{k' \neq k} s(o_{ki}, c'_{k'})/(K-1)}$

  - Entropy $E = -\frac{1}{|O|} \sum_{k=1}^{K} \sum_{o=1}^{\left|O_k\right|} b_k(o) \times \log(b_k(o))$. 
Quantifying Evolution

Continue rate of cluster \( c_i = \frac{1}{|O|} \sum_{o \in O} \min \left( \frac{b_i(o)_y}{b_i(o)_{y-1}}, 1 \right) \)

Merge rate of cluster \( c_i = \frac{1}{|O|} \sum_{o \in O} \max \left( \frac{b_i(o)_y - b_i(o)_{y-1}}{b_i(o)_y}, 0 \right) \)

Split rate of cluster \( c_i = \frac{1}{|O|} \sum_{o \in O} \max \left( \frac{b_i(o)_{y-1} - b_i(o)_y}{b_i(o)_{y-1}}, 0 \right) \)

Appearance rate \( = \frac{\sum_{o \in O'} b_c(o)_y}{\sum_{o \in O''} b_c(o)_y} \)

Disappearance rate \( = \frac{\sum_{o \in O'''} b_c(o)_y}{\sum_{o \in O''} b_c(o)_y} \)
Tracking object evolution

- Stability of objects
  - Simple temporal stability
  - Sequential temporal stability
  - Maximum sequential temporal stability
  - Simple social stability
  - Ranked social stability
  - Relative stability

- Sociability of objects

- Effect of social influence: normality
Experiments

Dataset
DBLP (1993 to 2008, 654K papers, 484K authors, 107K title terms and 3900 conferences)
Four area (DM, DB, IR, ML papers; 1993 to 2008, 24K papers, 26K authors, 12K title terms, 20 conferences)
Experiments ...

- Rates are higher at deeper levels.

- Objects maintain their cluster membership distribution up to a degree of 70.

- Most representative objects in the cluster continue to be stable, whereas the “modestly” representative objects may vary more significantly.

- There has been substantial mutual influence between the DD and IN areas. We also notice ML to IN influence. Mutual influence between DM and ML is quite natural.
Related work

- Clustering graphs: mincut, min-max cut, spectral, density-based, RankClus [Sun EDBT 09], NetClus [Sun KDD 09]
- Evolutionary clustering: k-means [Chak KDD06], spectral [Chi KDD07], text streams [Mei KDD05], social network structure [Kuma KDD06]
- Evolutionary graph studies: GraphScope [Sun KDD07], density-based [Kim VLDB09], analysis [Back KDD06, Lesk KDD05, Lesk KDD08], communities using FacetNet [Lin WWW08], individual objects [Asur KDD07]
Future directions

- Variable number of clusters at different time periods
- Study the effect of compactness for different time granularities and when priors are defined for different node types
- Identifying outliers in the network both in the static as well as evolutionary sense.
Cluster-based veracity analysis

**Truth finder algorithm sketch**

\[ s(f) = 1 - \prod_{w \in W(f)} (1 - t(w)) \]

\[ \sigma(f) = -\ln(1 - s(f)) \]

\[ \sigma^*(f) = \sigma(f) + \rho \cdot \sum_{\sigma(f') = \sigma(f)} \sigma(f') \cdot \text{imp}(f' \rightarrow f) \]

\[ s^*(f) = 1 - e^{-\sigma^*(f)} \]

\[ t(w) = \frac{\sum_{f \in F(w)} s(f)}{|F(w)|} \]
Approach

- Project 5: Trust CCRI: Designing Trusted Information Networks
- Trustworthy providers influence confident facts, facts influence themselves and providers, rankings of facts and providers influence clustering space.
- Different clusters have different levels of falsity. Also, different providers have different trustworthiness scores in different clusters.
- We consider 5 categories of books: business, children books, fiction, history, science.
<table>
<thead>
<tr>
<th>Provider</th>
<th>Rank in full</th>
<th>Rank in childrens_books</th>
<th>Rank in history</th>
<th>Rank in business</th>
<th>Rank in fiction</th>
<th>Rank in science</th>
</tr>
</thead>
<tbody>
<tr>
<td>goldstone books</td>
<td>1</td>
<td>54</td>
<td>272</td>
<td>47</td>
<td>2</td>
<td>63</td>
</tr>
<tr>
<td>ps books</td>
<td>2</td>
<td>29</td>
<td>121</td>
<td>20</td>
<td>24</td>
<td>NA</td>
</tr>
<tr>
<td>brit books</td>
<td>3</td>
<td>23</td>
<td>133</td>
<td>7</td>
<td>5</td>
<td>101</td>
</tr>
<tr>
<td>follett educational services</td>
<td>4</td>
<td>71</td>
<td>8</td>
<td>79</td>
<td>422</td>
<td>1</td>
</tr>
<tr>
<td>at a glance books</td>
<td>5</td>
<td>133</td>
<td>69</td>
<td>50</td>
<td>124</td>
<td>3</td>
</tr>
<tr>
<td>the used book hounds</td>
<td>6</td>
<td>7</td>
<td>174</td>
<td>66</td>
<td>70</td>
<td>89</td>
</tr>
<tr>
<td>bdwe ltd</td>
<td>7</td>
<td>254</td>
<td>NA</td>
<td>143</td>
<td>3</td>
<td>NA</td>
</tr>
<tr>
<td>free shipping books</td>
<td>8</td>
<td>22</td>
<td>13</td>
<td>10</td>
<td>17</td>
<td>107</td>
</tr>
<tr>
<td>break time books</td>
<td>9</td>
<td>11</td>
<td>37</td>
<td>29</td>
<td>13</td>
<td>127</td>
</tr>
<tr>
<td>the saint bookstore</td>
<td>10</td>
<td>1</td>
<td>27</td>
<td>1</td>
<td>28</td>
<td>2</td>
</tr>
<tr>
<td>kenago books</td>
<td>11</td>
<td>31</td>
<td>78</td>
<td>23</td>
<td>23</td>
<td>NA</td>
</tr>
<tr>
<td>ceredigion book shop</td>
<td>12</td>
<td>105</td>
<td>726</td>
<td>105</td>
<td>15</td>
<td>207</td>
</tr>
<tr>
<td>daleside books</td>
<td>13</td>
<td>110</td>
<td>784</td>
<td>99</td>
<td>7</td>
<td>205</td>
</tr>
<tr>
<td>great time books</td>
<td>14</td>
<td>15</td>
<td>19</td>
<td>24</td>
<td>19</td>
<td>72</td>
</tr>
<tr>
<td>book drives incorporated</td>
<td>15</td>
<td>NA</td>
<td>94</td>
<td>28</td>
<td>90</td>
<td>22</td>
</tr>
<tr>
<td>wv company</td>
<td>16</td>
<td>450</td>
<td>103</td>
<td>14</td>
<td>128</td>
<td>30</td>
</tr>
<tr>
<td>the book women</td>
<td>17</td>
<td>88</td>
<td>48</td>
<td>546</td>
<td>14</td>
<td>340</td>
</tr>
<tr>
<td>look at a book</td>
<td>18</td>
<td>63</td>
<td>31</td>
<td>63</td>
<td>92</td>
<td>24</td>
</tr>
<tr>
<td>allamericantextbooks.com</td>
<td>19</td>
<td>128</td>
<td>14</td>
<td>33</td>
<td>69</td>
<td>4</td>
</tr>
<tr>
<td>books_at_half</td>
<td>20</td>
<td>28</td>
<td>490</td>
<td>NA</td>
<td>55</td>
<td>308</td>
</tr>
</tbody>
</table>
Ongoing work

- Explore applicability of RankClus for veracity analysis. RankClus is an iterative ranking and clustering philosophy.
- Compute cluster-wide trust scores and smooth them by global scores.
- Generating an appropriate dataset which has natural clusters.
Thank you

May, 2010, Urbana-Champaign