Theory of LSI

老師：陳柏琳教授
報告人：郭榮芳
時間：12/26/2003

參考文獻

1. 程雋(2003)，應用線性代數，台北：文笙書局
2. Landauer, T. K., Foltz, P. W., Laham, D. (1998). An introduction to Latent Semantic Analysis . *Discourse Processes* , 25, 259-284.
3. G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. *Information retrieval using a singular value decomposition model of latent semantic structure*. In Proceedings of the Eleventh International Conference on Research & Development in Information Retrieval, pages 465--480, 1988.
4. M. W. Berry, S. T. Dumais, and G. W O'Brien. *Using linear algebra for intelligent information retrieval*. SIAM Review, 37(4):177-196, 1995.

潛在語意分析(latent semantic analysis)

1. 潛在語意分析((latent semantic analysis,LSA)是以線性代數中的奇異值分解(singular value decomposition,SVD)為基礎的模組，能夠將文字和文章投影到一個向量空間。
2. SVD是是一種矩陣的分解方法，經過SVD的轉換，可以將一個矩陣X分解成三個矩陣T_m、S_m、D_m，且$X=T_mS_mD_m$，T_m、S_m為正交(orthonormal)矩陣，S_m為對角(diagonal)矩陣。

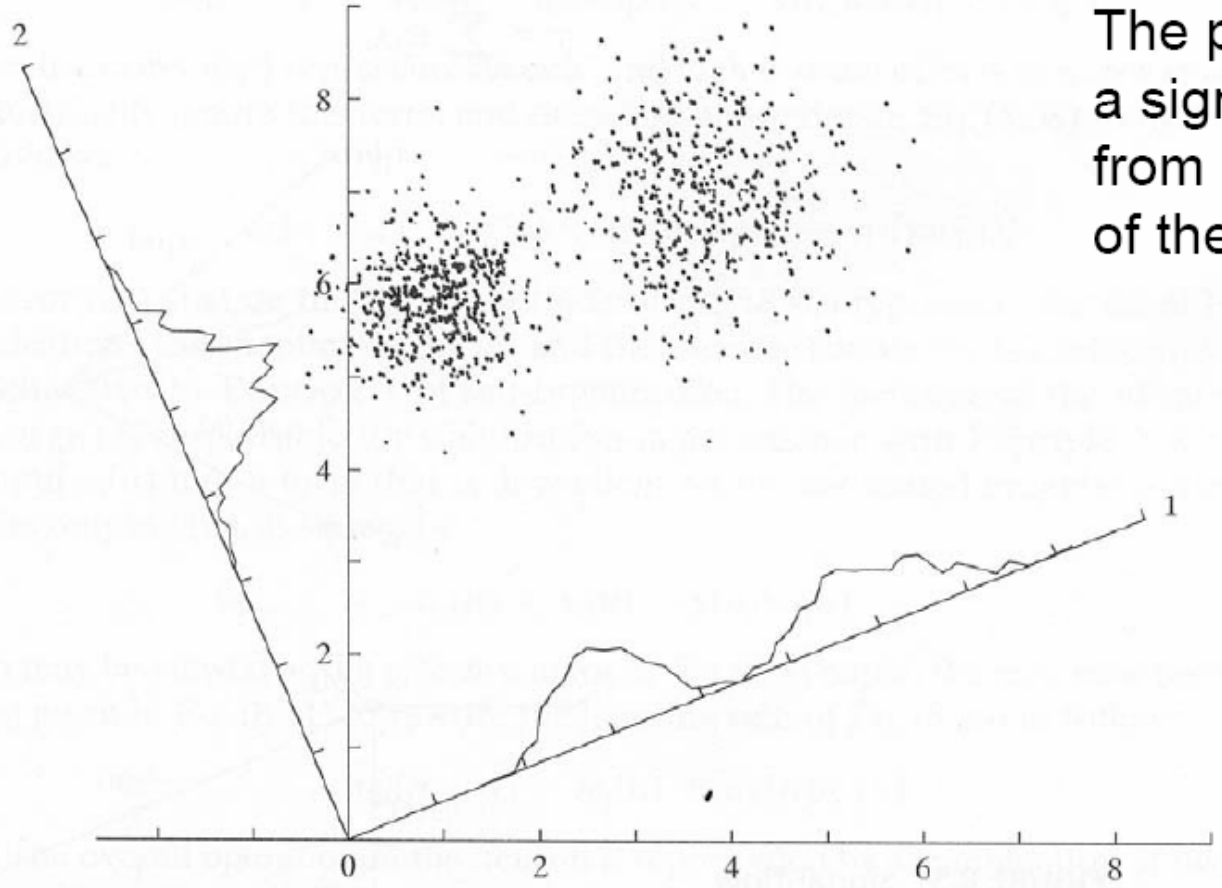


FIGURE 8.4 A cloud of data points is shown in two dimensions, and the density plots formed by projecting this cloud onto each of two axes, 1 and 2, are indicated. The projection onto axis 1 has maximum variance, and clearly shows the bimodal, or clustered character of the data.

The patterns show a significant difference from each other in one of the transformed axes

奇異值分解

1. 奇異值分解為一種基底變換，其基底為正交單範基底:
   1) 定義:對佈於內積空間V_F的一組向量{u1, u2, ..., un}而言，若其具有以下性質，則稱此集合為正交單範集（orthnomal set）：

$$\forall i, j \in 1 \sim n, < ui, uj >= \begin{cases} 0 \; if \; i \neq j \\ 1 \; if \; i = j \end{cases}$$

奇異值分解(續)

1. 將矩陣 $A \in C^{m \times n}$ 分解成 $U \sum V^{-1}$ 的形式，其中 U 與 V 都是么正矩陣（unitary matrix），因此 $V^{-1} = V^{H}$ ，且希望

$$\sum = [\sigma_{ij}]_{m \times n} \text{且} \sigma_{ij} = 0 \; if \; i \neq j$$

方法

因為 $AA^H \in C^{m\times m}$ 為正規(normal)矩陣，因此必存在么正矩陣 $U \in C^{m\times m}$ 而使得下式成立：

$$UU^H = U^H U = I$$

$$U^H(AA^H)U = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_m^2 \end{bmatrix} = D^2; \sigma_1 \sim \sigma_m \in R, \sigma_1 \sim \sigma_m \geq 0$$

$$\Rightarrow AA^H = UDDU^H = (UD)(UD)^H; D \in R^{m\times m}$$

方法（續）

1. 因為U為非奇異(nonsingular)方陣，所以恆有$r(AA^H)=r(D^2)$ 的性質，且因為 $r(AA^H)=r(A)$ 所以 $r(D^2)=r(A)$ 。
2. 設$r(A)=k$，則在 D^2 的對角線元素中，有k個不等於0，m-k個為零。將非零的元素排在 $\sigma_1 \sim \sigma_k$
3. 因為 $UD \in C^{m\times m}, A \in C^{m\times n}$ ，所以事實上A和UD並不相等，故須引入另一個么正矩陣V及一個特殊矩陣Σ。

方法（續）

$$AA^H = UDDU^H = UDD^HU^H \quad (D = D^H)$$

$$= U\Sigma\Sigma^H U^H = U\Sigma V^H V\Sigma^H U^H$$

$$(DD^H = \Sigma\Sigma^H, V^H V = I)$$

$$= (U\Sigma V^H)(U\Sigma V^H)^H$$

方法（續）

$$A = U \sum V^H$$

$$\Rightarrow A^H A = V \sum^H U^H U \sum V^H = V(\sum^H \sum) V^H$$

$$\Rightarrow V^H (A^H A) V = \sum^H \sum$$

$$Ax = \lambda x$$

由於 $V^H = V^{-1}, \sum^{\mathrm{H}} \sum$ 為對角矩陣，因而知道V的行向量是 $A^H A$ 的特徵向量。

方法（續）

$$A = U \sum V^H$$

$$\Rightarrow AA^H = U \sum V^H V \sum^H U^H = U(\sum^H \sum) U^H$$

$$\Rightarrow U^H (A^H A) U = \sum^H \sum$$

$$Ax = \lambda x$$

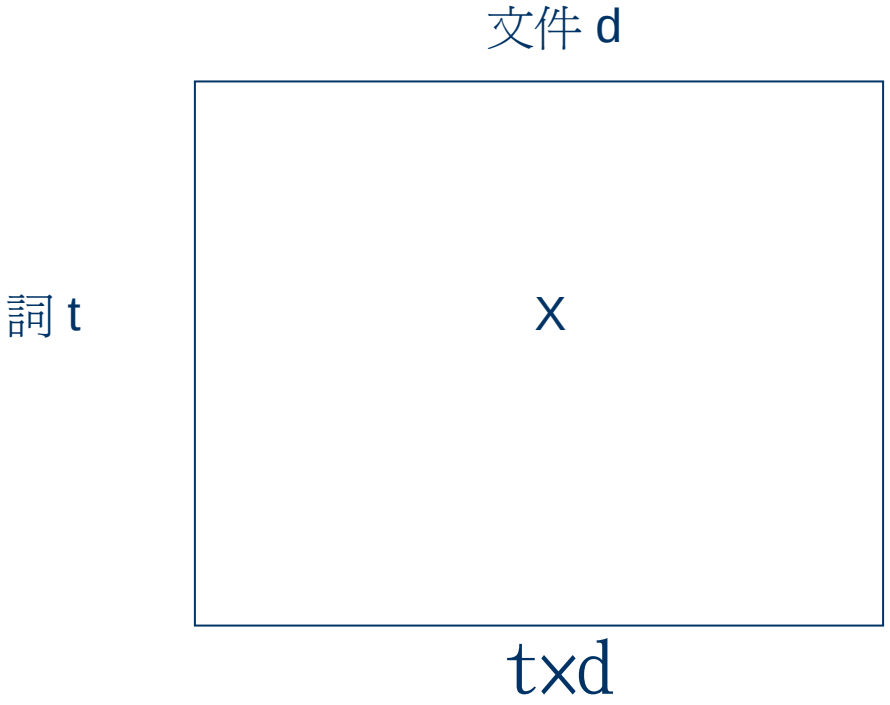
由於 $U^H = U^{-1}, \sum^{\mathrm{H}} \sum$ 為對角矩陣，因而知道U的行向量是 AA^H 的特徵向量。

定理

1. 已知$A \in C^{m \times n}$可被奇異值分解成$U \Sigma V^H$ ，若r(A)=k，則$\Sigma = [\sigma_{ij}]$中恰有k個σ_{ij} 不為零，將其零元素至於非零元素之右側，則在取$U = [u^1\ u^2 ... u^m], V = [v^1 v^2 ... v^n\ ]$後，U之行向量為$AA^H$ 的特徵向量，V之行向量為$A^H A$ 之特徵向量。

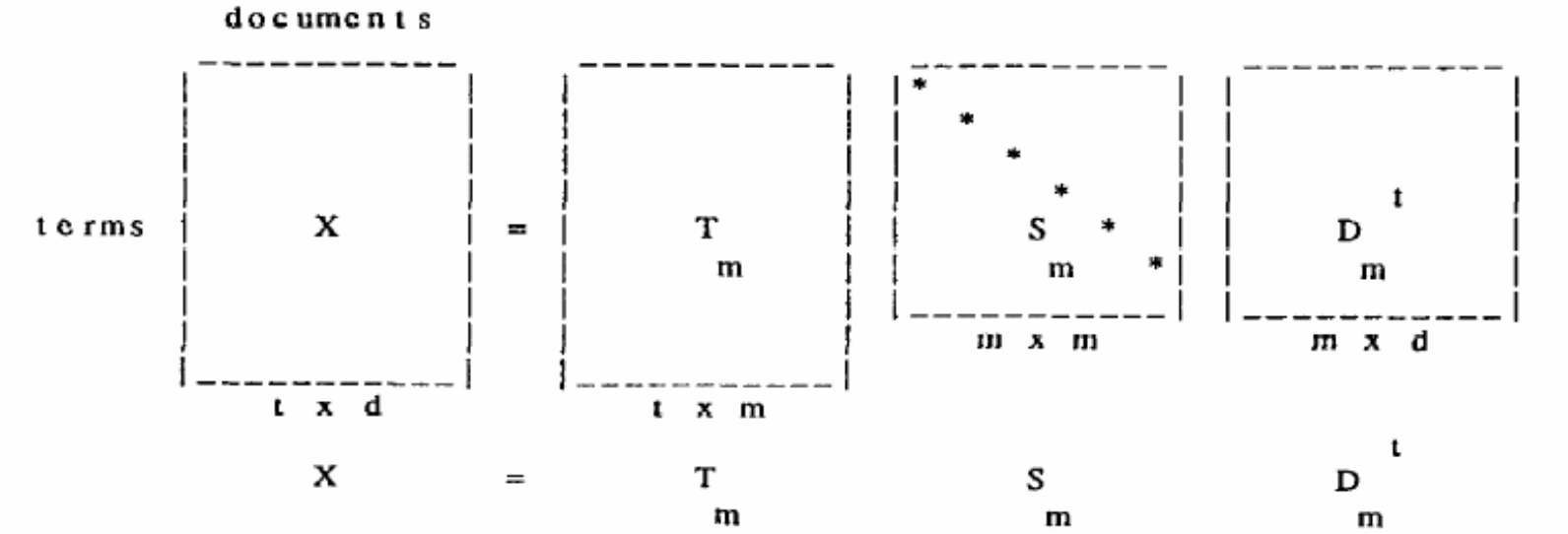
LSA的進行步驟

將文件和詞的關係表示成一個txd的矩陣X，矩陣X的每個元素x表示詞t在文件d出現的次數。



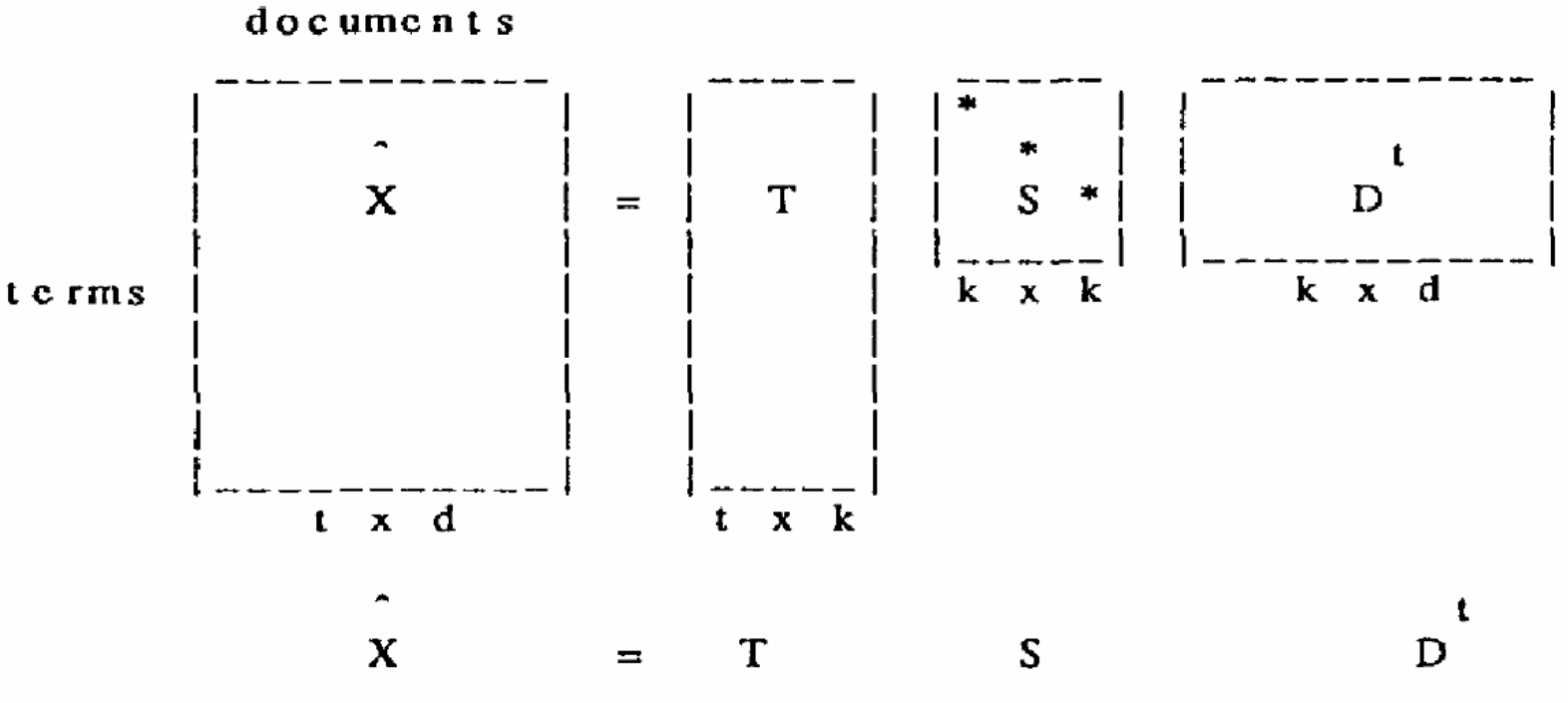
LSA的進行步驟(cont.)

1. 將矩陣X經過SVD分解後的到三個矩陣的連乘積，$X=T_m S_m D_m^t$ ，其中S為一對角矩陣，代表語意空間(semantic space)，矩陣T_m的某一列向量(row vector)為該詞在語意空間的表示法，矩陣D_m 的某一列向量為該文件在語意空間的表示法，m為矩陣X的rank。



LSA的進行步驟(cont.)

1. Reduced singular value decomposition.
2. $X' = T_k S_k D_k^t \simeq X$



範例

有九篇文件如下：

Technical Memo Example

Titles:

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

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*

範例(cont.)
計算每個詞在文件中出現的次數

Terms	Documents								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

r (human.user) = -.38

r (human.minors) = -.29

範例(cont.)

做SVD分解得到三個矩陣Tm、Sm、Dm

Tm=

0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	-0.68	0.23
0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18

範例(cont.)

Sm=

3.34								
	2.54							
		2.35						
			1.64					
				1.50				
					1.31			
						0.85		
							0.56	
								0.36

Dm

0.20	-0.06	0.11	-0.95	0.05	-0.08	0.18	-0.01	-0.06
0.61	0.17	-0.50	-0.03	-0.21	-0.26	-0.43	0.05	0.24
0.46	-0.13	0.21	0.04	0.38	0.72	-0.24	0.01	0.02
0.54	-0.23	0.57	0.27	-0.21	-0.37	0.26	-0.02	-0.08
0.28	0.11	-0.51	0.15	0.33	0.03	0.67	-0.06	-0.26
0.00	0.19	0.10	0.02	0.39	-0.30	-0.34	0.45	-0.62
0.01	0.44	0.19	0.02	0.35	-0.21	-0.15	-0.76	0.02
0.02	0.62	0.25	0.01	0.15	0.00	0.25	0.45	0.52
0.08	0.53	0.08	-0.03	-0.60	0.36	0.04	-0.07	-0.45

範例(cont.)

新維度設為2

T

0.22	-0.11
0.20	-0.07
0.24	0.04
0.40	0.06
0.64	-0.17
0.27	0.11
0.27	0.11
0.30	-0.14
0.21	0.27
0.01	0.49
0.04	0.62
0.03	0.45

S

3.34	
	2.54

D^t

0.20	0.61	0.46	0.54	0.28	0.00	0.02	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53

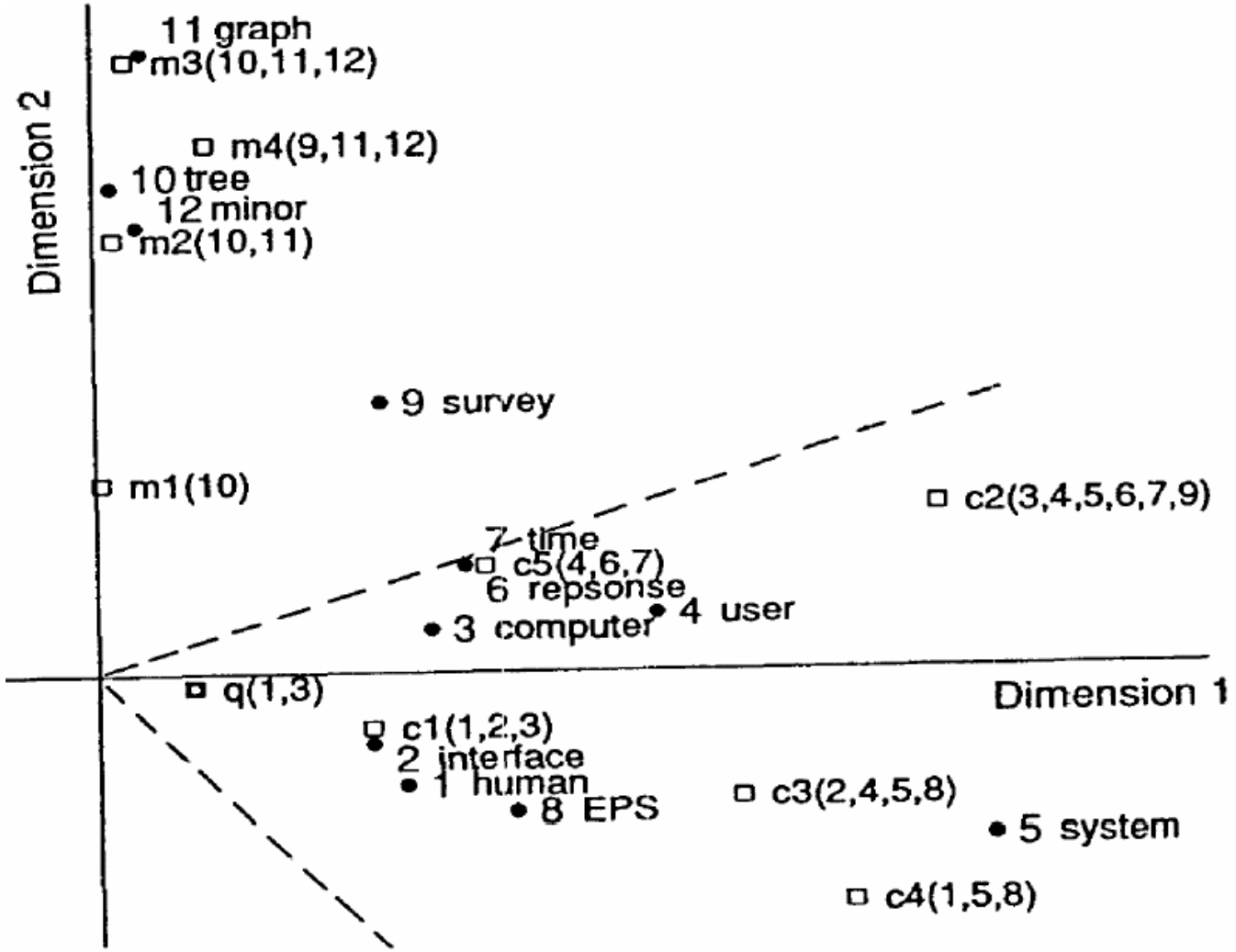


Figure 5. A 2-dimensional plot of 12 Terms and 9 Documents from the example set. Terms are represented by filled circles. Documents are shown as open squares, and component terms are indicated parenthetically. The query ("human computer interaction") is represented as a pseudo-document at point q. Axes are appropriately scaled for Document-Document or Term-Term comparisons The dotted cone contains all points within a cosine of .9 from the query q. All documents about human-computer (c1-c5) are within this cone, but none of the graph theory documents (m1-m4) are nearby. In this reduced space, even documents c3 and c5, which share no terms with the query, are very close to the query direction.

範例(cont.)

相乘後得到新矩陣X'

X'=

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

r (human.user) = .94

r (human.minors) = -.83

比較做LSA前後的兩個矩陣X和X'，可發現某些詞並未在文件中出現，但經LSA後分數明顯增加(例如tree)。

X=

Terms	Documents								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

X'=

human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

Correlations between titles in raw data:

	c1	c2	c3	c4	c5	m1	m2	m3
c2	-0.19							
c3	0.00	0.00						
c4	0.00	0.00	0.47					
c5	-0.33	0.58	0.00	-0.31				
m1	-0.17	-0.30	-0.21	-0.16	-0.17			
m2	-0.26	-0.45	-0.32	-0.24	-0.26	0.67		
m3	-0.33	-0.58	-0.41	-0.31	-0.33	0.52	0.77	
m4	-0.33	-0.19	-0.41	-0.31	-0.33	-0.17	0.26	0.56

Correlations in two dimensional space:

c2	0.91							
c3	1.00	0.91						
c4	1.00	0.88	1.00					
c5	0.85	0.99	0.85	0.81				
m1	-0.85	-0.56	-0.85	-0.88	-0.45			
m2	-0.85	-0.56	-0.85	-0.88	-0.44	1.00		
m3	-0.85	-0.56	-0.85	-0.88	-0.44	1.00	1.00	
m4	-0.81	-0.50	-0.81	-0.84	-0.37	1.00	1.00	1.00

加入新文件的方法

1. 重新計算
2. fold-in

$$t_i = T_i S D^t$$

$$\Rightarrow (t_i)^t = (T_i S D^t)^t \Rightarrow t_i^{\,t} = D S^t T_i^{\,t}, \quad (S^t = S)$$

$$\Rightarrow S^{-1} D^t t_i^{\,t} = T_i^{\,t} \Rightarrow (S^{-1} D^t t_i^{\,t})^t = (T_i^{\,t})^t \Rightarrow T_i = t_i D S^{-1}$$

$$d_j = T S D_j^{\,t}$$

$$\Rightarrow (d_j)^t = (T S D_j^{\,t})^t \Rightarrow d_j^{\,t} = D_j S^t T^t, \quad (S^t = S)$$

$$\Rightarrow d_j^{\,t} T S^{-1} = D_j$$

Fold-in(cont)

1. Two step
2. Step 1 : Fold-In term vector

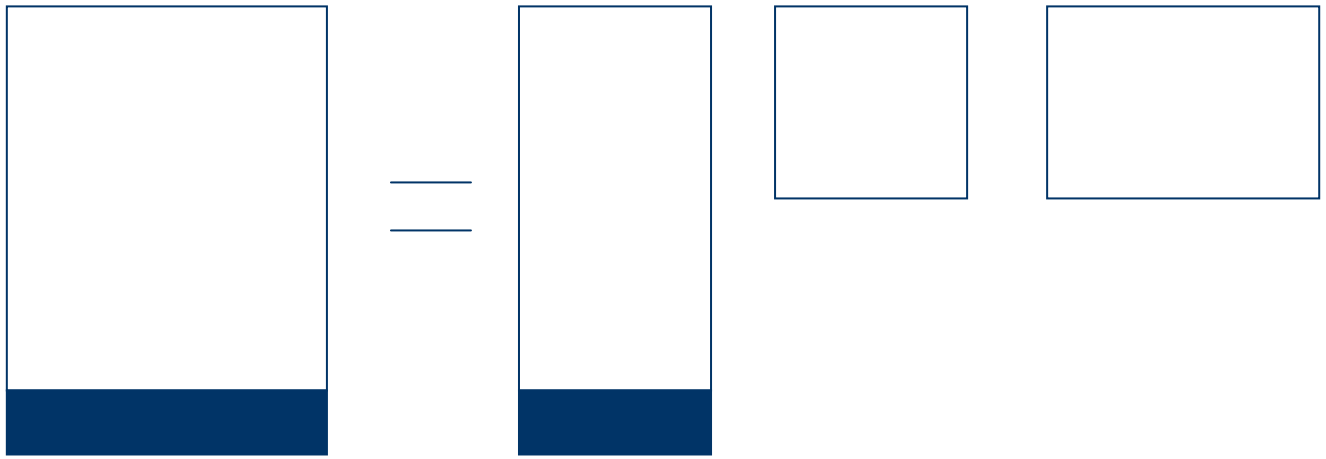
(1) Compute new term vector(D) using

$$T_i = t * D_k * S_k^{-1}$$

where t is term vector to be added

(2) Append T_i to the rows of T_k

Fold-in(cont)



X T_k S_k D_k^T

Fold-in

Step2:Fold-In document vector

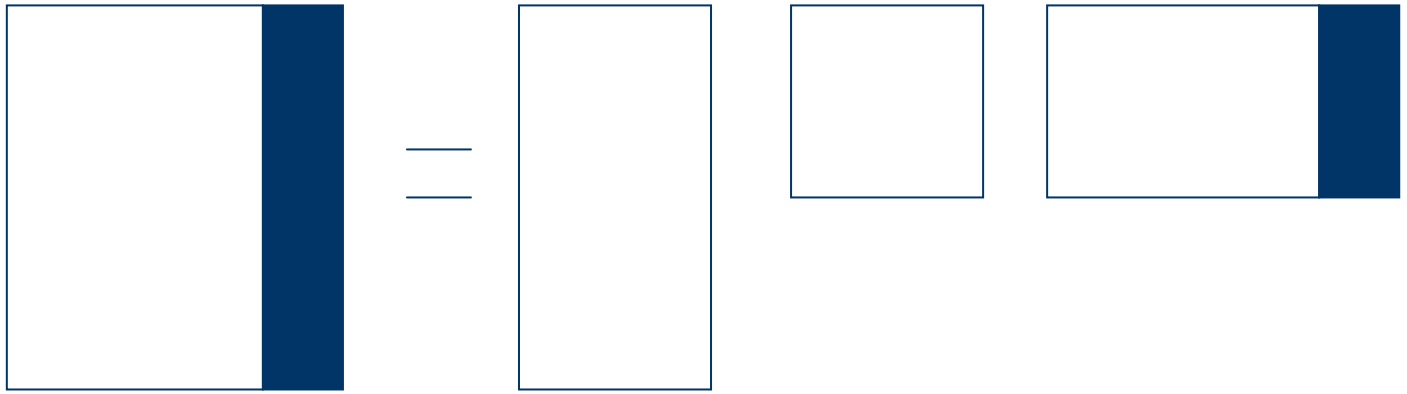
(1) Compute new document vector(T)

using $D_j = d^T * T_k * S_k^{-1}$

where d is document vector to be added

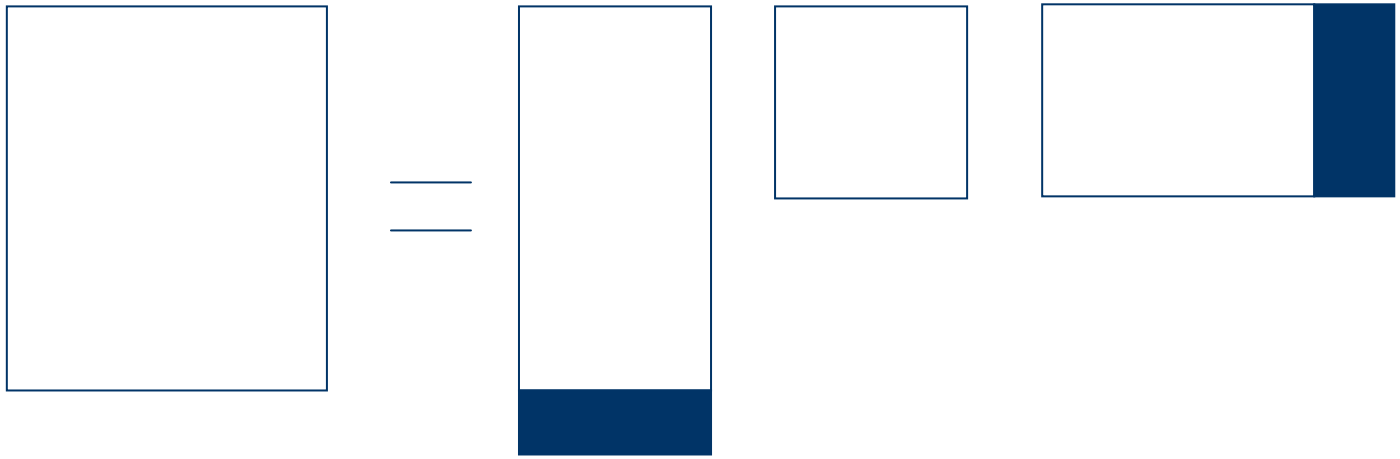
(2)Append D_j to the columns of D_k

Fold-in(cont)



x T_k S_k D_k^T

Fold-in(cont)



X_{new} T_k S_k D_k^T