第二章　文獻探討

本研究牽涉之理論，包含奇異值分解、潛在語意分析，以及中文斷詞，本章將依序說明之。

第一節 奇異值分解

LSA 係以奇異值分解(singular value decomposition, SVD)和維度約化(dimension reduction) 為理論基礎，SVD 是一種數學矩陣的分解技術，能將文件中所隱含的知識轉換到一個向量空間中；而維度約化能去除文件在此向量空間中的雜訊，如此 LSA 能較精確的表示文件中所隱含的資訊。SVD 可以將一個高維度的資料，經過轉換後降到一個 50-150 大小的維度空間 (Berry, Dumais, & 'Brien,1995)，並且將原矩陣 X 分解成三個矩陣 T、S、D，且 $X = TSD^t$ ，S 為奇異值矩陣 (singular value matrix)，代表語意空間(semantic space)，T、D 為正交(orthonormal)矩陣，矩陣 T 的某一行向量為該詞彙(term)在語意空間的表示法，矩陣 D 的某一列向量為該文件(document)在語意空間的表示法[t]。以下敘述其分解的原理：

奇異值分解為一種基底變換，其基底為正交單位基底 (orthonormal basis)，正交單位的定義為：

對佈於內積空間 V_F 的一組向量$\{u_1, u_2, \cdots, u_n\}$而言，若其具有以下性質，則稱此集合為正交單位集（orthonormal set）：

$$\forall i, j \in 1 \sim n, < u_i, u_j >= \begin{cases} 0 \text{ if } i \neq j \\ 1 \text{ if } i = j \end{cases} \cdots\cdots（1）$$

若將矩陣 $X \in C^{m\times n}$ 分解成 $X = TSD^H$ 的形式，其中 T 與 D 都是單式矩陣（unitary matrix)，即 $TT^H = T^H T = I, DD^H = D^H D = I$，且 S 為奇異值矩陣，奇異值矩陣的定義如下（廖亦德，2001）：

若 $m{\times}n$ 實數矩陣 S 滿足下列條件：

（1）$S = [\sigma_{ij}]_{m\times n}$ 且 $\sigma_{ij} = 0\ if\ i \neq j$ ……（2）

（2）且 $\exists r \leq \min\{m, n\}$，使得 $\begin{cases} \sigma_{ii} > 0\ if\ i \leq r \\ \sigma_{ii} = 0\ if\ i > r \end{cases}$ ……（3）

則稱 S 為奇異值矩陣。

若 $X \in R^{m\times n}$，則 T 與 D 可取為實數正交矩陣。

瞭解由矩陣 X 分解成的三個矩陣 T_m、S_m、D_m，其矩陣的性質後，接下來是分析這種分解是否存在，如果存在，如何經由矩陣 X 計算矩陣 T、S、D(廖亦德，2001；陳雋，2003)：

1. 對任何 $X \in C^{m\times n}$，$D^H D$ 恆為佈於 $C^{n\times n}$ 的 Hermitian 矩陣，因此必存在單式矩陣 D，使右式成立：$D^H X^H XD = diag(\mu_1, \mu_2, ..., \mu_n)$ ……（4）

因為D、D^H為非奇異（nonsingular）方陣，所以

$$rank(D^H X^H XD) = rank(X^H X) = rank(diag(\mu_1, \mu_2, ..., \mu_n)) \cdots\cdots\cdots\cdots\cdots\cdots (5)$$

又因為對於任意矩陣X，恆有$rank(X^H X) = rank(X)$，故可知

$$rank(diag(\mu_1, \mu_2, ..., \mu_n)) = rank(X) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (6)$$

若$rank(X) = r$，則有r個μ_i不為0，適當調整排列方式，可使

$$\mu_1, \mu_2, ..., \mu_r > 0,\ \mu_{r+1}, \mu_{r+2}, ..., \mu_n = 0 \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (7)$$

2. 令σ_i為μ_i的正平方根，

取 S=$\begin{bmatrix} \Sigma & O_{r\times(n-r)} \\ O_{(m-r)\times r} & O_{(m-r)\times(n-r)} \end{bmatrix}$，其中$\Sigma = diag(\sigma_1, \sigma_2, ..., \sigma_r)$，$O$為零矩陣。……（8）

3. 令$XD = [v_1, v_2, ..., v_n]_{m\times n} = [\ B\ \ |\ \ C\ ]$, B為 mxr 矩陣，C為 mx$(n-r)$矩陣。……（9）

由(4)式知 $D^H X^H XD = ((D^H X^H)^H)^H (XD) = (XD)^H (XD) = diag(\mu_1, \mu_2, ..., \mu_n)$

因為σ_i為μ_i的正平方根，所以 $diag(\mu_1, \mu_2, ..., \mu_n) = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2)$。

故 $(XD)^H (XD) = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2)$ ……（10）

（9）代入（10）可得$v_i^H v_i = \sigma_i^2$

且若$i \neq j$, 則$v_i^H v_j = 0$……（11）

由（8）式可知C為0矩陣（C=0）。……（12）

對 $i=1, 2, \cdots, r$ ，令 $t_i = \sigma_i^{-1} v_i$

則 $t_1, t_2, ..., t_r$ 形成正交單位集……（13）

將這 r 個行向量擴充為 $C^{m\times 1}$ 的正交單位基底 $\{t_1, t_2, ..., t_m\}$，

然後將 $t_1, t_2, ..., t_m$ 排成 $m\times m$ 的單式矩陣 T 。……（14）

4. 由(8)與(14)兩式可知

$$TS = \begin{bmatrix} t_1 & t_2 & \cdots & t_m \end{bmatrix}_{m\times m} \begin{bmatrix} \Sigma & O_{r\times(n-r)} \\ O_{(m-r)\times r} & O_{(m-r)\times(n-r)} \end{bmatrix}_{m\times n}$$

$$= \begin{bmatrix} t_1 & t_2 & \cdots & t_r \end{bmatrix}_{m\times r} \begin{bmatrix} \Sigma & O_{r\times(n-r)} \end{bmatrix}_{r\times n}$$

$$= \begin{bmatrix} \sigma_1 t_1 & \sigma_2 t_2 & \cdots & \sigma_r t_r \end{bmatrix} O_{m\times(n-r)} \Big]_{m\times n} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (15)$$

由(13)式知 $t_i = \sigma_i^{-1} v_i \Rightarrow \sigma_i t_i = v_i$ ……(16)

由(9)與(12)兩式可知

$XD = [v_1, v_2, ..., v_n]_{m\times n} = [\, B \;\; | \;\; C \,]$,

其中 B 為 $m\times r$ 矩陣，C 為 $m\times(n-r)$ 矩陣，且 $C=0$，……(17)

故 $XD = [\sigma_1 d_1, \sigma_2 d_2, ..., \sigma_r d_r \;|\; O_{m\times(n-r)}]_{m\times n}$ ……(18)

由(15)與(18)兩式知 $TS=XD$，$\Rightarrow TSD^H = XDD^H \Rightarrow TSD^H = X$,

故知 $X = TSD^H$ ……(19)

由前述奇異值分解，設奇異值依序為$\sigma_1,\sigma_2,...,\sigma_{\min\{m,n\}}$，前$r$個大於零，其他的為零，並設$T$的行是$t_1,t_2,...,t_m$，$D$的行是$d_1,d_2,...,d_n$，則分解後的矩陣有以下的性質：

(a) 因為 T，D為單式矩陣，所以 T，D可逆

故 $rankX = rank(TSD^H) = rankS = r$ ……(20)

(b)X 的非零奇異值是 $X^H X$ 的非零特徵值的正平方根，證明如下：

$X = TSD^H$

$\Rightarrow X^H X = (TSD^H)^H (TSD^H) = DS^H T^H TSD^H = DS^H SD^H$

$$= D\begin{bmatrix}\Sigma & O\\ O & O\end{bmatrix}^H\begin{bmatrix}\Sigma & O\\ O & O\end{bmatrix}D^H = D\begin{bmatrix}\Sigma^2 & O\\ O & O\end{bmatrix}D^H \quad \cdots\cdots(21)$$

又 $\det(X^H X - xI_n) = \det\left\{\begin{bmatrix}\Sigma^2 & O\\ O & O\end{bmatrix} - xI_n\right\}$

$= (\sigma_1^2 - x)(\sigma_2^2 - x)...(\sigma_r^2 - x)(-x)^{n-r}$ ……(22)

故得證。

(c) 由 $X^H X = (TSD^H)^H (TSD^H) = DS^H T^H TSD^H = DS^H SD^H = DS^H SD^{-1}$，

即知 d_i 是 $X^H X$ 的特徵向量（eigenvector）。……(23)

由 $XX^H = (TSD^H)(TSD^H)^H = TSD^H DS^H T^H = TSS^H T^H = TSS^H T^{-1}$，

即知 t_i 是 XX^H 的特徵向量。 ……(24)

第二節 潛在語意分析

瞭解奇異值分解的原理後，以下討論如何利用奇異值分解做潛在語意分析：

潛在語意分析(latent semantic analysis,LSA)的運作流程如下(Furnas et al., 1988)：

1.將文件集裡所有的文件表示成一個 m×n 的矩陣 X，矩陣 X 的每個元素 x 表示詞彙 t_i 在文件 d_i 中的重要性，如圖 2-1。

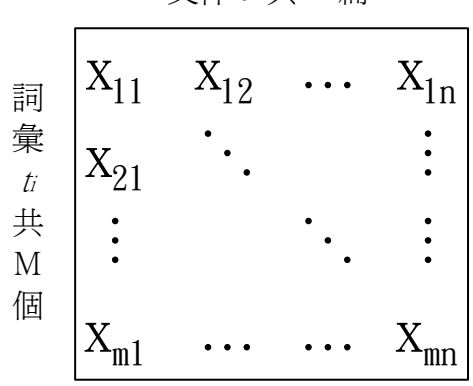

圖 2-1 詞彙與文件間的關係所形成的矩陣 X

2. 將矩陣 X 經過 SVD 分解後的到三個矩陣的連乘積，$X = T_r S_r D_r^t$，分解後如圖 2-2，其中 S_r 為一奇異值矩陣，代表語意空間(semantic space)，矩陣 T_r 的某一列向量(row vector)為該詞在語意空間的表示法，矩陣 D_r 的某一列向量為該文件在語意空間的表示法， m 為詞彙的數量，n 為文件的數量，r 為矩陣 X 的 rank。

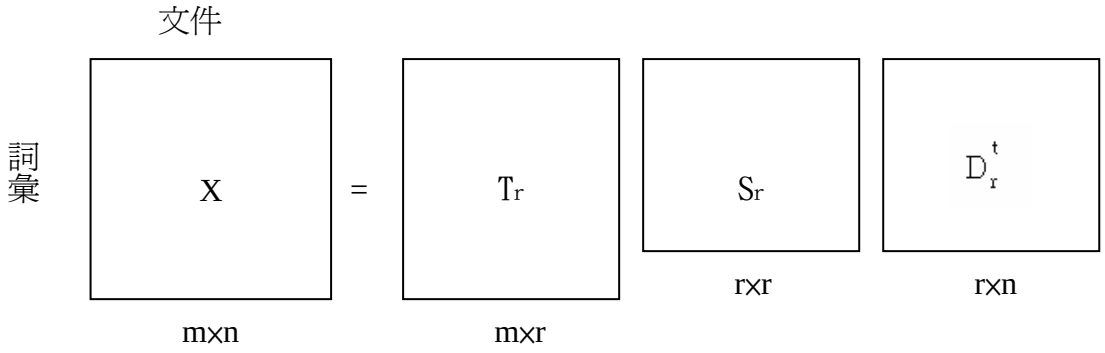

圖 2-2 詞彙與文件之關係所形成的矩陣 X 之奇異值分解

3.將 SVD 分解後得到的矩陣，取前 k 個最大的奇異值做維度約化，以消除此語意空間中的雜訊，並重建矩陣 $X \approx X' = T_k S_k D_k^t$，經過維度約化（dimension reduction）所得到的新語意空間 S_k 較能精確的表示詞彙與文件間的關係。

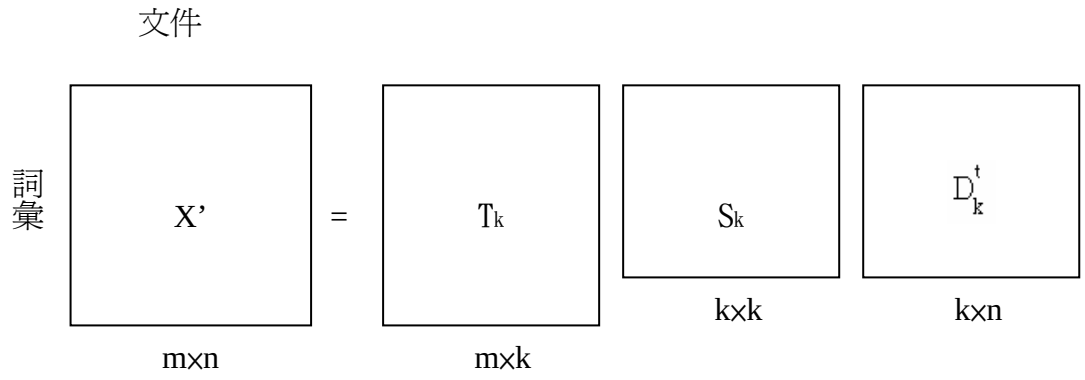

圖 2-3 維度約化的新矩陣 X'

以下利用實例說明 LSA 的運作流程（Landauer, Foltz, &Laham, 1998）：

假設有九篇文件如圖 2-4，其中 c1~c5 與人機互動（Human Computer Interaction, HCI）有關，m1~m4 與數學圖形理論（Mathematical Graph Theory）有關。

Example of text data: Titles of Some Technical Memos
c1: *Human* machine *interface* for 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, ordered *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*

圖 2-4 技術文件的標題

首先從這份文件中挑選出現兩次以上的詞彙（斜體部份），共計 12 個，將這些詞彙和文件建立一個矩陣 X，矩陣 X 每一列為這個詞彙在每一篇文件中出現的次數，每一行為這篇文件中出現多少次詞彙，如表 2-1。

表 2-1 利用圖四之文件所產生的詞彙與文件關係的矩陣 X

	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

要得到兩個辭彙間的相似度，可利用向量空間模型（Vector Space Model， VSM），其方式為計算矩陣中列向量(row vector)的內積值，公式如下（Saton,1968）：

$$\mathrm{sim}(t_i, t_j) = \cos(v_i, v_j) = \frac{v_i v_j^{t}}{\| v_i \| \| v_j \|} \quad \cdots\cdots(25)$$

t_i、t_j為詞彙i和詞彙j，$sim(t_i,t_j)$為兩個詞彙的相似度，v_i、v_j為兩個詞彙在矩陣X的列向量。

若要求的是兩篇文件的相似度，則利用相同的公式求矩陣中行向量（column vector）的內積即可。利用公式25與表2-1，因為user和human並沒有出現在同一篇文件，利用公式25得到user和human兩個詞彙的相似度為0，而user和survey因為曾經同時出現在c2這篇文件中，所以利用公式25得到user和survey兩個詞彙的相似度為0.4082，所以若只由表2-1的矩陣，得到的結果是user和human的相似度小於user和survey的相似度。

將矩陣X做奇異值分解，得到三個矩陣T_r、S_r、D_r，如圖2-5，且$X = T_r S_r D_r^{t}$。其中T_r為XX^H的特徵向量，D_r為$X^H X$的特徵向量，S_r為$X^H X$特徵值的正平方根，特徵值代表每一維度的變異程度，變異程度愈大者較具有鑑別力，因此可保留特徵值較大的維度即可，要保留特徵值多大的維度，或保留多少維度，並沒有理論上的最佳值，因此需計算保留維度不同時效果的差異，以找出此文件集最佳的保留維度(Wang & Nie,2003)。在本實例，暫不考慮不同維度之差異，因此研究者以保留2個特徵值最大的維度做計算，計算方式為取矩陣S_r中最高的兩個值，並取T_r的前兩行與D_r^{t}的前兩列，利用$X' = T_k S_k D_k^{t}$可得到到新矩陣X'如表2-2。

T =

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

S =

$$\begin{pmatrix}
3.34 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 2.54 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2.35 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1.64 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1.5 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1.31 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.85 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.56 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.36
\end{pmatrix}$$

D^t =

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

圖 2-5 將表 2-1 之矩陣 X 經奇異值分解所得到的三個矩陣 T、S、D^t

表 2-2 將圖 2-5 之三個矩陣取 2 個最大的特徵值做約化後所得到的新矩陣 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	-0.1	-0.21	0.15	0.22	0.50	0.71	0.62

利用新矩陣X'，可得到user和human兩個列向量間的內積為0.8878，user和survey間的內積為0.7752，比對利用原始矩陣X所得的結果如下：

表2-3 利用新舊矩陣所得到的辭彙相似度之比較

	原始矩陣 X	新矩陣 X'
user 和 human 的相似度	0	0.8878
user 和 survey 的相似度	0.4082	0.7752

由以上之例可知雖然user和human未出現在同一篇文件，但利用LSA可得到其相似程度高於user和survey的相似度。

利用原始矩陣X和新矩陣X'，求文件c1和其他文件間的相似度，與文件c2和其他文件間的相似度，可得到表2-4：

表2-4 利用新舊矩陣所得到的c1文件和其他文件相似度之比較

矩陣類別 / 文件編號	原始矩陣 X	新矩陣 X'
c2	0.2357	0.9998
c3	0.2887	0.9999
c4	0.2357	0.9999
c5	0	0.9996
m1	0	0.9996
m2	0	0.9947
m3	0	0.9941
m4	0	0.9889

由表2-4可知，利用原始矩陣X，文件c1和c5因為沒有任何相同的辭彙，所以算出的相似度為0，而利用LSA所求得的新矩陣X' ，則可求得文件c1和c5的相似度為0.9996，而且其值皆大於不同類別的文件m。

表2-5 利用新舊矩陣所得到的c2文件和其他文件相似度之比較

矩陣類別 / 文件編號	原始矩陣 X	新矩陣 X'
c1	0.2357	0.9998
c3	0.4082	0.9848
c4	0.3333	0.9166
c5	0.7071	0.9877
m1	0	0.8681
m2	0	0.8799
m3	0	0.8827
m4	0.2357	0.9142

由表2-5可知，利用原始矩陣X，文件c2和m4是不同類別，但其相似度為0.2357，與c2和c1的相似度相同，而利用LSA所求得的新矩陣X' ，則可求得文件c2和m4的相似度為0.9142，小於c2和c1的相似度0.9998。因此可知利用LSA所得到的新矩陣X' ，較能反映文件與文件間實際的相似情形。

第三節 中文斷詞

要將文件集表示成如圖 2-1 的矩陣時，由於中文句子的詞彙和詞彙之間，不像英文以空白做區隔，因此計算詞彙在文件中的重要性前，必須先對文件做斷詞處理，以下就中文斷詞會遇到的問題與解決的方法，簡述如後。

1. 中文斷詞的問題

在做中文斷詞處理時，常遇到下列兩個問題：

（一）未知詞：斷詞最簡單的方法是查辭典，但由於新詞不斷的產生（如姓名），若辭典未包含這些詞彙，會導致斷詞結果有誤，此即未知詞的問題。

（二）歧異性：即一個句子可能會有不同的斷詞組合，如「開創新生活」由於相鄰的兩個字皆可連接成為一個詞，所以可能的斷詞組合有：

「開創 新生 活」、「開創 新 生活」、「開 創新 生活」與「開創 新生活」等。

二、中文斷詞的方法

中文斷詞的方法目前最常見的有辭庫式斷詞、法則式斷詞和統計式斷詞：

（一）辭庫式斷詞：將句子中的詞與辭典做比對，以找出可能的詞彙，最常用的方法是正向最大匹配法（forward maximum matching algorithm）和反向（backward）最大匹配法（Lua & Gan ,1994），正向最大匹配法的做法是從句子的句首正向與辭典的最長詞比對，比對成功後再處理剩下的字串，直到全部比對完成。反向最大匹配法的方式和正向最大匹配法相似，不同的是它從句尾

反向做匹配。這兩種方法的優點是簡單且易於實作，缺點是會出現未知詞的問題。

（二）法則式斷詞：利用語言的特性，建立構詞法則，並配合辭庫做斷詞處理。如利用語法配合詞長與詞頻做斷詞（王良志等，1991），利用構詞原則配合組合率做斷詞（陳克健、陳正佳、林隆基，1986）。這些方法的優點是利用少量的語法規則即能處理大量的資料，但由於中文語法的複雜性，一套完整的語法規則並不容易建立，因此經常會有例外狀況產生。

（三）統計式斷詞：利用大量語料庫上的統計資訊，以鄰近字元同時出現頻率高低作為斷詞的依據，採用此法有統計斷詞法（Sproat& Shih,1990）和鬆弛法（范長康，1989），統計式斷詞法定義中文的「詞彙」為一串相鄰且出現機率較高的字，因此從大量語料庫得到相鄰字出現的頻率後，再利用一階馬可夫機率模式來決定斷詞點。鬆弛法（relaxation）為影像處理常用的方法，此法將斷詞視做一種對句中各字做「字詞指派」的過程。利用句中字詞間的組成關係做為指派方式的約束條件，在執行鬆弛程序時，這些約束條件將剔除不相容的指派，以找出正確的斷詞結果。這些方法由於沒有考慮詞的正確性，所以經常會產生無意義的斷詞結果。