Graph Theoretical Clustering

Agglomerative: Bottom up

We use the object to start the initial clusters then use the cluster to obtain higher level clusters.

Types: Single-link, complete-link, average-link. Produced clustering bstructure is called dendrogram.

Single-link: The similarity between a pair of clusters is taken to be the similarity between the most similar pair of documents one of which appears in each cluster; thus each cluster member will be more similar to at least one member in the same cluster then to any member of another cluster.

Complete-link: The similarity between the least similar pair of items from the clusters is used as the cluster similarity. Each cluster member is more similar to the most dissimilar member of its own cluster than the most similar of any other cluster.

Single-link example:

\[
S = \begin{bmatrix}
A & B & C & D \\
A & 1.0 & 0.3 & 0.5 & 0.6 \\
B & - & 1.0 & 0.4 & 0.5 \\
C & - & - & 1.0 & 0.3 \\
D & - & - & - & 1.0 \\
\end{bmatrix}
\]

<table>
<thead>
<tr>
<th>Step</th>
<th>Pair</th>
<th>Sim</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AD</td>
<td>0.6</td>
</tr>
<tr>
<td>2</td>
<td>AC</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>BD</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>BC</td>
<td>0.4</td>
</tr>
<tr>
<td>5</td>
<td>AB</td>
<td>0.3</td>
</tr>
<tr>
<td>6</td>
<td>CD</td>
<td>0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step</th>
<th>Sim.Pair</th>
<th>Single-Link Str.</th>
<th>Items Covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AD,0.6</td>
<td>0.6</td>
<td>A,D</td>
</tr>
<tr>
<td>2</td>
<td>AC,0.5</td>
<td>0.5</td>
<td>A,C,D</td>
</tr>
<tr>
<td>3</td>
<td>BD,0.5</td>
<td>0.5</td>
<td>A,B,C,D</td>
</tr>
</tbody>
</table>

Similarity matrix implied by the clustering structure

\[
S' = \begin{bmatrix}
A & B & C & D \\
A & 1.0 & 0.5 & 0.5 & 0.6 \\
B & - & 1.0 & 0.5 & 0.5 \\
C & - & - & 1.0 & 0.5 \\
D & - & - & - & 1.0 \\
\end{bmatrix}
\]

This has higher values than the original matrix.
Measure the similarity between the original S matrix and the S’ matrix derived from the clustering structure. If it is high we can say that clustering represents/retains the characteristics of the data sets.

**Complete-link example:**

<table>
<thead>
<tr>
<th>Step</th>
<th>Sim.Pair</th>
<th>Complete-Link Str.</th>
<th>Items Covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AD,0.6</td>
<td>0.6</td>
<td>AD</td>
</tr>
<tr>
<td>2</td>
<td>AC,0.5</td>
<td>Too early to connect since we do not know C,D.</td>
<td>AD,AC</td>
</tr>
<tr>
<td>3</td>
<td>BD,0.5</td>
<td>Too early; we do not know A,B.</td>
<td>AD,AC,BD</td>
</tr>
<tr>
<td>4</td>
<td>BC,0.4</td>
<td>0.6 0.4</td>
<td>AD, AC, BC, BD</td>
</tr>
<tr>
<td>5</td>
<td>AB,0.3</td>
<td>0.6 0.4</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CD,0.3</td>
<td>0.6 0.4</td>
<td></td>
</tr>
</tbody>
</table>

Similarity matrix implied by the clustering structure:

\[
S' = \begin{bmatrix}
A & 1.0 & 0.3 & 0.3 & 0.6 \\
B & - & 1.0 & 0.4 & 0.3 \\
C & - & - & 1.0 & 0.3 \\
D & - & - & - & 1.0
\end{bmatrix}
\]

**Cover Coefficient based Clustering Methodology (C³M)**

ACM Trans- on Database Systems (Can, Özkarahan, "Concepts and effectiveness of the cover coefficient-based clustering methodology for text databases." ACM TODS, 1990.)

- Single-pass
- Seed based: Select some objects as cluster initiators (seeds) and assign non-seed s to these clusters.
- Partitioning type clustering algorithm (No overlap)
- Provides a way of obtaining no. of clusters
C³M Steps
1. Find no of clusters
2. Select cluster seeds
3. Assign non-seeds to seeds

Example
\[
D = \begin{bmatrix}
1 & 0 & 0 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 0 & 0 \\
\end{bmatrix}
\]

\[t_1 \quad t_2 \quad t_3 \quad t_4 \quad t_5 \quad t_6\]

\[d_1 \]
\[d_2 \]
\[d_3 \]
\[d_4 \]
\[d_5 \]

\[A: \text{Inverse of row sums}\]
\[
\frac{1}{3}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{1}{2}\]

\[B: \text{Inverse of column sums}\]
\[
\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, \frac{1}{4}, \frac{1}{2}, \frac{1}{3}\]

\[m: \text{no. of docs (5)}\]
\[n: \text{no. of terms (6)}\]

\[
S = \begin{bmatrix}
\frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\
\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\
\frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\
0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & 0 & 1/2 & 0 & 0 & 0 \\
\end{bmatrix},
S' = \begin{bmatrix}
\frac{1}{4} & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{3} \\
\frac{1}{4} & 1 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\
\frac{1}{4} & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{3} \\
0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{3} \\
\frac{1}{4} & 0 & 1/2 & 0 & 0 & 0 \\
\end{bmatrix}
\]

\[S \text{ for rows, } S' \text{ for columns.}\]

\[C = S \times S'^T \Rightarrow S_{m \times n} \times S'^T_{n \times m} = C_{m \times m} \]

\[C: \text{cover coefficient matrix}\]

\[
\begin{array}{c}
(1/3) t_1 \\
(1/3) t_4 \\
(1/3) t_6 \\
\end{array}
\begin{array}{c}
\frac{1}{3} d_1 \\
\frac{1}{3} d_3 \\
\frac{1}{3} d_4 \\
\end{array}
\]

Double Stage Probability experiment
We either use S and S' or $C_{ij}$ calculation above to get the C.

$$c_{ij} = \alpha_i \times \sum_{k=1}^{n}(d_{ik} \times \beta_k \times d_{jk})$$

$c_{ij}$ = probability of selecting a term of $d_i$ from $d_j$

$$c_{11} = \alpha_1 \times \sum_{k=1}^{6}(d_{1k} \times \beta_k \times d_{1k}) = \frac{1}{3} \times \left(\frac{1}{4} \times \frac{1}{2} \times \frac{1}{3}\right) = 0.369$$

We either use S and S’ or $C_{ij}$ calculation above to get the C.

$$C = \begin{bmatrix}
0.361 & 0.250 & 0.194 & 0.111 & 0.083 \\
0.188 & 0.563 & 0.063 & 0.000 & 0.188 \\
0.194 & 0.083 & 0.361 & 0.277 & 0.083 \\
0.167 & 0.000 & 0.417 & 0.417 & 0.000 \\
0.125 & 0.000 & 0.125 & 0.000 & 0.375
\end{bmatrix}$$

- $c_{ij} = 0$ then $c_{ji} = 0$
- $c_{ij} > 0$ then $c_{ji} > 0$
- If $c_{ij} = 0.000$ then $d_i$ and $d_j$ do not have any term in common.
- Row sums = 1
- $c_{ii} \geq c_{ij}$, $i \neq j$, ($1 \leq i, j \leq m$)
- If $d_i$ and $d_j$ are identical, their rows are identical
- For a unique document (its terms do not appear in other docs)
  $c_{ii} = 1$, $c_{ij} = 0$ for $i \neq j$
- If all documents are identical then $c_{ij} = \frac{1}{m}$, for ($1 \leq i, j \leq m$)
- If all documents are unique then $C = \begin{bmatrix}1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1\end{bmatrix}$ (identity matrix)

$$n_c = m$$
$$C = \begin{bmatrix}1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1\end{bmatrix}$$

$$n_c = 1$$
$$C = \begin{bmatrix}\frac{1}{m} & \frac{1}{m} & \frac{1}{m} \\ \frac{1}{m} & \ddots & \frac{1}{m} \\ \frac{1}{m} & \frac{1}{m} & \frac{1}{m}\end{bmatrix}$$

$n_c$: no of clusters

- $\min(c_{ii}) = 1/m$
- If $d_i$ and $d_j$ are identical, $c_{ii} = c_{jj} = c_{ij} = c_{ji}$
- If a document has a few number of common terms with the other documents, its $c_{ii}$ is relatively high.
\[ c_{ii} = \text{decoupling coefficient} = \delta_i \] (How this document is separated from other documents)

\[ 1 - c_{ii} = \Psi_i: \text{coupling coefficient} \]

\[ n_c = \sum_{i=1}^{m} c_{ii} = \sum_{i=1}^{m} \delta_i \quad \text{avg. Decoupling} \quad \delta = \frac{\sum \delta_i}{m} \quad n_c = m \times \delta \]

**Cover coefficient values for terms**
\[ D \rightarrow C'_n \times n \]

\[ C' = S^T \times S \Rightarrow S^T_{n \times m} \times S_{m \times n} = C'_{n \times n} \quad C': \text{cover coefficient matrix for terms} \]

\[ c'_{ij} = \beta_i \times \sum_{k=1}^{n} (d_{kl} \times \alpha_k \times d_{kj}) \quad (1 \leq i, j \leq n) \]

**Number of term clusters**
\[ c'_{ii} = \delta'_i \]

\[ n'_c = \sum_{i=1}^{n} C'_{ii} = \sum_{i=1}^{n} \delta'_i , \quad \delta' = \frac{\sum \delta'_i}{n} , \quad n'_c = n \times \delta' \]

\[ n_c = n'_c \]

\[ 1 \leq n_c, n'_c \leq \min (m, n) \]

**Average cluster size**
\[ d_c : \text{average no. of docs / cluster} \]

\[ d_c = \frac{m}{n_c} \quad \frac{m}{\min (m, n)} \leq d_c \leq m \quad \text{from terms point of view} \quad \frac{n}{\min (m, n)} \leq d'_c \leq n \]

**Cluster seed power**

Seeds
- Coupled
- Decoupled

How to select cluster seeds?
- Assign a cluster seed value to individual documents
- We can use column sums of \( c \)
  - Too expensive.. need an approximation

Cluster seed power \( p_i \)

\[ p_i = \delta_i \times \Psi_i \times X_{d_i} = \delta_i \times \Psi_i \times \alpha^{-1} \]
Selection of Seeds

1. Calculate $p_i$ values for all docs
2. Sort them in descending order
3. Select top $n_c$ of them as cluster seeds

Questions

1. Which one is more efficient in terms of time: Complete-link or single-link?
   - Single-link is more time efficient. In single link it is enough for the item to be added to have a similarity value greater than 0 to any of the cluster members in one of the clusters. However, in complete link we continue to iterate until we have the same similarity of one of the next items to the existing clusters.

2. Consider the resulting $S'$ matrices out of single-link and complete-link clustering. What are some observations?
   - Resulting $S'$ matrix from single-link algorithm has higher values than the original matrix whereas complete-link clustering has somewhat similar values. This indicates that with single-link algorithm, we lose precision. However with the complete-link algorithm, there is the trade-off of computational time since it lasts longer than the single-link algorithm.

3. What might be the problem with the selection of seeds procedure written above?
   - Since the list is sorted, identical or nearly identical docs will form separate clusters, which is a problem because they need to be in the same cluster.

4. What is meant by a seed to be coupled and decoupled at the same time?
   - A seed must be similar to the ones in its own cluster and must be dissimilar from the seeds of other clusters.

5. Considering the complete-link, single-link algorithms, how can we complete the tree when there is no similarity? How about single items with no similarity to the others?
   - We can connect the tree at the root with the 0.0 value.
   - That type of item can stay singleton (meaning that it’s a cluster on its own) or we can connect them similarly with 0.0 at the top.