【19】中華民國　　【12】專利公報　(B)

【11】證書號數：I254880

【45】公告日：中華民國95(2006) 年 5 月 11 日

【51】Int. Cl.[7]：　G06F7/06,　G06F17/27

發明　　全 7 頁

【54】名稱：電子化文件分群解析方法

METHOD FOR CLASSIFYING ELECTRONIC DOCUMENT ANALYSIS

【21】申請案號：093131521　　【22】申請日：中華民國93(2004)年10月18日

【11】公開編號：200614065　　【43】公開日：中華民國95(2006)年5月1日

【72】發明人：許芙瑜 HSU, FU CHIANG；侯建良 HOU, JIANG LIANG；何佩勳 HO, PEI HSUN；張瑞芬 TRAPPEY, AMY J.C.；張力元 TRAPPEY, CHARLES V.；劉尚志 LIU, SHANG JYH

【71】申請人：亞頌科技股份有限公司 AVECTEC. COM, INC.
新竹市埔頂路29號3樓之2

【74】代理人：詹銘文；蕭錫清

1　　2

[57]申請專利範圍：

1.一種電子化文件分群解析方法，包括下列步驟：

取得一電子文件庫內之一電子化文件，該電子化文件係包含多數個重要詞彙；

擷取該電子化文件之內容之該些重要詞彙；

根據該電子化文件中之該些重要詞彙的出現頻率以計算該些重要詞彙間之相關性；以及

根據該些重要詞彙間之相關性，將該些重要詞彙分成至少一技術群組。

2.如申請專利範圍第1項所述之電子化文件分群解析方法，其中擷取該電子化文件中之該些重要詞彙之步驟，包括字節解析、字詞解析、字詞比對、字詞頻率維護、候選詞庫重要詞彙擷取與待確認詞庫之重要詞彙擷取之至少一者。

3.如申請專利範圍第1項所述之電子化文件分群解析方法，其中根據該電子化文件中之該些重要詞彙的出現頻率以計算該些重要詞彙間之相關性之步驟包括：
合併該些重要詞彙中之相同者之出現頻率；以及
計算合併後之該些重要詞彙的出現頻率之相關性。

4.如申請專利範圍第3項所述之電子化文件分群解析方法，其中合併該些重要詞彙中之相同者之出現頻率之步驟包括：
從該電子化文件中取出該些重要詞彙；
整併該些重要詞彙中重複出現之部分；以及
重新計算該些重要詞彙之出現頻率。

5.如申請專利範圍第3項所述之電子化文件分群解析方法，其中計算合併後之該些重要詞彙的出現頻率之相關性之步驟包括：
取得該些重要詞彙之出現頻率；以及
計算該些重要詞彙兩兩間之出現頻率之一相關係數，並以該相關係數為該些重要詞彙之出現頻率之相關性。

6.如申請專利範圍第1項所述之電子化文件分群解析方法，其中將該些重要詞彙分群成該些技術群組之步驟包括：
根據該些重要詞彙間之相關性，利用該些重要詞彙之個數取得相對應之多數個詞彙維度，將該些重要詞彙依照該些詞彙維度形成一詞彙資料，並以該詞彙資料當作分群解析之輸入；以及
利用 K-Means 演算法根據該詞彙資料將該些重要詞彙區隔為該些技術群組。



7.如申請專利範圍第1項所述之電子化文件分群解析方法，更包括根據該電子化文件內之該些重要詞彙的詞彙個數、已分類之技術下之電子化文件個數、與已分類之技術下的技術詞彙個數，可求得此技術群組之成熟度。

8.一種電子化文件分群解析方法，包括：
取得一電子文件庫內之多數個電子化文件，該些電子化文件中之一係包含至少一技術群組；
取得該些電子化文件中之該些技術群組；
合併統計該些電子化文件中之該些技術群組的出現次數；以及
根據已合併統計之該些電子化文件中之該些技術群組的出現次數，將該些電子化文件分群成至少一文件群。

9.如申請專利範圍第8項所述之電子化文件分群解析方法，其中取得該些電子化文件中之該些技術群組之步驟，包括：
擷取該些電子化文件中之多數個重要詞彙；
根據該些電子化文件中之該些重要詞彙的出現頻率以計算該些重要詞彙間之相關性；以及
根據該些重要詞彙間之相關性，將該些重要詞彙分群以取得該些技術群組。

10.如申請專利範圍第9項所述之電子化文件分群解析方法，其中取得該些電子化文件中之該些重要詞彙之步驟，包括字節解析、字詞解析、字詞比對、字詞頻率維護、候選詞庫重要詞彙擷取與待確認詞庫之重

要詞彙擷取之至少一者。

11.如申請專利範圍第 9 項所述之電子化文件分群解析方法，其中根據該電子化文件中之該些重要詞彙的出現頻率以計算該些重要詞彙間之相關性之步驟包括：

合併該些重要詞彙中之相同者之出現頻率；以及

計算合併後之該些重要詞彙的出現頻率之相關性。

12.如申請專利範圍第 11 項所述之電子化文件分群解析方法，其中合併該些重要詞彙中之相同者之出現頻率之步驟包括：

取出該些重要詞彙；

整併該些重要詞彙中重複出現之部分；以及

重新計算該些重要詞彙之出現頻率。

13.如申請專利範圍第 11 項所述之電子化文件分群解析方法，其中計算合併後之該些重要詞彙的出現頻率之相關性的步驟包括：

取得該些重要詞彙之出現頻率；以及

計算該些重要詞彙兩兩間之出現頻率之一相關係數，並以該相關係數為該些重要詞彙之出現頻率之相關性。

14.如申請專利範圍第 9 項所述之電子化文件分群解析方法，其中將該些重要詞彙分群以取得該些技術群組之步驟為：

根據該些重要詞彙間之相關性，利用該些重要詞彙之個數形成相對應之多數個詞彙維度，將該些重要詞彙依照該些詞彙維度形成一詞彙資料，並以該詞彙資料當作分群解析之輸入；以及

利用 K-Means 演算法根據該詞彙資料將該些重要詞彙區隔為該些技術群組。

15.如申請專利範圍第 8 項所述之電子化文件分群解析方法，其中將該些電子化文件分群之步驟為：

根據已統計之該些電子化文件中之該些技術群組的出現次數，利用該些技術群組之個數取得相對應之多數個技術維度，將該些電子化文件依照該些技術維度形成一技術資料，並以該技術資料當作分群解析之輸入；以及

利用 K-Means 演算法根據該技術資料將該些電子化文件分群為多數個之該些文件群。



圖式簡單說明：

圖 1 是習知技術之文件解析方法流程圖。

圖 2 係依照本發明一較佳實施例所繪示的電子化文件分群解析方法之流程圖。

圖 3 是繪示圖 2 中步驟 S203 之詳細流程圖。

圖 4 係依照本發明一較佳實施例所繪示的重要詞彙相關係數表。

圖5與圖6係本發明之一較佳實施例之 K-Means 演算法示意圖。

圖 7 係本發明之一較佳實施例之電子化文件內具有之技術群組統計表。

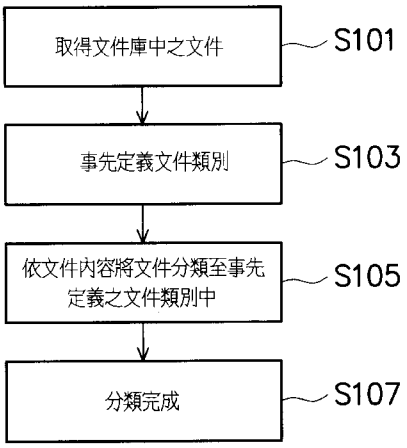
取得文件庫中之文件
S101
事先定義文件類別
S103
依文件內容將文件分類至事先定義之文件類別中
S105
分類完成
S107

圖 1

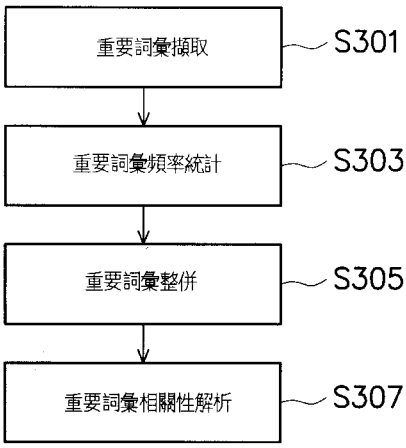
重要詞彙擷取
S301
重要詞彙頻率統計
S303
重要詞彙整併
S305
重要詞彙相關性解析
S307

圖 3

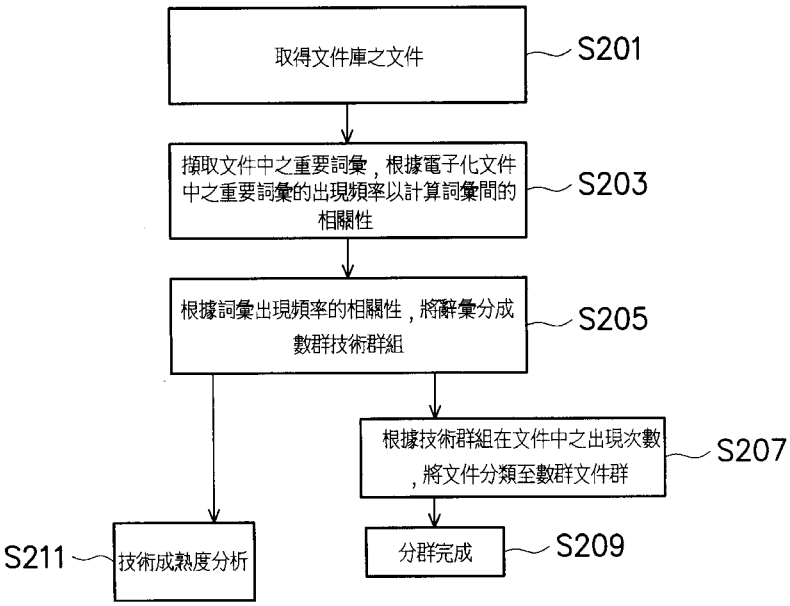
取得文件庫之文件
S201
擷取文件中之重要詞彙，根據電子化文件中之重要詞彙的出現頻率以計算詞彙間的相關性
S203
根據詞彙出現頻率的相關性，將辭彙分成數群技術群組
S205
根據技術群組在文件中之出現次數，將文件分類至數群文件群
S207
分群完成
S209
S211
技術成熟度分析

圖 2

| | 重要詞彙1 | 重要詞彙2 | 重要詞彙3 | 重要詞彙4 | 重要詞彙5 | 重要詞彙6 | 重要詞彙7 | 重要詞彙8 | 重要詞彙9 | 重要詞彙10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 重要詞彙1 | 1 | 0.5 | 0.8 | 0.1 | 0.2 | 0.1 | 0.6 | 0.2 | 0.1 | 0.1 |
| 重要詞彙2 | 0.5 | 1 | 0.7 | 0.1 | 0.3 | 0.2 | 0.7 | 0.1 | 0.1 | 0.3 |
| 重要詞彙3 | 0.8 | 0.7 | 1 | 0.2 | 0.1 | 0.1 | 0.8 | 0.2 | 0.2 | 0.1 |
| 重要詞彙4 | 0.1 | 0.1 | 0.2 | 1 | 0.2 | 0.2 | 0.1 | 0.8 | 0.3 | 0.1 |
| 重要詞彙5 | 0.2 | 0.3 | 0.1 | 0.2 | 1 | 0.1 | 0.2 | 0.7 | 0.1 | 0.2 |
| 重要詞彙6 | 0.1 | 0.3 | 0.1 | 0.2 | 0.1 | 1 | 0.1 | 0.9 | 0.1 | 0.1 |
| 重要詞彙7 | 0.6 | 0.7 | 0.8 | 0.1 | 0.2 | 0.1 | 1 | 0.7 | 0.2 | 0.2 |
| 重要詞彙8 | 0.2 | 0.1 | 0.2 | 0.8 | 0.7 | 0.9 | 0.7 | 1 | 0.1 | 0.1 |
| 重要詞彙9 | 0.1 | 0.1 | 0.2 | 0.3 | 0.1 | 0.1 | 0.2 | 0.1 | 1 | 0.8 |
| 重要詞彙10 | 0.1 | 0.3 | 0.1 | 0.1 | 0.2 | 0.1 | 0.2 | 0.1 | 0.8 | 1 |

圖 4

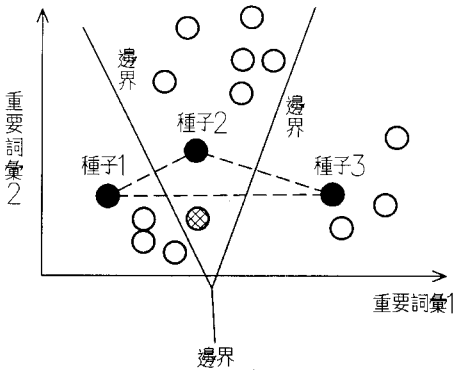
重要詞彙2
邊界
種子2
邊界
種子1
種子3
重要詞彙1
邊界

圖 5

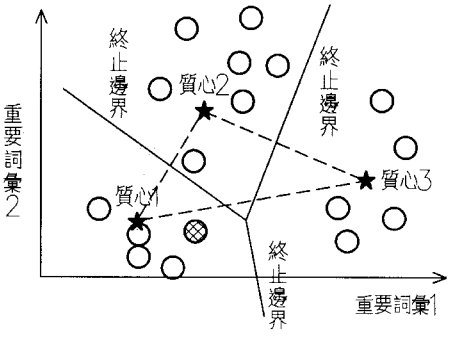
重要詞彙2
終止邊界
質心2
終止邊界
質心3
質心1
終止邊界
重要詞彙1

圖 6

| | 文件1 | 文件2 | 文件3 | 文件4 | 文件5 | 文件6 | 文件7 | 文件8 | 文件9 | 文件10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 技術群組1 | 2 | | 1 | 1 | 1 | | | 1 | 1 | |
| 技術群組2 | | | | 2 | | | | | | |
| 技術群組3 | 1 | | 1 | | 3 | | | 4 | 1 | |
| 技術群組4 | | 2 | | | | | | | | 2 |
| 技術群組5 | | 4 | | | | | | | | |
| 技術群組6 | | | | | | 4 | 1 | | | 1 |
| 技術群組7 | 3 | | | | | | | | | |
| 技術群組8 | | | | | | | | | | 2 |

圖 7