應用潛在語意分析探究詞彙對語料庫之重要性

白鎧誌
國立臺中教育大學
教育測驗統計研究所
碩士生
minbai0926@gmail.com

李政軒
國立交通大學
電控工程研究所
博士生
ChengHsuanLi@gmail.com

郭伯臣
國立臺中教育大學
教育測驗統計研究所
教授
kbc@mail.ntcu.edu.tw

廖晨惠
國立臺中教育大學
特殊教育學系
副教授
chenhueiliao@gmail.com

摘要

過去潛在語意分析(Latent semantic analysis, LSA)相關研究中，已有探討詞彙對於文件的重要程度，但尚未有探討詞彙對於整個語料庫的重要性，因此本研究主要應用 LSA 技術，探討中文語意空間詞彙對於語料庫的重要性程度。本研究依據探討詞彙對於文件重要程度相關研究成果，擬訂六種不同的指標，並使用三種詞彙權重方法建立不同的中文語意空間。依據六種指標探討每一指標與詞頻、詞彙出現的文件數之間的相關性，本研究使用相關分析，給予詞頻、詞彙出現的文件數不同權重係數以探討對於指標的相關性，期望能找到最相關的指標，作為詞彙對於語料庫的重要性程度的依據。

關鍵詞：潛在語意分析、語料庫、中文語意空間、相關分析

Abstract

In the past of research about latent semantic analysis, some had discussed the importance of terms in the document, but had no discussion about the importance of terms in the corpus. In accordance with some research results for the importance of terms in the document, this study investigated the importance of terms in Chinese semantic space applying latent semantic analysis. This research developed six different indicators, and used three term weighting schemes to investigate the correlation between these indicators, term frequency and term appears in the number of all documents. The study used correlation analysis to find the highest correlation of these indicators for the importance of terms.

Keywords: Latent semantic analysis, corpus, Chinese semantic space, correlation analysis.

1. 前言

潛在語意分析(Latent semantic analysis, LSA)是一種基於向量空間模型概念之技術，是使用線性代數且能有效自動化資訊檢索的方法[9]。LSA 最早是由 Deerwester 等人所提出，主要應用於資訊檢索方面。LSA 以統計的方法為基礎知識模型，其運作方式與類神經網路(Neural Net)極為相似，差別在於類神經網路是以權重的傳遞與回饋來修正本身的學習，LSA 則是以奇異值分解(Singular Value Decomposition, SVD)與維度約化(Dimension Reduction)為核心作為邏輯推演的方式[4]。而根據 Landauer 與 Dumais 提出的概念，LSA 可藉由統計計算方法應用到大量文本建置的語料庫，用來提取和表示上下文詞彙的意義。而最近有研究將 LSA 應用於心理語言分析的研究領域[8]，目的在利用大型語料庫分析提出一種定義詞或段落之間的相似度的方法，使用 LSA 建置一個可以反映語彙知識背後語意關係的語意空間。

詞彙對於文件重要性相關研究中，在資訊檢索或資料探勘研究主要是探討詞彙權重部分。相關研究提出 TF-IDF 的統計方法，用以評估詞彙對於一個文件集或一個語料庫中的其中一份文件的重要程度。TF 指的是某一個給定的詞彙在該文件中出現的次數，而 IDF 表示一個逆向的文件頻率，加入 IDF 此方法是考慮到依詞頻來判斷該詞彙對於文件之重要性是不夠客觀的，因此加入考量詞彙出現在所有文件的次數，即一個詞彙出現在有所文件的頻率倒數，作為加入詞彙權重的計算[11]。此外，在 LSA 相關研究中，詞彙權重給定方式大多是採用 Log-Entropy 方法，目的也是考慮到詞彙對於文件的重要性程度。

綜合上述，本研究使用中研院建置的現代漢語平衡語料庫(3.1)版，共有 9227 篇文章，並篩選在語料庫中出現次數兩次以上之詞彙作為本研究選取之詞彙，共有 78444 個詞彙，利用 LSA 技術與不同詞彙權重方法建立維度為

400、350、300、250、200、150、100、50 的中文語意空間，並訂定六種指標方法，利用相關分析方法，探討六種指標分別對於詞頻與詞彙出現的文件數之間的相關性。

本研究於文獻探討中分別說明應用 LSA 於中文語意空間建置與相似度計算、LSA 詞彙對於文件的重要性程度相關研究，並於研究方法中說明本研究使用之三種詞彙權重方法與擬提出之六種指標方法，最後於實驗結果中探討本研究依據相關研究成果擬訂的六種指標，期望找出與詞頻、詞彙出現的文件數中最相關的指標。

2. 文獻探討

2.1 中文語意空間建置與相似度計算

使用潛在語意分析技術建置語意空間，需要以下幾個步驟：(1)建立詞彙-文件共生矩陣；(2)詞彙權重計算；(3)運用 SVD 轉換矩陣；(4)維度約化[9]。

建立詞彙-文件共生矩陣是從文件中找出關鍵詞來做處理與比對的依據，在LSA中對關鍵詞的定義為在欲處理的文件中出現過兩次以上的字。從建置的詞彙-文件共生矩陣可以看出某一個關鍵詞在不同的句子、段落或文件中所出現的情形與每一個句子、段落或文件中有那些重要的關鍵詞。

詞彙權重計算是考慮並非所有的辭彙和所有的文件都有相同的重要性，根據不同的應用研究，關鍵詞彙其重要性也並不完全相同，因此需給予不同的權重。LSA詞彙權重的調整方式可分為local權重和global權重兩部份。local權重是指詞彙在文件中的重要性，通常以詞彙出現次數代表，出現次數愈多代表愈重要；global 權重則是詞彙在整個語料庫的重要性，與local 權重相反，出現次數愈多，對於文件的重要性相對愈低 [2][5][7] 。

在 LSA 中最常用來分解詞彙－文件共生矩陣 A 的方法為 SVD，藉由 SVD 的運算過程，可以計算出每個詞彙在對角矩陣中的特徵值，一般來說特徵值愈大的向量，表示具有較大的訊息量，反之則只有微小的訊息量。經過 SVD 轉換後的矩陣，關鍵詞和文件的關係，就不是原本出現次數的關係，取而代之的是表徵關鍵詞在文件中的語意關係。經過 SVD 轉換一個 $m \times n$ 的詞彙－文件共生矩陣 A ，可被拆解成

$$A = U \sum V^T \tag{1}$$

其中 U 是正交矩陣(orthogonal matrix)或稱為左奇異向量(Left singular value)，V 為正交矩陣(orthogonal matrix)或稱為右奇異向量(Right singular value)，$\sum$ 為由奇異特徵值組成的對角矩陣($\sum = \text{diagonal}(\lambda_1, \lambda_2, \cdots, \lambda_r)$，其於元素皆為 0)[7]，U 矩陣的列向量稱為詞彙向量(type vector)，而 V 矩陣的列向量稱為文件向量(document vector)[8]。

經由 SVD 後，因為語意空間的矩陣過大或是太多所謂的雜訊(noise)，則可以利用維度約化(dimension Reduction)，消除語意空間中不重要之雜訊，其方式是取出 SVD 後前 k 個最大的特徵奇異值，和 U 矩陣、V 矩陣前 k 個行向量($k < r$)，並重建矩陣 $\tilde{A} = W_k \sum_k V_k^T$，如圖 1 所示[1] [12]。

經由 SVD 重新建置的矩陣 $\tilde{A}$，是將詞彙、段落句子或文章以向量形式呈現該詞彙、段落句子或文章在語意空間的相對位置，以兩向量角度的餘弦值表示兩個文件之間的相似度，餘弦值愈大表示兩向量夾角愈小，在語意空間表示兩向量之間的語意愈相似。所以由 LSA 建置的語意空間，詞彙與詞彙、詞彙與段落、段落與段落間語意相似性都能藉由兩向量之間的餘弦值表示[3]。

假設要判斷第 *i* 個詞彙和第 *j* 個詞彙的相似度，則可利用 VSM(vector space model)求兩向量的夾角(餘弦)，即可求得其相似度，公式如下：

$$\cos(t_i, t_j) = \frac{t_i t_j^T}{\|t_i\| \|t_j\|} \tag{2}$$

其中 t_i 表示第 *i* 個詞彙在語意空間的向量表徵，$\|t_i\|$ 為 t_i 向量的長度。

假設若要利用 LSA 語意空間判斷兩篇文章(文章一用向量 T_1 表示和文章二用向量 T_2 表示，非語料庫之文章)的相似度，可以利用下述方式計算出：

$$T_1 = (a_1 t_1, a_2 t_2, \cdots, a_n t_n) \tag{3}$$

$$T_2 = (b_1 t_1, b_2 t_2, \cdots, b_n t_n) \tag{4}$$

a_i 表示第 *i* 個詞彙在文章一出現的次數，b_i 表示第 *i* 個詞彙在文章二出現的次數，t_i 表示第 *i* 個詞彙在語意空間的向量表徵。基於上述兩篇文章在 LSA 語意空間的向量，其相似程度可由下列公式計算得出：

$$\cos(T_1,T_2)=\frac{T_1T_2^T}{\|T_1\|\|T_2\|}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}a_ib_jt_it_j^T}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jt_it_j^T}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}b_ib_jt_it_j^T}} \quad (5)$$

$\|T_i\|$ 為 T_i 向量的長度。

2.2 LSA 詞彙對於文件之重要性

McNamara、Cai and Louwerse 的研究[5]，提出在文件中每個詞彙對於文件是否都有同樣的重要性的問題；一般認知是認為給予詞彙加權，權重愈大的詞彙對文件的影響較大。其研究所定義的文件與文件之間相似度公式如下：

$$sim(T_1,T_2)=\frac{(\sum_{i=1}^{n}a_iv_i)(\sum_{i=1}^{n}b_iv_i)^{\mathrm{T}}}{\left\|\sum_{i=1}^{n}a_iv_i\right\|\left\|\sum_{i=1}^{n}b_iv_i\right\|} \quad (6)$$

n 表示此語意空間的詞彙，v_i 表示第 i 個詞彙向量，a_i 表示在 T_1 文件中出現的詞彙的次數，而 b_i 表示在 T_2 文件中出現的詞彙的次數。

此研究定義文件向量為將文件中出現的詞彙的向量加總，假設一文件中有語意空間中十個詞彙出現，即將這十個詞彙的詞彙向量加總，其公式如下：

$$v=v_1+v_2+...+v_{10} \quad (7)$$

此研究實驗分別刪除詞彙對於刪除該詞彙的文件與原來文件做相似度(cosine 值)比對，其研究結果發現經由 SVD 轉置重新建置的詞彙文件矩陣，詞彙的向量長度對於 LSA 的 cosine 值有負高相關(相關性=-0.94，$p<0.01$)，即詞彙向量長度愈大，其刪除該詞彙之文件與原來文件兩者之間的 cosine 值愈低，而詞彙權重對於文件影響則呈現負低相關(相關性=-0.19)，表示詞彙向量長度對於文件的影響有高相關性。

表 1 是其研究結果，LSA 為刪除該詞彙之文件與原來文件之 cosine 值，length 為該詞彙之向量長度，α_i 為該詞彙計算之權重。

表 1 McNamara 等人(2007)研究結果

	LSA	Length	α_i
as	.996	.10	.04
they	.969	.32	.07
were	.958	.34	.09
being	.962	.31	.20
dragged	.999	.04	.51
off	.953	.35	.19
cried	.987	.19	.35
for	.998	.08	.03
help	.872	.56	.20

3.研究方法

本研究使用之語料庫為中央研究院建置的現代漢語平衡語料庫(3.1)版，共有 9227 份文件，五百萬詞。本研究進行關鍵詞篩選後挑選出 78444 個詞彙作為本研究定義之關鍵詞，並使用三種不同詞彙權重方法建立中文語意空間，分別是：Log-Entropy、Log-IDF、TF-IDF，並擬訂六種指標方法與五種線性組合加權方式，將詞頻、詞彙出現的文件數線性組合，並給予不同的加權係數。以下是本研究訂定的 Y 變項公式描述：

$$Y=\alpha_1Y_1+\alpha_2Y_2 \quad (8)$$

Y_1 表示詞頻，Y_2 表示詞彙出現文件數，本研究給與五種加權係數：(1) α_1=1、α_2=0，(2) α_1=0.75、α_2=0.25，(3) α_1=0.5、α_2=0.5，(4) α_1=0.25、α_2=0.75，(5) α_1=0、α_2=1，分別探討此五種加權，其 Y 變項對於指標之相關性。

3.1 至 3.3 小節分別說明詞彙權重給定公式，3.4 至 3.9 小節針對各指標方法做說明。

3.1 Log-Entropy 詞彙權重

本研究選擇的詞彙的 local 權重與 global 權重公式如下：

$$L(i,j)=\log(tf_{ij}+1) \quad (9)$$

$L(i,j)$ 表示 local 權重，tf_{ij} 表示第 i 個詞彙在第 j 個文件出現的次數。

$$G(i)=1+\sum_{\mathrm{j}}\frac{p_{ij}\log_2(p_{ij})}{\log_2 n}\ ,\ p_{ij}=\frac{tf_{ij}}{gf_i} \quad (10)$$

$G(i)$ 表示 global 權重，gf_i 是第 i 個詞彙在所有文件中出現次數的總和。

3.2 Log-IDF 詞彙權重

Log-IDF 給定的詞彙權重公式如下：

$$L(i,j)=\log(tf_{ij}+1) \quad (11)$$

$$G(i)=\log(m/df(i)) \quad (12)$$

tf_{ij} 表示第 i 個詞彙在第 j 個文件出現的次數，m 表示文件的個數，$df(i)$表示每一個詞彙 i 出現在多少個文件中。

3.3 TF-IDF 詞彙權重

TF-IDF 的詞彙權重給定公式如下：

$$L(i,j)=\frac{n_{i,j}}{\sum_{k}^{N} n_{k,j}} \quad (13)$$

$$G(i)=\log(m/df(i)) \quad (14)$$

$L(i,j)$ 表示 local 權重，$n_{i,j}$ 表示第 i 個詞彙在第 j 個文件中出現的次數，$\sum_{k}^{N} n_{k,j}$ 表示在第 j 個文件中所有詞彙的出現次數的總和。$G(i)$ 表示 global 權重，m 表示文件的個數，$df(i)$ 表示第 i 詞彙出現在多少個文件中。

3.4 SVD 重新建置的矩陣之詞彙向量長度

將 U、Σ、V^T 重新相乘形成新的矩陣 Ã，如公式 1 所示，其 Ã 矩陣中的每一列向量表示每個詞彙的向量，公式定義如下：

$$T_i=U_i \Sigma V^T \quad (15)$$

T_i 表示第 i 個詞彙的向量，其每一個詞彙向量長度之定義如下：

$$\|T_i\|=\|U_i \Sigma V^T\| \quad (16)$$

$\|T_i\|$ 為 T_i 向量的長度，本研究將每一詞彙向量長度當作詞彙對於corpus的重要性程度參考之指標，即作為本研究的指標一。

3.5 SVD 分解的詞彙向量矩陣

經由公式 1 所得到的 U 矩陣即為詞彙向量矩陣，因此本研究取 U 矩陣中每個詞彙的向量長度當作詞彙對於 corpus 的重要性程度，作為本研究的指標二。

3.6 U、V 矩陣之 cosine 平均值

經由 SVD 得到的 U 矩陣為詞彙向量矩陣，V 為文件向量矩陣，本研究計算 U 與 V 兩者之間的餘弦值，即每個詞彙對於每份文件的相似度(cosine 值)，其公式如下：

$$\cos(U_i,V_j)=\frac{U_i V_j^T}{\|U_i\|\|V_j^T\|} \quad (17)$$

U_i 表示第 i 個詞彙向量，V_j 表示第 j 個文件向量。本研究取每詞彙與每份文件之相似度的平均值當作每個詞彙對於整個 corpus 的重要性程度，作為本研究的指標三。

3.7 詞彙與其他詞彙之間之 cosine 平均值

經由 U、Σ、V^T 重新相乘建置的的新矩陣 Ã，計算矩陣中每一詞彙向量與其他詞彙向量之 cosine 值，其 cosine 值計算如公式 2，再取每一詞彙與其他詞彙之 cosine 值的平均值作為本研究的指標四。

3.8 詞彙與詞彙向量加總之 cosine 值

根據[5]的研究，定義文件向量為文件中出現的詞彙向量之加總，因此本研究利用經由 U、Σ、V^T 重新相乘建置的新矩陣 Ã，將 corpus 中所有詞彙的向量加總做為整個 corpus 文件向量，其公式定義如下：

$$\sum_{i=1}^{n} v_i \quad (18)$$

假設共有 n 個詞彙，v_i 表示第 i 個詞彙的向量，將所有詞彙向量加總作為整份文件之文件向量，再計算每一詞彙與文件向量之 cosine 值，作為本研究的指標五。

3.9 詞彙與文件之 cosine 平均值

依據[5]的研究中，定義文件的向量為將文件中出現的詞彙的向量加總，作為該文件的向量，因此本研究將每份文件所出現之詞彙向量加總做為該文件向量，其公式如下：

$$v=v_1+v_2+\ldots+v_i \quad (19)$$

v_i 表示所有詞彙中第 i 個詞彙出現在該文件中之向量。本研究比對每一詞彙與每一文件之 cosine 值，其公式如下：

$$sim(t_i,v_j)=\frac{t_i v_j^T}{\|t_i\|\|v_j^T\|} \quad (20)$$

t_i 表示第 i 個詞彙向量表徵，v_j 表示第 j 份文件中出現的詞彙向量之加總。本研究取每一詞彙

與每份文件的 cosine 值之平均值作為該詞彙對 corpus 之重要性程度，即作為本研究的指標六。

4. 實驗結果

表 2 至表 16 為本研究使用三種不同詞彙權重方法、在不同維度下與不同權重係數，各指標與詞頻、詞彙出現的文件數之相關性實驗結果，以下針對各表作說明。

表 2 至表 4 為給予詞頻權重為 1，詞彙出現文件數為 0，其與各指標相關分析結果顯示在 Log-Entropy 詞彙權重中指標一與詞頻、詞彙出現文件數相關性較高，相關性皆達 0.5 以上。表 5 至表 7 為詞頻權重 0.75，詞彙出現文件數 0.25，其相關分析結果呈現 Log-Entropy 詞彙權重中指標一與詞頻、詞彙出現文件數相關性較高，相關性皆達 0.55 以上

表 8 至表 10 給予詞頻權重 0.5，詞彙出現文件數 0.5，其相關分析結果發現 Log-Entropy 詞彙權重中的指標一與詞頻、詞彙文件數相關性較高，相關係數幾乎達 0.6。表 11 至表 13 給予詞頻權重 0.25，詞彙出現文件數.75，其分析結果 Log-Entropy 詞彙權重中的指標一與詞頻、詞彙文件數相關性較高，相關性皆達 0.6 以上。表 14 至表 16 是詞頻權重給予 0，詞彙出現文件數為 1，其相關分析結果也呈現以 Log-Entropy 詞彙權重方法的指標一與詞頻、詞彙文件數有較高相關性。

由上述本研究實驗結果可以發現，本研究使用三種詞彙權重方法與在不同的維度空間下，其實驗結果皆呈現指標一與詞頻、詞彙出現的文件數有較高相關性，而以三種詞彙權重方法比較結果，Log-Entropy 詞彙權重方法的各指標與詞頻、詞彙出現文件數有較高相關性。這與過去研究探討 LSA 詞彙對於文件重要程度影響之因素大致呈現一致，可以作為詞彙對於語料庫的重要性程度之參考。

5. 結論

本研究主要是應用 LSA 技術探討中文語意空間詞彙對於語料庫的重要性程度。並使用三種不同詞彙權重方法與六種不同比較指標，利用相關分析方法探討指標與詞頻、詞彙出現的文件數之間的相關性，實驗結果發現指標一，利用 SVD 重新建置的矩陣之詞彙向量長度與詞頻、詞彙出現的文件數之間有較高相關性。詞彙權重方法的比較發現 Log-Entropy 詞彙權重方法在各指標與詞頻、詞彙出現文件數之相關性較高。綜合上述，利用 Log-Entropy 詞彙權重方法與指標一，可以作為詞彙對於語料庫的重要性程度之參考。未來本研究期望能找出與詞頻、詞彙出現在所有文件的次數之間更高相關性的新指標，並加入專家對於詞彙重要性程度評分，藉由與專家評分結果驗證本研究之實驗結果，以做更進一步的探討與研究。

致謝

本研究為國科會之研究成果之計畫，編號 NSC-100-2420-H-142-001-MY3，由於國科會的支持，使本研究得以順利進行，特此致上感謝之意。

本研究感謝中華民國計算語言學會、中央研究院詞庫小組與中華民國國家科學委員會。

參考文獻

[1] Berry, M.W., Dumais, S., & O'Brien, G. (1995). Using linear algebra for intelligent information retrieval. ***SIAM Review***, 37, 573–595.

[2] Berry, M.W., & Browne, M. (2005). Understanding search engines: Mathematical modeling and text retrieval. ***Philadelphia: SIAM***, 2.

[3] Cooley, W. W. & Lohnes, P. R. (1971) ***Multivariate Data Analysis.*** Wiley, New York, NY.

[4] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis, ***Journal of the American Society for Information Science***, 41, 391-407.

[5] Dumais, S. (1991). Improving the retrieval of information from external sources. ***Behavior Research Methods, Instruments, and Computers***, 23, 229–236.

[6] Golub, G. and Van Loan, C. (1989). ***Matrix Computations***, Baltimore, MD: Johns Hopkins, 2.

[7] Letsche, T., & Berry, M. W. (1997). Large-scale information retrieval with latent semantic indexing. ***Information Sciences***, 100, 105–137.

[8] Landauer, T.K., Foltz, P. W., & Laham, D. (1998). An introduction to latentsemantic analysis. ***Discourse Processes***, 25, 259-284.

[9] Martin, D.I., & Berry, M. W.(2007). Mathematical Foundations Behind Lantent Semantic Analysis. In T. K. Landauer, D. S.

McNamara, S. Dennis, & W. Kintsch (Eds.), ***Handbook of Latent Semantic Analysis.*** (pp. 35-55). Mahwah, NJ: Lawrence Erlbaum Associates.

[10] McNamara, D. S., Cai, Z., & Louwerse, M. M. (2007). Optimizing LSA measures of cohesion. In Landauer, T., D.S. McNamara, S. Dennis, & W. Kintsch (Eds.), ***Handbook of Latent Semantic Analysis***. (pp. 379-400). Mahwah, NJ: Erlbaum Associates.

[11] Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. ***Information Processing and Management***, 24(5), 513-523.

[12] Witter, D., & Berry, M. W. (1998). Downdating the latent semantic indexing model for conceptual information retrieval. ***The Computer Journal***, 41, 589–6

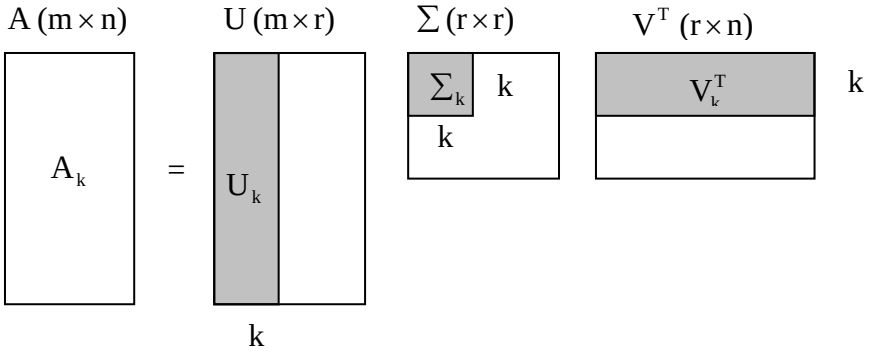

圖 1 語意維度約化圖示

表 2 Log-Entropy 詞彙權重各指標與詞頻、詞彙出現的文件數相關分析結果

(詞頻權重=1、詞彙出現的文件數權重=0)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.532	0.450	0.351	0.234	0.273	0.262
350	0.538	0.456	0.341	0.226	0.264	0.253
300	0.545	0.461	0.328	0.217	0.255	0.244
250	0.552	0.465	0.314	0.207	0.246	0.235
200	0.559	0.467	0.300	0.197	0.237	0.224
150	0.568	0.466	0.285	0.185	0.226	0.213
100	0.580	0.463	0.265	0.168	0.211	0.197
50	0.608	0.467	0.227	0.137	0.182	0.165

表 3 Log-IDF 詞彙權重各指標與詞頻、詞彙出現的文件數相關分析結果

(詞頻權重=1、詞彙出現的文件數權重=0)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.492	0.424	0.386	0.241	0.280	0.269
350	0.498	0.429	0.374	0.232	0.271	0.260
300	0.504	0.433	0.358	0.222	0.262	0.250
250	0.509	0.435	0.343	0.212	0.252	0.240
200	0.516	0.436	0.325	0.201	0.242	0.229
150	0.523	0.435	0.305	0.187	0.229	0.216
100	0.534	0.430	0.280	0.170	0.213	0.199
50	0.558	0.431	0.234	0.137	0.182	0.166

表 4 TF-IDF 權重各指標與詞頻、詞彙出現的文件數之相關分析結果

(詞頻權重=1、詞彙出現的文件數權重=0)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.460	0.394	0.242	0.194	0.246	0.230
350	0.463	0.393	0.234	0.187	0.239	0.223
300	0.465	0.391	0.227	0.180	0.233	0.216
250	0.469	0.389	0.217	0.172	0.225	0.208
200	0.474	0.386	0.207	0.163	0.216	0.199
150	0.482	0.383	0.195	0.151	0.205	0.186
100	0.499	0.388	0.180	0.134	0.188	0.169
50	0.535	0.410	0.148	0.103	0.154	0.133

表 5 Log-Entropy 詞彙權重各指標與詞頻、詞彙出現的文件數相關分析結果

(詞頻權重=0.75、詞彙出現的文件數權重=0.25)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.565	0.487	0.361	0.250	0.292	0.280
350	0.571	0.492	0.351	0.242	0.283	0.271
300	0.577	0.496	0.338	0.232	0.274	0.262
250	0.583	0.500	0.325	0.223	0.265	0.253
200	0.590	0.500	0.312	0.212	0.256	0.242
150	0.597	0.497	0.298	0.200	0.245	0.230
100	0.609	0.491	0.280	0.183	0.229	0.214
50	0.636	0.492	0.242	0.149	0.198	0.181

表 6 Log-IDF 詞彙權重各指標與詞頻、詞彙出現的文件數相關分析結果

(詞頻權重=0.75、詞彙出現的文件數權重=0.25)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.527	0.461	0.396	0.257	0.299	0.287
350	0.532	0.465	0.384	0.248	0.290	0.278
300	0.538	0.469	0.368	0.238	0.281	0.268
250	0.543	0.470	0.353	0.228	0.271	0.258
200	0.549	0.470	0.337	0.216	0.261	0.247
150	0.555	0.467	0.318	0.203	0.248	0.234
100	0.565	0.460	0.294	0.184	0.232	0.216
50	0.589	0.459	0.248	0.150	0.199	0.182

表 7 TF-IDF 權重各指標與詞頻、詞彙出現的文件數之相關分析結果

(詞頻權重=0.75、詞彙出現的文件數權重=0.25)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.484	0.421	0.253	0.209	0.265	0.248
350	0.486	0.420	0.245	0.202	0.259	0.242
300	0.488	0.416	0.238	0.195	0.253	0.235
250	0.491	0.413	0.229	0.187	0.245	0.226
200	0.495	0.408	0.219	0.178	0.236	0.217
150	0.502	0.404	0.207	0.165	0.224	0.204
100	0.518	0.406	0.193	0.147	0.206	0.185
50	0.554	0.426	0.160	0.113	0.169	0.147

表 8 Log-Entropy 詞彙權重各指標與詞頻、詞彙出現的文件數相關分析結果

(詞頻權重=0.5、詞彙出現的文件數權重=0.5)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.595	0.521	0.369	0.265	0.309	0.297
350	0.601	0.526	0.359	0.256	0.301	0.288
300	0.607	0.530	0.347	0.247	0.292	0.279
250	0.612	0.532	0.334	0.238	0.283	0.269
200	0.618	0.531	0.322	0.227	0.274	0.259
150	0.624	0.526	0.310	0.214	0.262	0.247
100	0.634	0.517	0.292	0.197	0.247	0.230
50	0.660	0.514	0.255	0.161	0.214	0.195

表 9 Log-IDF 詞彙權重各指標與詞頻、詞彙出現的文件數相關分析結果

(詞頻權重=0. 5、詞彙出現的文件數權重=0.5)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.560	0.496	0.403	0.272	0.317	0.304
350	0.565	0.500	0.392	0.263	0.308	0.295
300	0.570	0.502	0.377	0.253	0.298	0.285
250	0.574	0.503	0.363	0.242	0.289	0.275
200	0.579	0.502	0.347	0.231	0.278	0.264
150	0.585	0.498	0.329	0.217	0.266	0.251
100	0.593	0.487	0.306	0.198	0.249	0.233
50	0.617	0.484	0.261	0.161	0.215	0.196

表 10 TF-IDF 權重各指標與詞頻、詞彙出現的文件數之相關分析結果

(詞頻權重=0.5、詞彙出現的文件數權重=0.5)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.506	0.446	0.262	0.223	0.284	0.265
350	0.507	0.444	0.255	0.217	0.278	0.259
300	0.508	0.439	0.249	0.210	0.272	0.252
250	0.510	0.435	0.240	0.201	0.264	0.244
200	0.513	0.428	0.231	0.191	0.254	0.234
150	0.520	0.422	0.219	0.178	0.242	0.220
100	0.535	0.422	0.205	0.159	0.223	0.200
50	0.571	0.439	0.172	0.123	0.184	0.160

表 11 Log-Entropy 詞彙權重各指標與詞頻、詞彙出現的文件數相關分析結果

(詞頻權重=0.25、詞彙出現的文件數權重=0.75)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.622	0.552	0.374	0.279	0.325	0.312
350	0.628	0.557	0.365	0.270	0.317	0.303
300	0.633	0.560	0.354	0.261	0.308	0.294
250	0.638	0.561	0.342	0.251	0.299	0.285
200	0.642	0.559	0.330	0.241	0.290	0.275
150	0.648	0.552	0.320	0.228	0.279	0.263
100	0.656	0.540	0.304	0.210	0.263	0.245
50	0.681	0.534	0.267	0.172	0.229	0.209

表 12 Log-IDF 詞彙權重各指標與詞頻、詞彙出現的文件數相關分析結果

(詞頻權重=0.25、詞彙出現的文件數權重=0.75)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.589	0.528	0.408	0.285	0.333	0.319
350	0.594	0.531	0.398	0.276	0.324	0.310
300	0.598	0.534	0.383	0.266	0.314	0.300
250	0.602	0.534	0.370	0.256	0.305	0.290
200	0.606	0.531	0.355	0.244	0.295	0.280
150	0.611	0.525	0.339	0.230	0.282	0.266
100	0.618	0.512	0.317	0.211	0.265	0.248
50	0.642	0.506	0.272	0.172	0.229	0.209

表 13 TF-IDF 權重各指標與詞頻、詞彙出現的文件數之相關分析結果

(詞頻權重=0.25、詞彙出現的文件數權重=0.75)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.525	0.469	0.270	0.236	0.301	0.281
350	0.526	0.466	0.263	0.230	0.295	0.275
300	0.526	0.460	0.258	0.223	0.289	0.268
250	0.527	0.454	0.249	0.214	0.281	0.260
200	0.529	0.446	0.241	0.204	0.272	0.250
150	0.535	0.438	0.230	0.190	0.259	0.236
100	0.549	0.436	0.216	0.171	0.239	0.215
50	0.584	0.449	0.182	0.133	0.198	0.172

表 14 Log-Entropy 詞彙權重各指標與詞頻、詞彙出現的文件數相關分析結果

(詞頻權重=0、詞彙出現的文件數權重=1)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.646	0.580	0.378	0.291	0.339	0.325
350	0.651	0.584	0.369	0.282	0.331	0.316
300	0.656	0.587	0.358	0.273	0.322	0.308
250	0.660	0.588	0.347	0.263	0.314	0.299
200	0.663	0.584	0.337	0.253	0.305	0.289
150	0.667	0.575	0.328	0.240	0.294	0.277
100	0.674	0.560	0.313	0.221	0.278	0.259
50	0.697	0.550	0.278	0.183	0.243	0.221

表 15 Log-IDF 詞彙權重各指標與詞頻、詞彙出現的文件數相關分析結果

(詞頻權重=0、詞彙出現的文件數權重=1)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.615	0.557	0.411	0.297	0.347	0.333
350	0.620	0.560	0.401	0.288	0.338	0.323
300	0.624	0.562	0.388	0.278	0.328	0.314
250	0.627	0.561	0.375	0.268	0.319	0.304
200	0.630	0.557	0.362	0.256	0.309	0.293
150	0.634	0.550	0.346	0.242	0.297	0.280
100	0.640	0.534	0.327	0.222	0.280	0.261
50	0.663	0.526	0.282	0.182	0.243	0.221

表 16 TF-IDF 權重各指標與詞頻、詞彙出現的文件數之相關分析結果

(詞頻權重=0、詞彙出現的文件數權重=1)

方法 / 維度	指標一	指標二	指標三	指標四	指標五	指標六
400	0.541	0.489	0.277	0.248	0.317	0.296
350	0.541	0.485	0.270	0.242	0.311	0.289
300	0.540	0.478	0.265	0.235	0.304	0.282
250	0.541	0.471	0.257	0.226	0.297	0.274
200	0.542	0.461	0.250	0.216	0.287	0.264
150	0.546	0.451	0.239	0.201	0.274	0.250
100	0.560	0.447	0.225	0.181	0.254	0.228
50	0.594	0.458	0.192	0.141	0.211	0.183