【19】中華民國　　【12】專利公報　（B）

【11】證書號數：I354904
【45】公告日： 中華民國 100 (2011) 年 12 月 21 日
【51】Int. Cl.：　*G06F17/21　(2006.01)*

發明　　全 7 頁

【54】名　稱：文件自動分類方法、電腦可讀取之儲存媒體及其電腦
METHOD OF CLASSIFYING DOCUMENTS AUTOMATICALLY, COMPUTER READABLE STORAGE MEDIUM, AND COMPUTER THEREOF

【21】申請案號：097102799　　【22】申請日： 中華民國 97 (2008) 年 01 月 25 日
【11】公開編號：200933394　　【43】公開日期： 中華民國 98 (2009) 年 08 月 01 日

【72】發 明 人： 耿筠 (TW) KEN, YUN；吳智鴻 (TW) WU, CHIH HUNG；黃道 (TW) HUANG, TAO
【71】申 請 人： 耿筠　KEN, YUN
臺中市西區忠勤街 348 號 7 樓
吳智鴻　WU, CHIH HUNG
新北市三重區重新路 2 段 14 號 12 樓之 10
黃道　HUANG, TAO
苗栗縣苗栗市高苗里木鐸山 18 鄰 26 號
【74】代 理 人： 陳啟桐；廖和信；張祖康

【56】參考文獻：

TW	469386	TW	542993
TW	571251	TW	I286709
US	7194471B1	US	7289982B2

[57]申請專利範圍

1. 一種文件自動分類方法，係用於一電腦運算出一待分類文字資訊是否屬於一目標領域，該方法包括下列步驟：學習階段：(A)接收複數相關文字資訊與複數非相關文字資訊，其中該複數相關文字資訊係屬於該目標領域，該複數非相關文字資訊係不屬於該目標領域；(B)擷取該複數相關文字資訊與該複數非相關文字資訊之複數單位字詞，並且計算各單位字詞於該複數相關文字資訊與該複數非相關文字資訊之出現次數；(C)產生各單位字詞之一權重值，該權重值係代表各單位字詞與該目標領域之間的相關程度；自動分類階段：(D)接收該待分類文字資訊，擷取該待分類文字資訊之複數單位字詞，並且計算各單位字詞於該待分類文字資訊之出現次數；以及(E)藉由該待分類文字資訊之各單位字詞與各權重值以計算出該待分類文字資訊之一相似度數值。
2. 如申請專利範圍第 1 項所述之文件自動分類方法，其中該電腦係利用下列公式以計算出

$$W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j\left(k(\mathrm{x_i},\mathrm{x_j})\right)\ ;$$

該相似度數值：

$$0\le\alpha_i\le C,\qquad i=1,...,l\ ;$$

$$\sum_{i=1}^{l}\alpha_i y_i=0\ ;$$

$$w^* = \sum_{i=1}^{l} \alpha_i^* y_i x_i \quad ;$$

$$b^* = y_i - w^{*T} x_i \quad ;$$

$$f(x) = sign\left(\sum_{i=1}^{l} y_i \alpha_i^* k(x, x_i) + b^* \right) \quad ;$$

其中，$i=1,...n$，n 代表資料個數，$x_i \in R^d$，x_i 代表輸入的變數向量矩陣，$y_i \in \{1,-1\}$，y_i 代表輸出的變數向量矩陣，$k(x_i,x_j)$代表核心函式(Kernel function)轉換，α_i 代表拉氏乘數(Largrange Multiplier)，w^*表示最佳權重值。

3. 如申請專利範圍第 2 項所述之文件自動分類方法，其中：若該相似度數值為+1，表示該待分類文字資訊屬於該目標領域；若該相似度數值為-1，表示該待分類文字資訊不屬於該目標領域。
4. 如申請專利範圍第 3 項所述之文件自動分類方法，其中 $k(x_i,x_j)$係利用以下公式：

$$k\left(x_i, x_j\right) = \exp\left(-\frac{\left\| x_i - x_j \right\|^2}{2\sigma^2} \right) \quad ;$$

$k(x_i,x_j)=(1+x_i \cdot x_j)^d$；或 $k(x_i, x_j) = x_i^T x_j$ 。

5. 如申請專利範圍第 2 項所述之文件自動分類方法，其中 $k(x_i,x_j)$係利用以下公式：

$$k\left(x_i, x_j\right) = \exp\left(-\frac{\left\| x_i - x_j \right\|^2}{2\sigma^2} \right) \quad ;$$

$k(x_i,x_j)=(1+x_i \cdot x_j)^d$；或 $k(x_i, x_j) = x_i^T x_j$ 。

6. 如申請專利範圍第 1 項所述之文件自動分類方法，其中更包括：(F)設定該待分類文字資訊屬於該目標領域之一門檻數值；其中，若該相似度數值大於或等於該門檻數值，表示該待分類文字資訊屬於該目標領域；若該相似度數值小於該門檻數值，表示該待分類文字資訊不屬於該目標領域。
7. 一種電腦可讀取之儲存媒體，係儲存一自動分類程式，藉由一電腦執行該自動分類程式以運算出一待分類文字資訊是否屬於一目標領域，該儲存媒體包括：一第一程式碼，用以接收複數相關文字資訊與複數非相關文字資訊，其中該複數相關文字資訊係屬於該目標領域，該複數非相關文字資訊係不屬於該目標領域；一第二程式碼，用以擷取該複數相關文字資訊與該複數非相關文字資訊之複數單位字詞，並且計算各單位字詞於該複數相關文字資訊與該複數非相關文字資訊之出現次數；一第三程式碼，用以產生各單位字詞之一權重值，該權重值係代表各單位字詞與該目標領域之間的相關程度；一第四程式碼，用以接收該待分類文字資訊，擷取該待分類文字資訊之複數單位字詞，並且計算各單位字詞於該待分類文字資訊之出現次數；以及一第五程式碼，藉由該待分類文字資訊之各單位字詞與各權重值以計算出該待分類文字資訊之一相似度數值。

8. 如申請專利範圍第 7 項所述之儲存媒體，其中該電腦係 利用下列公式以計算出該相似度

$$W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j\left(k(\mathrm{x_i},\mathrm{x_j})\right)\ ;$$

$$0\le\alpha_i\le C,\qquad i=1,...,l\ ;$$

數值：
$$\sum_{i=1}^{l}\alpha_i y_i=0\ ;$$
$$w^*=\sum_{i=1}^{l}\alpha_i^* y_i x_i\ ;$$
$$b^*=y_i-w^{*T}x_i\ ;$$
$$f(x)=sign\left(\sum_{i=1}^{l}y_i\alpha_i^* k(x,x_i)+b^*\right)\ ;$$

其中，$x_i\in R^d$，$y_i\in\{1,-1\}$，x_i 代表輸入的變數向量矩陣，$i=1,...n$，n 代表資料個數，y_i 代表輸出的變數向量矩陣，d 代表資料的維度，$K(x_i,x_j)$代表核心函式(Kernel function)轉換，α_i 代表拉氏乘數(Largrange Multiplier)，w^*表示最佳權重值。

9. 如申請專利範圍第 8 項所述之儲存媒體，其中：若該相似度數值為+1，表示該待分類文字資訊屬於該目標領域；若該相似度數值為-1，表示該待分類文字資訊不屬於該目標領域。

10. 如申請專利範圍第 9 項所述之儲存媒體，其中 $k(x_i,x_j)$係利用以下公式：

$$k\left(x_i,x_j\right)=\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{2\sigma^2}\right)\ ;$$

$k(x_i,x_j)=(1+x_i\cdot x_j)^d$；或 $k(x_i,x_j)=x_i^T x_j$。

11. 如申請專利範圍第 8 項所述之儲存媒體，其中 $k(x_i,x_j)$係利用以下公式：

$$k\left(x_i,x_j\right)=\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{2\sigma^2}\right)\ ;$$

或 $k(x_i,x_j)=(1+x_i\cdot x_j)^d$；或 $k(x_i,x_j)=x_i^T x_j$。

12. 如申請專利範圍第 7 項所述之儲存媒體，其中更包括：一第六程式碼，用以設定該待分類文字資訊屬於該目標領域之一門檻數值，其中，若該相似度數值大於或等於該門檻數值，表示該待分類文字資訊屬於該目標領域；若該相似度數值小於該門檻數值，表示該待分類文字資訊不屬於該目標領域。

13. 一種電腦，用以運算出一待分類文字資訊是否屬於一目標領域，該電腦包括：一記憶體，該記憶體包括一軟體程式；一處理器，該處理器與該記憶體電性連接，該處理器可執行該軟體程式以達成下列機制：(A)接收複數相關文字資訊與複數非相關文字資訊，其中該複數相關文字資訊係屬於該目標領域，該複數非相關文字資訊係不屬於該目標領域；(B)擷取該複數相關文字資訊與該複數非相關文字資訊之複數單位字詞，並且計算各單位字詞於該複數相關文字資訊與該複數非相關文字資訊之出現次數；(C)產生各單位字詞之一權重值，該權重值係代表各單位字詞與該目標領域之間的相關程度；(D)接收該待分類文字資訊，擷取該待分類文字資訊之複數單位字詞，並且計算各單位字詞於該待分

類文字資訊之出現次數；以及(E)藉由該待分類文字資訊之各單位字詞與各權重值以計算出該待分類文字資訊之一相似度數值。

14. 如申請專利範圍第 13 項所述之電腦，其中該電腦係利用下列公式以計算出該相似度數值：

$$W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j\left(k(\mathrm{x_i},\mathrm{x_j})\right) \quad ;$$

$$0\le\alpha_i\le C,\qquad i=1,...,l \quad ;$$

$$\sum_{i=1}^{l}\alpha_i y_i=0 \quad ;$$

$$w^*=\sum_{i=1}^{l}\alpha_i^* y_i x_i \quad ;$$

$$b^*=y_i-w^{*T}x_i \quad ;$$

$$f(x)=sign\left(\sum_{i=1}^{l}y_i\alpha_i^* k(x,x_i)+b^*\right) \quad ;$$

其中，$x_i\in R^d$，$y_i\in\{1,-1\}$，x_i 代表輸入的變數向量矩陣，$i=1,...n$，n 代表資料個數，y_i 代表輸出的變數向量矩陣，d 代表資料的維度，$K(x_i,x_j)$代表核心函式(Kernel function)轉換，α_i 代表拉氏乘數(Largrange Multiplier)，w^*表示最佳權重值。

15. 如申請專利範圍第 14 項所述之電腦，其中：若該相似度數值為+1，表示該待分類文字資訊屬於該目標領域；若該相似度數值為-1，表示該待分類文字資訊不屬於該目標領域。

16. 如申請專利範圍第 15 項所述之電腦，其中 $k(x_i,x_j)$係利用以下公式：

$$k(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right) \quad ;\ \text{或}$$

$k(x_i,x_i)=(1+x_i\cdot x_j)^d$；或 $k(x_i,x_j)=x_i^T x_j$ 。

17. 如申請專利範圍第 14 項所述之電腦，其中 $k(x_i,x_j)$*係利用以下公式：*

$$k(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right) \quad ;\ \text{或}$$

$k(x_i,x_j)=(1+x_i\cdot x_j)^d$；或 $k(x_i,x_j)=x_i^T x_j$ 。

18. 如申請專利範圍第 13 項所述之電腦，其中更包括：(F)設定該待分類文字資訊屬於該目標領域之一門檻數值；其中，若該相似度數值大於或等於該門檻數值，表示該待分類文字資訊屬於該目標領域；若該相似度數值小於該門檻數值，表示該待分類文字資訊不屬於該目標領域。

圖式簡單說明

圖 1 係本發明之電腦之架構圖。

圖 2 係本發明之文件自動分類之方法之流程圖。

圖 3 係文字矩陣實施例。

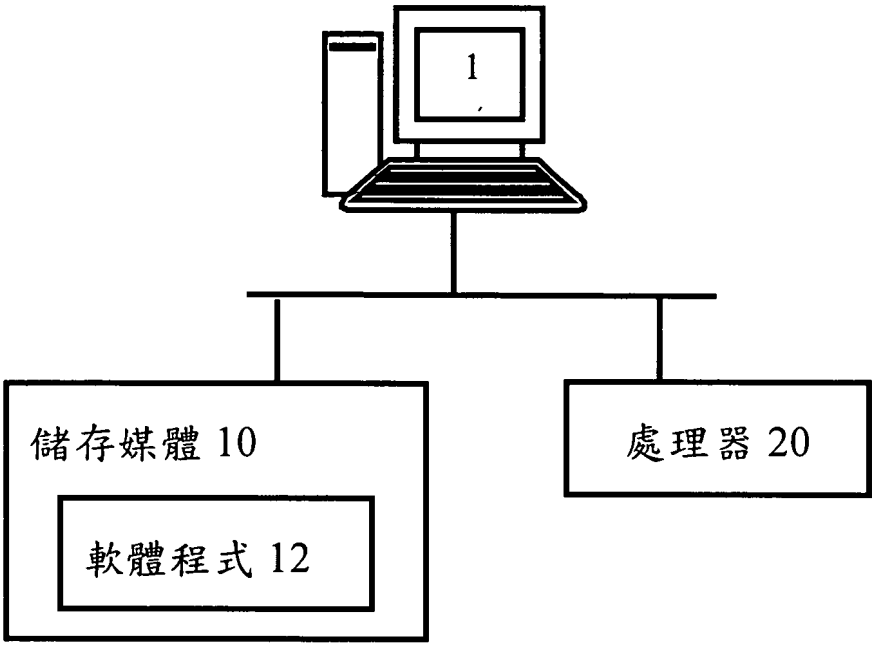


圖 1

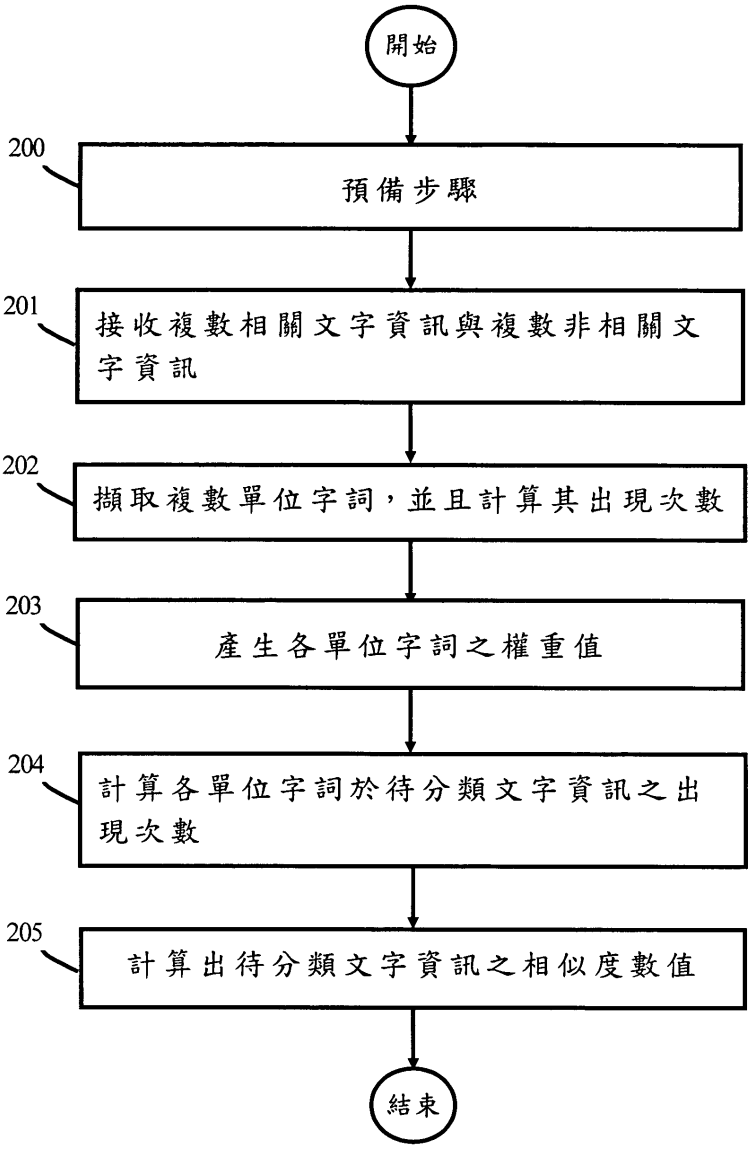
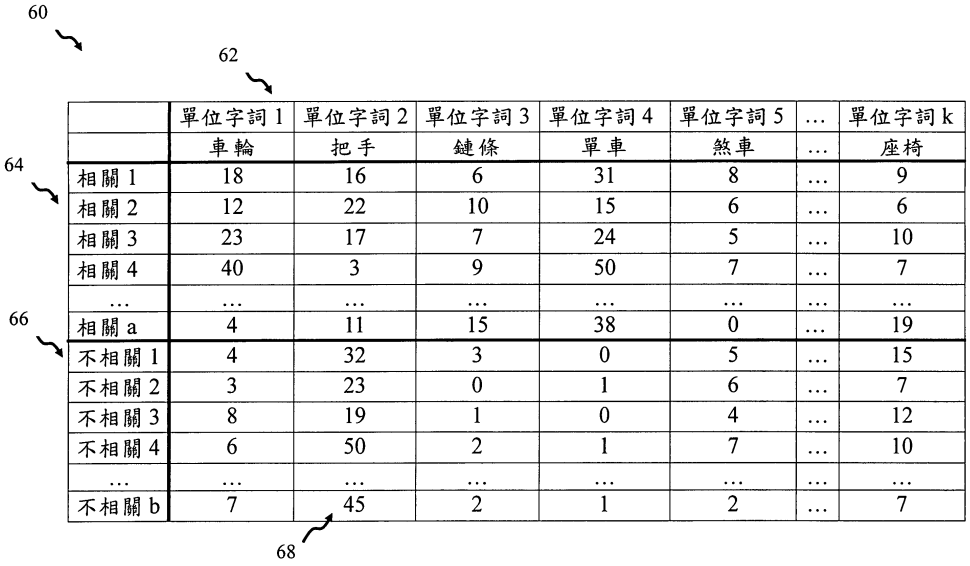

開始
200
預備步驟
201
接收複數相關文字資訊與複數非相關文字資訊
202
擷取複數單位字詞，並且計算其出現次數
203
產生各單位字詞之權重值
204
計算各單位字詞於待分類文字資訊之出現次數
205
計算出待分類文字資訊之相似度數值
結束

圖 2



	單位字詞 1	單位字詞 2	單位字詞 3	單位字詞 4	單位字詞 5	…	單位字詞 k
	車輪	把手	鏈條	單車	煞車	…	座椅
相關 1	18	16	6	31	8	…	9
相關 2	12	22	10	15	6	…	6
相關 3	23	17	7	24	5	…	10
相關 4	40	3	9	50	7	…	7
…	…	…	…	…	…	…	…
相關 a	4	11	15	38	0	…	19
不相關 1	4	32	3	0	5	…	15
不相關 2	3	23	0	1	6	…	7
不相關 3	8	19	1	0	4	…	12
不相關 4	6	50	2	1	7	…	10
…	…	…	…	…	…	…	…
不相關 b	7	45	2	1	2	…	7

圖 3