

Additional SVD Examples

(for deeper insight into LSI's impact on IR)

Example Documents (From Bellcore)

- Human-Computer Interaction

c1:Human machine interface for Lab ABC computer applications

c2:A survey of user opinion of computer system response time

c3:The EPS user interface management system

c4:System and human system engineering testing of EPS

c5:Relation of user-perceived response time to error measurement

- Graphs

m1:The generation of random, binary, unordered trees

m2:The intersection graph of paths in trees

m3:Graph minors IV: Widths of trees and well-quasi-ordering

m4:Graph minors: A survey

Example Documents – Matrix Representation

Term-by-Document Matrix

c1	c2	c3	c4	c5	m1	m2	m3	m4	
1			1						human
1		1							interface
1	1								computer
	1	1		1					user
	1	1	2						system
	1			1					response
	1			1					time
		1	1						EPS
	1							1	survey
					1	1	1		trees
						1	1	1	graph
							1	1	minors

Initial Document-Document Similarity (original matrix)

$$\text{sim}(x, y) = \cos(x, y) = \frac{\sum_{i=1}^V x_i y_i}{|x| |y|}$$

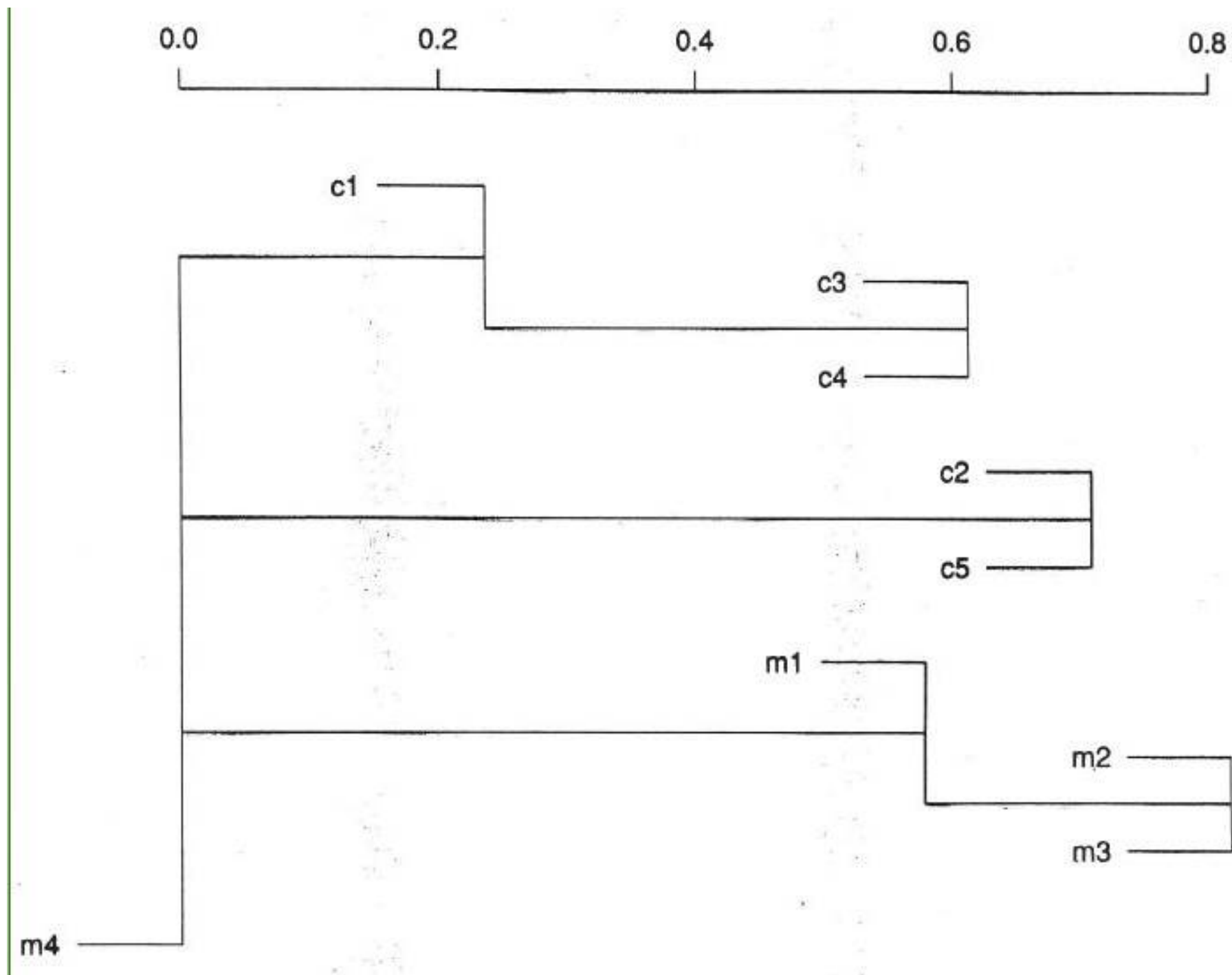
$$\cos(c1, c2) = \frac{1}{\sqrt{3} \times \sqrt{6}} = 0.2$$

	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1	1.0	0.2	0.3	0.2					
c2	0.2	1.0	0.4	0.3	0.7				0.2
c3	0.3	0.4	1.0	0.6	0.3				
c4	0.2	0.3	0.6	1.0					
c5		0.7	0.3		1.0				
m1						1.0	0.7	0.6	
m2						0.7	1.0	0.8	0.4
m3						0.6	0.8	1.0	0.7
m4		0.2					0.4	0.7	1.0

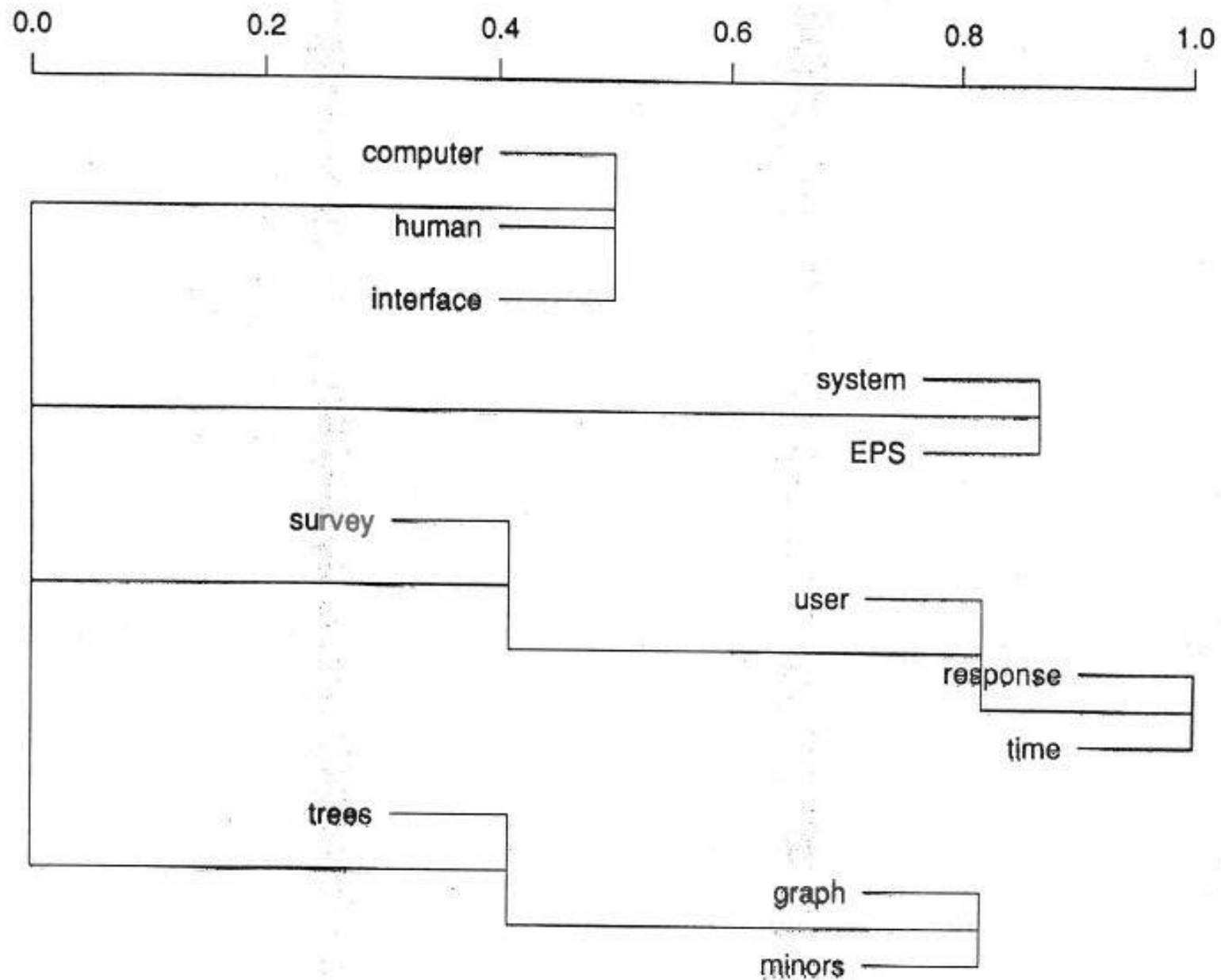
Initial Term-Term Cosines (original matrix)

human	interface	computer	user	system	response	time	EPS	survey	trees	graph	minors	
1.0	0.5	0.5		0.6			0.5					human
0.5	1.0	0.5	0.4	0.3			0.5					interface
0.5	0.5	1.0	0.4	0.3	0.5	0.5		0.5				computer
	0.4	0.4	1.0	0.5	0.8	0.8	0.4	0.4				user
0.6	0.3	0.3	0.5	1.0	0.3	0.3	0.9	0.3				system
		0.5	0.8	0.3	1.0	1.0		0.5				response
		0.5	0.8	0.3	1.0	1.0		0.5				time
0.5	0.5		0.4	0.9			1.0					EPS
		0.5	0.4	0.3	0.5	0.5		1.0		0.4	0.5	survey
									1.0	0.7	0.4	trees
								0.4	0.7	1.0	0.8	graph
								0.5	0.4	0.8	1.0	minors

Hierarchical Clustering of Documents (original matrix)



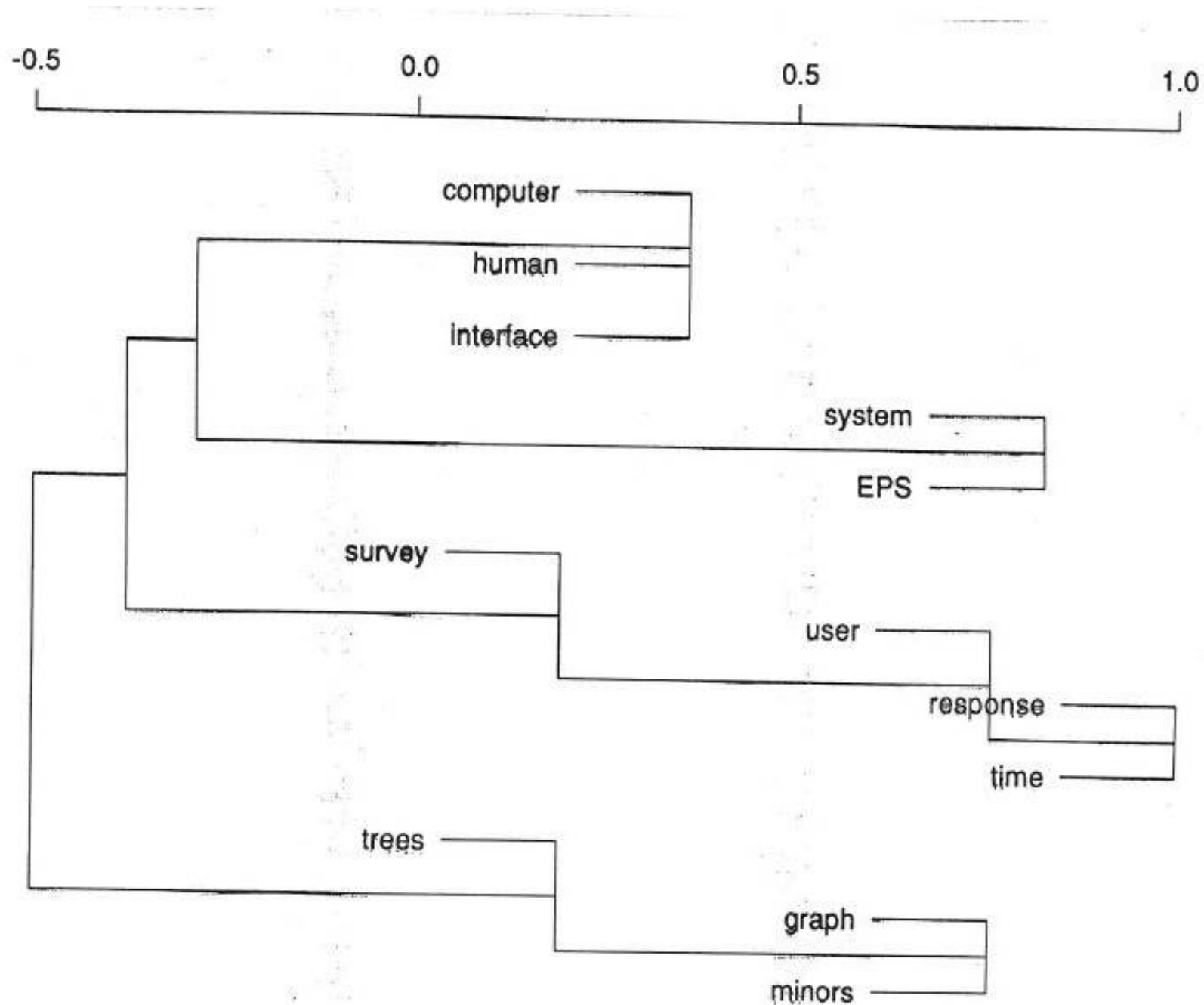
Hierarchical Clustering of Terms (original matrix)



Term-Term correlations

human	interface	computer	user	system	response	time	EPS	survey	trees	graph	minors	
1.0	0.4	0.4	-0.4	0.4	-0.3	-0.3	0.4	-0.3	-0.4	-0.4	-0.3	human
0.4	1.0	0.4	0.2	0.04	-0.3	-0.3	0.4	-0.3	-0.4	-0.4	-0.3	interfa
0.4	0.4	1.0	0.2	0.04	0.4	0.4	-0.3	0.4	-0.4	-0.4	-0.3	compu
-0.4	0.2	0.2	1.0	0.2	0.8	0.8	0.2	0.2	-0.5	-0.5	-0.4	user
0.4	0.04	0.04	0.2	1.0	0.04	0.04	0.8	0.04	-0.5	-0.5	-0.3	system
-0.3	-0.3	0.4	0.8	0.04	1.0	1.0	-0.3	0.4	-0.4	-0.4	-0.3	respon
-0.3	-0.3	0.4	0.8	0.04	1.0	1.0	-0.3	0.4	-0.4	-0.4	-0.3	time
0.4	0.4	-0.3	0.2	0.8	-0.3	-0.3	1.0	-0.3	-0.4	-0.4	-0.3	EPS
-0.3	-0.3	0.4	0.2	0.04	0.4	0.4	-0.3	1.0	-0.4	0.2	0.4	survey
-0.4	-0.4	-0.4	-0.5	-0.5	-0.4	-0.4	-0.4	-0.4	1.0	0.5	0.2	trees
-0.4	-0.4	-0.4	-0.5	-0.5	-0.4	-0.4	-0.4	0.2	0.5	1.0	0.8	graph
-0.3	-0.3	-0.3	-0.4	-0.4	-0.3	-0.3	-0.3	0.4	0.2	0.8	1.0	minors

Hierarchical Clustering of Term Correlation Coefficients



Hard clustering (of correlated terms)

“Automatic” Thesaurus

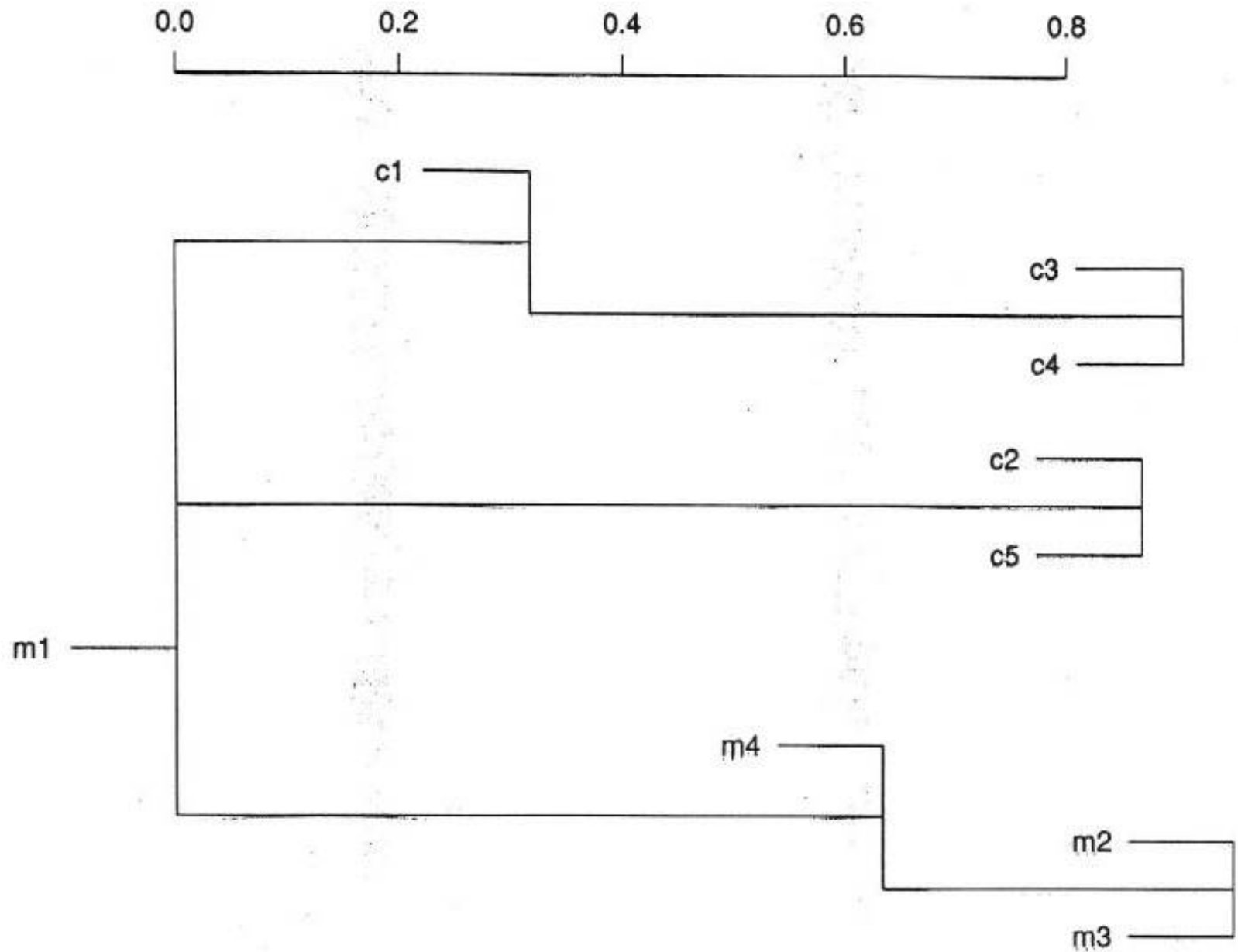
- Merge terms that cluster together
 - human/ interface/ computer
 - user/ response/ time
 - system/ EPS
 - graph/ minors

Term-Document Matrices before/after hard clustering

Before									
c1	c2	c3	c4	c5	m1	m2	m3	m4	
1			1						human
1		1							interface
1	1								computer
	1	1		1					user
	1	1	2						system
	1			1					response
	1			1					time
		1	1						EPS
	1							1	survey
					1	1	1		trees
						1	1	1	graph
							1	1	minors

After									
c1	c2	c3	c4	c5	m1	m2	m3	m4	
3	1	1	1						com/hum/inter
	3	1		3					user/res/time
	1	2	3						system/EPS
	1							1	survey
					1	1	1		trees
						1	2	2	graph/minors

Document-Document clustering after Hard Term Clustering



Document-Document Cosine Similarity after Term Clustering

Cosines after Term Reduction									
	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1	1.0	0.3	0.4	0.3					
c2	0.3	1.0	0.7	0.4	0.9				0.1
c3	0.4	0.7	1.0	0.9	0.4				
c4	0.3	0.4	0.9	1.0					
c5		0.9	0.4		1.0				
m1						1.0	0.7	0.4	
m1						0.7	1.0	0.9	0.6
m3						0.4	0.9	1.0	0.8
m4		0.1					0.6	0.8	1.0

Differences (before – after)									
	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1		-0.05	-0.1	-0.08					
c2	-0.05		-0.3	-0.03	-0.2				0.1
c3	-0.1	-0.3		-0.3	-0.1				
c4	-0.08	-0.03	-0.3						
c5		-0.2	-0.1						
m1								0.1	
m1								-0.1	-0.2
m3						0.1	-0.1		-0.1
m4		0.1					-0.2	-0.1	

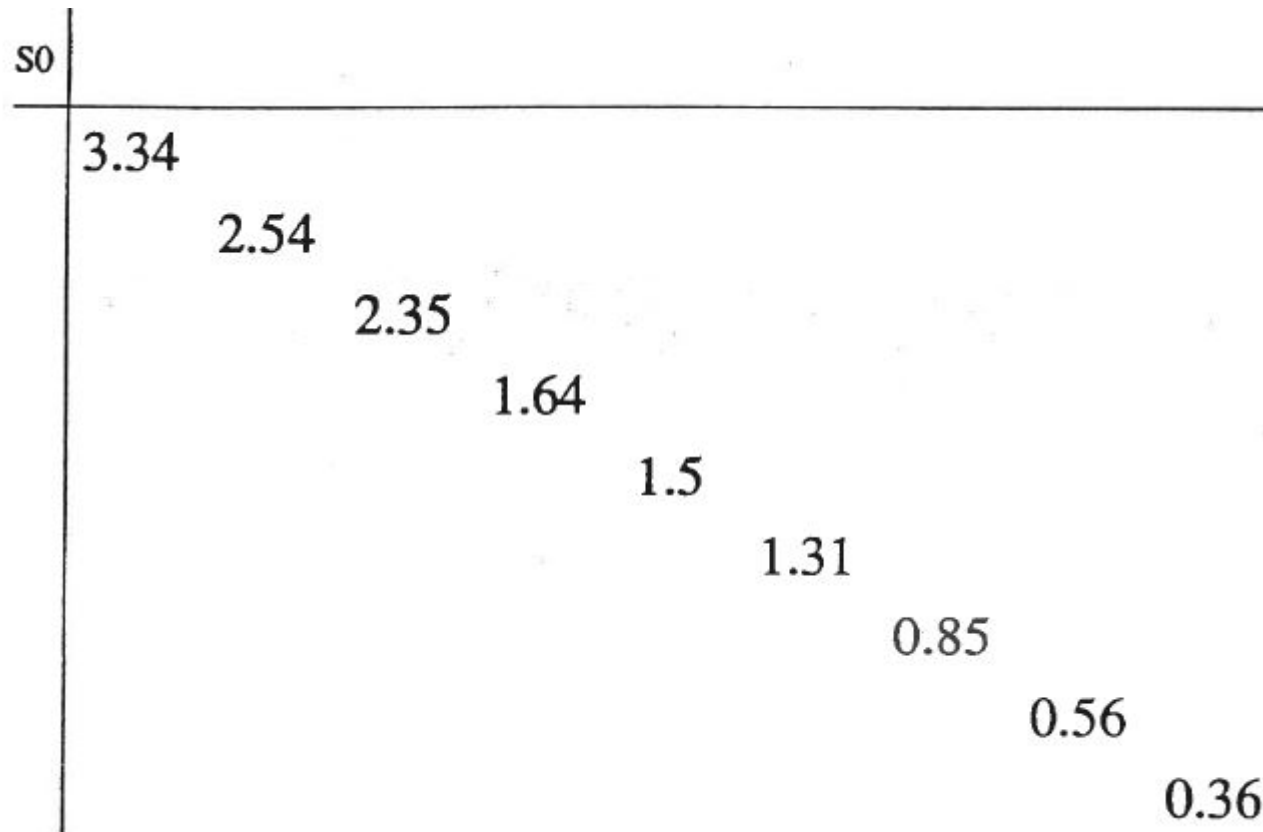
Singular Value Decomposition

- Let X be the term-by-doc matrix described above
- X can be decomposed into the product of three matrices, such that:
 - $X \approx T_0 S_0 D_0'$
 - T_0 and D_0 have orthogonal unit length cols so that $T_0 T_0' \approx I$ and $D_0 D_0' \approx I$
 - S_0 is a diagonal matrix.

SVD Matrix Decomposition

T0									
human	0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
interface	0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
computer	0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
user	0.40	0.06	-0.34	0.10	0.33	0.38			0.01
system	0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
response	0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
time	0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
EPS	0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
survey	0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
trees	0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
graph	0.04	0.62	0.22		-0.07	0.11	0.16	-0.68	0.23
minors	0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18

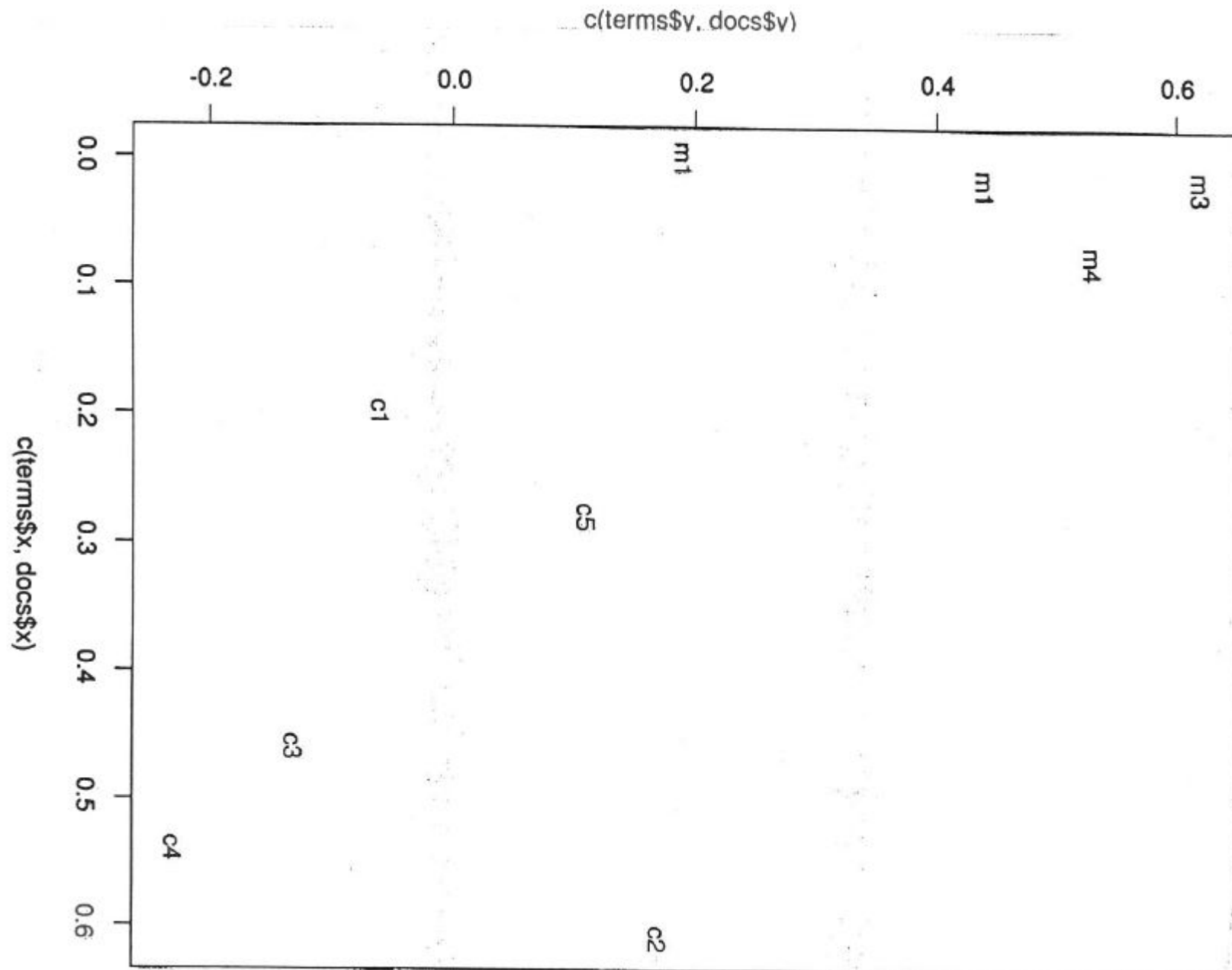
SVD Matrix Decomposition



SVD Matrix Decomposition

D0									
c1	0.20	-0.06	0.11	-0.95	0.05	0.08	0.11	-0.01	-0.06
c2	0.61	0.17	-0.50	-0.03	-0.21	-0.26	-0.43	0.05	0.24
c3	0.46	-0.13	0.21	0.04	0.38	0.72	-0.24	0.01	0.02
c4	0.54	-0.23	0.57	0.27	-0.21	-0.37	0.26	-0.02	-0.08
c5	0.28	0.11	-0.51	0.15	0.33	0.03	0.67	-0.06	-0.26
c6		0.19	0.10	0.02	0.39	-0.30	-0.34	0.45	-0.62
c7	0.01	0.44	0.19	0.02	0.35	-0.21	-0.15	-0.76	0.02
c8	0.02	0.62	0.25	0.01	0.15		0.25	0.45	0.52
c9	0.08	0.53	0.08	-0.03	-0.60	0.36	0.04	-0.07	-0.45

Two-Parameter representation of Documents



SVD Matrices before Truncation

T0									
human	0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
interface	0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
computer	0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
user	0.40	0.06	-0.34	0.10	0.33	0.38			0.01
system	0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
response	0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
time	0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
EPS	0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
survey	0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
trees	0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
graph	0.04	0.62	0.22		-0.07	0.11	0.16	-0.68	0.23
minors	0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18

S0									
	3.34								
		2.54							
			2.35						
				1.64					
					1.5				
						1.31			
							0.85		
								0.56	
									0.36

D0									
c1	0.20	-0.06	0.11	-0.95	0.05	0.08	0.11	-0.01	-0.06
c2	0.61	0.17	-0.50	-0.03	-0.21	-0.26	-0.43	0.05	0.24
c3	0.46	-0.13	0.21	0.04	0.38	0.72	-0.24	0.01	0.02
c4	0.54	-0.23	0.57	0.27	-0.21	-0.37	0.26	-0.02	-0.08
c5	0.28	0.11	-0.51	0.15	0.33	0.03	0.67	-0.06	-0.26
c6		0.19	0.10	0.02	0.39	-0.30	-0.34	0.45	-0.62
c7	0.01	0.44	0.19	0.02	0.35	-0.21	-0.15	-0.76	0.02
c8	0.02	0.62	0.25	0.01	0.15		0.25	0.45	0.52
c9	0.08	0.53	0.08	-0.03	-0.60	0.36	0.04	-0.07	-0.45

SVD Matrices after Truncation (2 singular values)

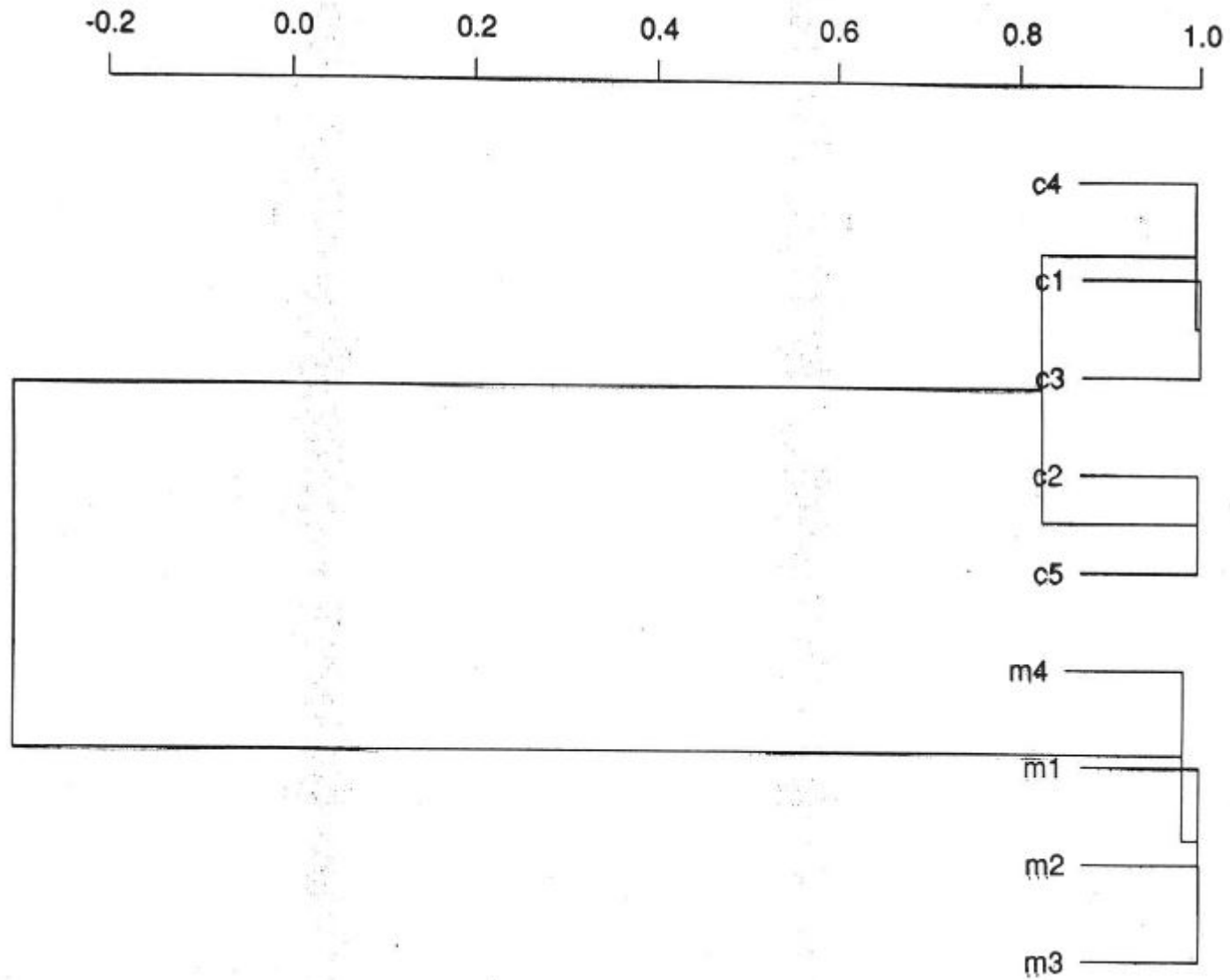
	T		S			D	
human	0.22	-0.11	3.34	0	c1	0.20	-0.06
interface	0.20	-0.07	0	2.54	c2	0.61	0.17
computer	0.24	0.04			c3	0.46	-0.13
user	0.40	0.06			c4	0.54	-0.23
system	0.64	-0.17			c5	0.28	0.11
response	0.27	0.11			c6		0.19
time	0.27	0.11			c7	0.01	0.44
EPS	0.30	-0.14			c8	0.02	0.62
survey	0.21	0.27			c9	0.08	0.53
trees	0.01	0.49					
graph	0.04	0.62					
minors	0.03	0.45					

Improved Document-Document Cosines after SVD

- Reconstitute $X \approx \hat{X} = T S D'$, and compute cosines for all pairs of documents.
- Note that the cosines now do a better job of separating the c docs from the m docs.

	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1	1.0	0.9	1.0	1.0	0.9	-0.2	-0.2	-0.2	-0.02
c2	0.9	1.0	0.9	0.9	1.0	0.2	0.2	0.3	0.4
c3	1.0	0.9	1.0	1.0	0.9	-0.2	-0.2	-0.2	-0.01
c4	1.0	0.9	1.0	1.0	0.8	-0.3	-0.3	-0.3	-0.1
c5	0.9	1.0	0.9	0.8	1.0	0.3	0.3	0.3	0.5
m1	-0.2	0.2	-0.2	-0.3	0.3	1.0	1.0	1.0	1.0
m2	-0.2	0.2	-0.2	-0.3	0.3	1.0	1.0	1.0	1.0
m3	-0.2	0.3	-0.2	-0.3	0.3	1.0	1.0	1.0	1.0
m4	-0.02	0.4	-0.01	-0.1	0.5	1.0	1.0	1.0	1.0

Document-Document Clustering after SVD



Improved Document-Document Cosines after SVD

- Reconstitute $X \approx \hat{X} = T S D'$, and compute cosines for all pairs of documents.
- Note that the cosines now do a better job of separating the c docs from the m docs.

	c1	c2	c3	c4	c5	m1	m2	m3	m4
c1	1.0	0.9	1.0	1.0	0.9	-0.2	-0.2	-0.2	-0.02
c2	0.9	1.0	0.9	0.9	1.0	0.2	0.2	0.3	0.4
c3	1.0	0.9	1.0	1.0	0.9	-0.2	-0.2	-0.2	-0.01
c4	1.0	0.9	1.0	1.0	0.8	-0.3	-0.3	-0.3	-0.1
c5	0.9	1.0	0.9	0.8	1.0	0.3	0.3	0.3	0.5
m1	-0.2	0.2	-0.2	-0.3	0.3	1.0	1.0	1.0	1.0
m2	-0.2	0.2	-0.2	-0.3	0.3	1.0	1.0	1.0	1.0
m3	-0.2	0.3	-0.2	-0.3	0.3	1.0	1.0	1.0	1.0
m4	-0.02	0.4	-0.01	-0.1	0.5	1.0	1.0	1.0	1.0