CS 533 Spring 2014 Class Notes ({March 3 & 5,.2014)

Tolga Yılmaz

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.

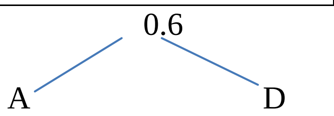
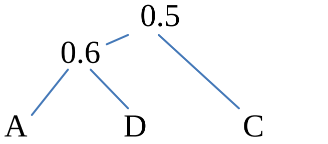
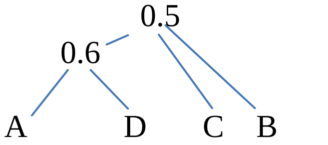
Single-link: The similarity between a pair of clusters is taken to be the similarity between the most similar pair of documents one of which appears in each cluster; thus each cluster member will be more similar to at least one member in the same cluster then to any member of another cluster.

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:

$$S = \begin{array}{cc} & \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{bmatrix} 1.0 & 0.3 & 0.5 & 0.6 \\ - & 1.0 & 0.4 & 0.5 \\ - & - & 1.0 & 0.3 \\ - & - & - & 1.0 \end{bmatrix} \end{array}$$

Step	Pair	Sim
1	AD	0.6
2	AC	0.5
3	BD	0.5
4	BC	0.4
5	AB	0.3
6	CD	0.3

Step	Sim.Pair	Single-Link Str.	Items Covered
1	AD,0.6		A,D
2	AC,0.5		A,C,D
3	BD,0.5		A,B,C,D

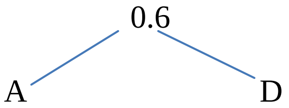
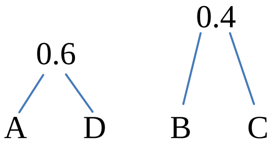
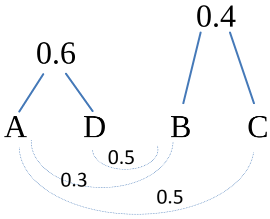
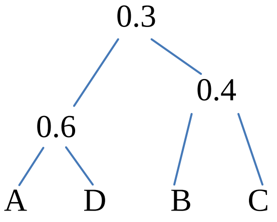
Similarity matrix implied by the clustering structure

$$S' = \begin{array}{cc} & \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{bmatrix} 1.0 & 0.5 & 0.5 & 0.6 \\ - & 1.0 & 0.5 & 0.5 \\ - & - & 1.0 & 0.5 \\ - & - & - & 1.0 \end{bmatrix} \end{array}$$

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:

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

Similarity matrix implied by the clustering structure

$$S' = \begin{array}{c|cccc} & A & B & C & D \\ \hline A & 1.0 & 0.3 & 0.3 & 0.6 \\ B & - & 1.0 & 0.4 & 0.3 \\ C & - & - & 1.0 & 0.3 \\ D & - & - & - & 1.0 \end{array}$$

Cover Coefficient based Clustering Methodology(C^3M)

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

C^{3}M Steps

1. Find no of clusters
2. Select cluster seeds
3. Assign non-seeds to seeds

Example

$$D = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{array} \begin{array}{c} \begin{matrix} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

α: Inverse of row sums
α_1= 1/3, α_2=   1/4, α_3=   1/3, α_4= 1/2, α_5= ½

β: inverse of column sums
β_1 = ¼, β_2 = 1/1, β_3 = ½, β_4 = ¼, β_5 = ½, β_6 = 1/3

m: no. of docs (5)
n: no. of terms (6)

$$S = \begin{bmatrix} 1/3 & 0 & 0 & 1/3 & 0 & 1/3 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \end{bmatrix} \qquad S' = \begin{bmatrix} 1/4 & 0 & 0 & 1/2 & 0 & 1/3 \\ 1/4 & 1 & 1/2 & 1/2 & 0 & 0 \\ 1/4 & 0 & 0 & 0 & 1/2 & 1/3 \\ 0 & 0 & 0 & 0 & 1/2 & 1/3 \\ 1/4 & 0 & 1/2 & 0 & 0 & 0 \end{bmatrix}$$

S for rows, S' for columns.

$C = S \times S'^T$ => $S_{mxn} \times S'^T_{nxm} = C_{mxm}$ *C:* cover coefficient matrix

d_1	(1/3) t_1	(1/4) d_1
		(1/4) d_2
		(1/4) d_3
		(1/4) d_5
	(1/3) t_4	(1/2) d_1
		(1/2) d_2
	(1/3) t_6	(1/3) d_1
		(1/3) d_3
		(1/3) d_4

Double Stage Probability experiment

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

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

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

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)

All docs are unique $\longleftrightarrow$ All docs are identical

$n_c\ =\ m$

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

$n_c\ =\ 1$

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

$n_c\!: no\ of\ clusters$

- $\min(c_{ii}) = 1/\mathrm{m}$
- 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.

$c_{ii} = decoupling\ coefficient = \delta_i$ (How this document is seperated from other documents)

$1 - c_{ii} = \Psi_i: coupling\ coefficient$

$n_c = \sum_{i=1}^{m} c_{ii} = \sum_{i=1}^{m} \delta_i$ avg. Decoupling $\delta = \sum \delta_i / m$ $n_c = m \times \delta$

Cover coefficient values for terms

$D \to C'_{nxn}$

$C' = S'^T \times S \Rightarrow S'^T_{nxm} \times S_{mxn} = C'_{nxn}$ C': cover coefficient matrix for terms

$$c'_{ij} = \beta_i \times \sum_{k=1}^{n} (d_{ki} \times \alpha_k \times d_{kj}) \qquad (1 \le i, j \le n)$$

Number of term clusters

$c'_{ii} = \delta'_i$

$n'_c = \sum_{i=1}^{n} C'_{ii} = \sum_{i=1}^{n} \delta'_i$, $\delta' = \sum \delta'_i / n$, $n'_c = n \times \delta'$

$$n_c = n'_c$$

$1 \le n_c, n'_c \le \min(m, n)$

Average cluster size

d_c : average no. of docs / cluster

$d_c = m / n_c$ $\frac{m}{\min(m,n)} \le d_c \le$ m → from terms point of view $\frac{n}{\min(m,n)} \le d'_c \le$ 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

$p_i = \delta_i \times \Psi_i \times X_{d_i} = \delta_i \times \Psi_i \times \alpha^{-1}$

X_{d_i}: no of unique terms in d_i (depth of indexing)

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.