I396990

(19)中華民國智慧財產局

(12)發明說明書公告本 **(11)證書號數：TW I396990B1**

(45)公告日： 中華民國 102 (2013) 年 05 月 21 日

(21)申請案號：098126042 (22)申請日： 中華民國 98 (2009) 年 08 月 03 日

(51)Int. Cl.： ***G06F17/40 (2006.01)***

(71)申請人：國立臺灣科技大學 (中華民國) NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY (TW)
臺北市大安區基隆路 4 段 43 號

(72)發明人：李漢銘 LEE, HAHN MING (TW) ；何建明 HO, JAN MING (TW) ；陳水石 CHEN, SHUI SHI (TW) ；楊凱翔 YANG, KAI HSIANG (TW) ；王瑞遠 WANG, RUEI YUAN (TW) ；葉治宏 YEH, JE ROME (TW)

(74)代理人：洪澄文；顏錦順

(56)參考文獻：

TW I293737 US 2002/0156760A1

US 2006/0210157A1

審查人員：謝進忠

申請專利範圍項數：13 項 圖式數：4 共 0 頁

(54)名稱

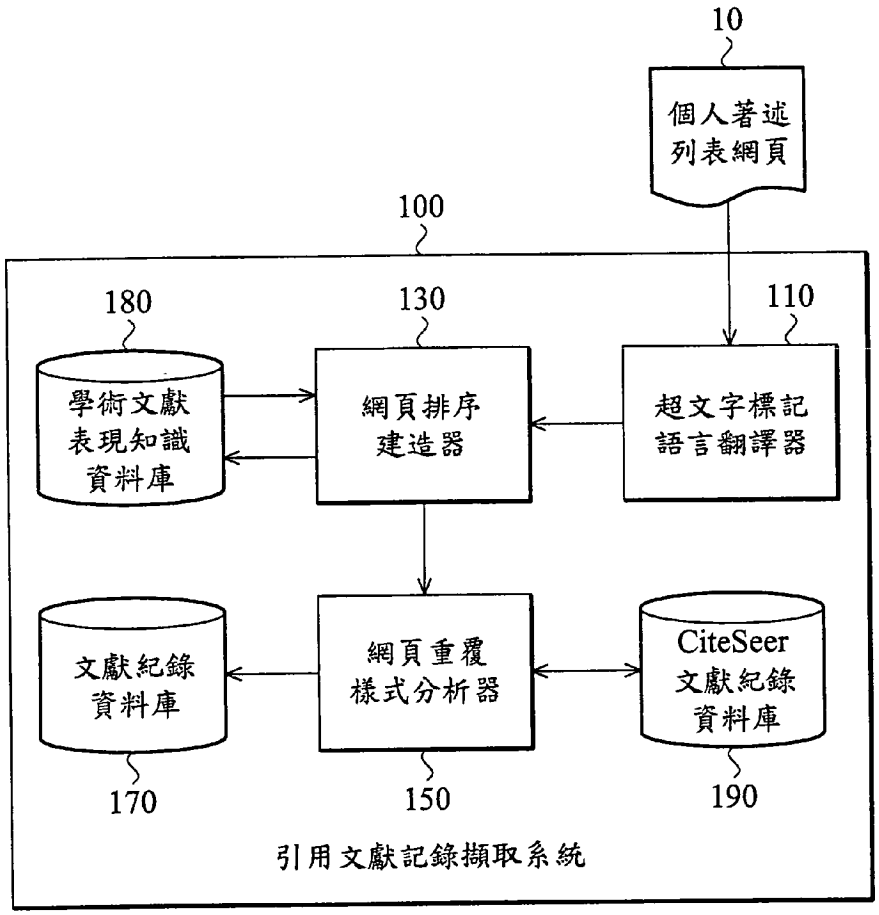
引用文獻記錄擷取系統、方法及程式產品

CITATION RECORD EXTRACTION SYSTEM AND METHOD, AND PROGRAM PRODUCT

(57)摘要

一種引用文獻記錄擷取系統，其包括：超文字標記語言翻譯器，接收一個人著述列表網頁，分析該個人著述列表網頁，以取得該個人著述列表網頁頁面的樣式資訊；一網頁排序建造器，依據該樣式資訊，將該個人著述列表網頁表示成多種記號的特徵序列；及一網頁重覆樣式分析器，分析該特徵序列中重覆的樣式，並過濾掉非引用文獻紀錄，以取得該個人著述列表網頁的一引用文獻記錄。

A Citation Record Extraction System is provided. An HTML rendering engine receives a publication list web page, parses the publication list web page to obtain layout information of the web page. A web page sequence builder generates a web page characteristic sequence for the web page according to the layout information. A web page repeated pattern analyzer analyzes repeated pattern presented in the web page characteristic sequence, screens out non-citation record therefrom, and obtains a citation record of the publication list web page.



第 1 圖

10・・・個人著述列表網頁

100・・・引用文獻記錄擷取系統

110・・・超文字標記語言翻譯器

130・・・網頁排序建造器

150・・・網頁重覆樣式分析器

170・・・文獻紀錄資料庫

180・・・學術文獻表現知識資料庫

190・・・CiteSeer 文獻紀錄資料庫

I396990

公告本

發明專利說明書

（本說明書格式、順序，請勿任意更動，※記號部分請勿填寫）

※ 申請案號：98126042

※ 申請日：98.8.3　※IPC 分類：G06F17/60 (2006.01)

一、發明名稱：(中文/英文)

引用文獻記錄擷取系統、方法及程式產品

Citation Record Extraction System and Method, and Program Product

二、中文發明摘要：

一種引用文獻記錄擷取系統，其包括：超文字標記語言翻譯器，接收一個人著述列表網頁，分析該個人著述列表網頁，以取得該個人著述列表網頁頁面的樣式資訊；一網頁排序建造器，依據該樣式資訊，將該個人著述列表網頁表示成多種記號的特徵序列；及一網頁重覆樣式分析器，分析該特徵序列中重覆的樣式，並過濾掉非引用文獻紀錄，以取得該個人著述列表網頁的一引用文獻記錄。

三、英文發明摘要：

A Citation Record Extraction System is provided. An HTML rendering engine receives a publication list web page, parses the publication list web page to obtain layout information of the web page. A web page sequence builder generates a web page characteristic sequence for the web page according to the layout information. A web page repeated pattern analyzer analyzes repeated pattern presented in the web page characteristic sequence, screens out non-citation record therefrom, and obtains a citation record of the publication list web page.

四、指定代表圖：

(一)本案指定代表圖為：第（1）圖。

(二)本代表圖之元件符號簡單說明：

個人著述列表網頁 10

引用文獻記錄擷取系統 100

超文字標記語言翻譯器 110

網頁排序建造器 130

網頁重覆樣式分析器 150

文獻紀錄資料庫 170

學術文獻表現知識資料庫 180

CiteSeer 文獻紀錄資料庫 190

五、本案若有化學式時，請揭示最能顯示發明特徵的化學式：無。

六、發明說明：

【發明所屬之技術領域】

本發明係有關於一種電腦系統與方法，特別是有關於一種自動擷取引用文獻記錄的系統與方法。

【先前技術】

一般而言，可從學者的個人著述列表網頁(publication list page)中獲得該學者的研究著述紀錄，例如:發表過的學術論文及書籍等。藉由分析學者的研究著述紀錄可延伸出許多應用，例如學術社群分析，利益衝突迴避等。然而大部分的個人著述列表網頁是由學者自行編排設計，因此，不同學者的個人著述列表網頁的格式各不相同。所以如何有效率且正確地從個人著述列表網頁中擷取有用的資訊供分析使用仍然是個深具挑戰性的研究課題。

因此，需要一種自動擷取引用文獻記錄的系統與方法。

【發明內容】

本發明提供一種引用文獻記錄擷取系統，其包括：超文字標記語言翻譯器，接收一個人著述列表網頁，分析該個人著述列表網頁，以取得該個人著述列表網頁頁面的樣式資訊；一網頁排序建造器，依據該樣式資訊，將該個人著述列表網頁表示成多種記號的特徵序列；及一網頁重覆樣式分析器，分析該特徵序列中重覆的樣式，並過濾掉非引用文獻紀錄，以取得該個人著述列表網頁的一引用文獻記錄。

本發明另提供一種引用文獻記錄擷取方法，其包括：接收一個人著述列表網頁；分析該個人著述列表網頁，以取得該個人著述列表網頁頁面的樣式資訊；依據該樣式資訊，將該個人著述列表網頁表示成多種記號的特徵序列；及分析該特徵序列中重覆的樣式，並過濾掉非引用文獻紀錄，以取得該個人著述列表網頁的一引用文獻記錄。

本發明更提供一種程式產品，用以被一機器載入且執行一引用文獻記錄擷取方法，該程式產品包括：一第一程式碼，接收一個人著述列表網頁，分析該個人著述列表網頁，以取得該個人著述列表網頁頁面的樣式資訊；一第二程式碼，依據該樣式資訊，將該個人著述列表網頁表示成多種記號的特徵序列；一第三程式碼，分析該特徵序列中重覆的樣式，並過濾掉非引用文獻紀錄，以取得該個人著述列表網頁的一引用文獻記錄。

為讓本發明之上述和其他目的、特徵、和優點能更明顯易懂，下文特舉出較佳實施例，並配合所附圖式，作詳細說明如下：

【實施方式】

第 1 圖顯示依據本發明實施例之引用文獻記錄擷取系統之方塊圖。

如第 1 圖所示，引用文獻記錄擷取系統 100 主要包括：一超文字標記語言翻譯器 110、一網頁排序建造器 130、一網頁重覆樣式分析器 150、一文獻紀錄資料庫 170、一學術文獻表現知識資料庫 180、及一 CiteSeer 文獻紀錄資料庫 190。

超文字標記語言翻譯器 110 從個人著述列表搜尋系統

(PLF)接收一個人著述列表網頁 10，並分析該個人著述列表網頁，以取得個人著述列表網頁 10 頁面的樣式資訊。

網頁排序建造器 130 接收超文字標記語言翻譯器 110 取得之該樣式資訊，並據以將個人著述列表網頁 10 表示成多種記號的特徵序列，以得知個人著述列表網頁 10 的架構。

第 2 圖顯示第 1 圖中網頁排序建造器的方塊示意圖。

網頁排序建造器 130 包括：超文字標記語言物件表示器 131、基本聚類建造器 133、圖形聚類器 135、網頁排序器 137、及內部聚類相似性估算器 139。

超文字標記語言物件表示器 131 接收超文字標記語言翻譯器 110 取得的該樣式資訊，利用該樣式資訊中的多個屬性來表示該個人著述列表網頁中的物件。

基本聚類建造器 133 參考一學術文獻表現知識資料庫 180，依據文字屬性的特徵，把相似度高的網頁特徵向量聚集成一個聚類，以建造多個範圍聚類。

圖形聚類器 135 接收由基本聚類建造器 133 聚類出來的物件，整合範圍聚類的結果。

網頁排序器 137 依據圖形聚類器 135 整合範圍聚類的結果，把個人著述列表網頁 10 中的特徵表示成一序列的樣式。其中每個樣式表示一個物件，而且一群的物件都有相同的樣式。

內部聚類相似性估算器 139 依據圖形聚類器 135 整合範圍聚類的結果，計算出兩聚類之間的連結量。該連結量係作為在網頁重覆樣式分析器 150 中建立調整分數矩陣時的一個指標。

網頁重覆樣式分析器 150 接收網頁排序建造器 130 產

生之該特徵序列，分析該特徵序列中重覆的樣式，並過濾掉非引用文獻紀錄。

第 3 圖顯示第 1 圖中網頁重覆樣式分析器的方塊示意圖。

網頁重覆樣式分析器 150 包括：調整分數矩陣計算器 151、重覆樣式搜尋器 153、樣式排序器 155、及非引用文獻紀錄過濾器 157。

調整分數矩陣計算器 151 接收內部聚類相似性估算器 139 計算的該連結量，並據以標示出表示交換、配對及裂縫成本的分數矩陣。

重覆樣式搜尋器 153 接收網頁排序器 137 產生的網頁特徵序列，找出該個人著述列表網頁 10 中重覆出現的樣式。

樣式排序器 155 依據該重覆的樣式查詢 CiteSeer 文獻紀錄資料庫 190，以確認其是否為引用文獻紀錄，並依據引用文獻紀錄的文字特徵給決定各樣式集的分數，以求出最高分數的樣式集，分數最高的樣式集便為擷取出來的引用文獻紀錄。

非引用文獻紀錄過濾器 157 以新進來的引用文獻紀錄與正確的引用文獻紀錄進行排序，排序越高者為對的引用文獻紀錄，以決定該新的引用文獻紀錄是否為正確的引用文獻紀錄，進而從個人著述列表網頁 10 中擷取到引用文獻紀錄。

經過網頁重覆樣式分析器 150 過濾處理後的結果，即為從該個人著述列表網頁中擷取出來的引用文獻紀錄。所得到的該引用文獻記錄被儲存在文獻紀錄資料庫 170 中。

第 2 圖顯示依據本發明實施例之引用文獻記錄擷取方

法的流程圖。

步驟 S201 中，接收一個人著述列表網頁。

步驟 S203 中，抽取該個人著述列表網頁的片段，以得到個人著述列表網頁的物件及編排樣式的資訊。詳言之，其首先分析該個人著述列表網頁，以取得個人著述列表網頁頁面的樣式資訊，並據以將個人著述列表網頁表示成多種記號的特徵序列，以得知個人著述列表網頁的架構。

步驟 S205 中，將相似度高的網頁片斷聚類並標示為叢集。詳言之，其首先依據文字屬性的特徵，把相似度高的網頁特徵向量聚集成一個聚類，以建造多個範圍聚類。

步驟 S207 中，將該個人著述列表網頁中多次重複的網頁片斷組合成文獻紀錄。

步驟 S209 中，過濾出步驟 S207 中得到的文獻紀錄中非屬文獻紀錄的內容。

如上述，本發明實施例之自動化引用文獻紀錄之擷取系統與方法，此系統與方法可解決在現今多樣化的個人著述列表網頁中，因為多種格式的文獻紀錄，使得擷取文獻紀錄時可能需要人力的介入，或者傳統的擷取系統無法做有效的擷取。本系統是依據在個人著述列表網頁中的文獻紀錄排序格式做萃取的動作。

雖然本發明已以較佳實施例揭露如上，然其並非用以限定本發明，任何熟習此技藝者，在不脫離本發明之精神和範圍內，當可作些許之更動與潤飾，因此本發明之保護範圍當視後附之申請專利範圍所界定者為準。

【圖式簡單說明】

第 1 圖顯示依據本發明實施例之引用文獻記錄擷取系統之方塊圖。

第 2 圖顯示第 1 圖中網頁排序建造器的方塊示意圖。

第 3 圖顯示第 1 圖中網頁重覆樣式分析器的方塊示意圖。

第 4 圖顯示依據本發明實施例之引用文獻記錄擷取方法的流程圖。

【主要元件符號說明】

個人著述列表網頁 10

引用文獻記錄擷取系統 100

超文字標記語言翻譯器 110

網頁排序建造器 130

超文字標記語言物件表示器 131

基本聚類建造器 133

基於圖形的聚類器 135

網頁排序器 137

內部聚類相似性估算器 139

網頁重覆樣式分析器 150

調整分數矩陣計算器 151

重覆樣式搜尋器 153

樣式排序器 155

非引用文獻紀錄過濾器 157

文獻紀錄資料庫 170

學術文獻表現知識資料庫 180

CiteSeer 文獻紀錄資料庫 190

七、申請專利範圍：

1.一種引用文獻記錄擷取系統，其包括：

一超文字標記語言翻譯器，接收一個人著述列表網頁，分析該個人著述列表網頁，以取得該個人著述列表網頁頁面的樣式資訊；

一網頁排序建造器，依據該樣式資訊，將該個人著述列表網頁表示成多種記號的特徵序列；及

一網頁重覆樣式分析器，分析該特徵序列中重覆的樣式，並過濾掉非引用文獻紀錄，以取得該個人著述列表網頁的一引用文獻記錄，其中該網頁重覆樣式分析器包括：

一調整分數矩陣計算器，接收該網頁排序建造器計算的一連結量，並據以標示出表示交換、配對及裂縫成本的分數矩陣；

一重覆樣式搜尋器，接收該網頁排序建造器產生的該網頁特徵序列，找出該個人著述列表網頁中重覆出現的樣式；

一樣式排序器，依據該重覆的樣式確認其是否為引用文獻紀錄，並依據引用文獻紀錄的文字特徵給決定各樣式集的分數，以求出最高分數的樣式集；及

一非引用文獻紀錄過濾器，以一新進來的引用文獻紀錄與正確的引用文獻紀錄進行排序，以決定該新進來的引用文獻紀錄是否為正確的引用文獻紀錄，進而從該個人著述列表網頁中擷取到該引用文獻紀錄。

2.如申請專利範圍第 1 項所述之引用文獻記錄擷取系統，其中該網頁排序建造器更包括：

一超文字標記語言物件表示器，利用該樣式資訊中的

多個屬性來表示該個人著述列表網頁中的物件；

一基本聚類建造器，依據上述屬性，把相似度高的網頁特徵向量聚集成一個聚類，以建造多個範圍聚類；

一圖形聚類器，整合該範圍聚類的結果；

一網頁排序器，依據該整合範圍聚類的結果，把該個人著述列表網頁中的特徵表示成一序列的樣式；及

一內部聚類相似性估算器，依據該整合範圍聚類的結果，計算出兩聚類之間的該連結量。

3.如申請專利範圍第 2 項所述之引用文獻記錄擷取系統，其中該網頁排序器表示之每個樣式表示一個物件，而且一群的物件都有相同的樣式。

4.如申請專利範圍第 1 項所述之引用文獻記錄擷取系統，該樣式排序器依據該重覆的樣式查詢一文獻紀錄資料庫，以確認其是否為引用文獻紀錄。

5.如申請專利範圍第 1 項所述之引用文獻記錄擷取系統，更包含一文獻紀錄資料庫，用以儲存得到的該引用文獻記錄。

6.一種引用文獻記錄擷取方法，其包括：

接收一個人著述列表網頁；

分析該個人著述列表網頁，以取得該個人著述列表網頁頁面的樣式資訊；

依據該樣式資訊，將該個人著述列表網頁表示成多種記號的特徵序列；及

分析該特徵序列中重覆的樣式，並過濾掉非引用文獻紀錄，以取得該個人著述列表網頁的一引用文獻記錄，該步驟包括：

接收一連結量，並據以標示出表示交換、配對及裂縫成本的分數矩陣；

接收該網頁特徵序列，找出該個人著述列表網頁中重覆出現的樣式；

依據該重覆的樣式確認其是否為引用文獻紀錄，並依據引用文獻紀錄的文字特徵給決定各樣式集的分數，以求出最高分數的樣式集；及

以一新進來的引用文獻紀錄與正確的引用文獻紀錄進行排序，以決定該新進來的引用文獻紀錄是否為正確的引用文獻紀錄，進而從該個人著述列表網頁中擷取到該引用文獻紀錄。

7.如申請專利範圍第 6 項所述之引用文獻記錄擷取方法，更包括：

利用該樣式資訊中的多個屬性來表示該個人著述列表網頁中的物件；

依據上述屬性，把相似度高的網頁特徵向量聚集成一個聚類，以建造多個範圍聚類；

整合該範圍聚類的結果；

依據該整合範圍聚類的結果，把該個人著述列表網頁中的特徵表示成一序列的樣式；及

依據該整合範圍聚類的結果，計算出兩聚類之間的該連結量。

8.如申請專利範圍第 7 項所述之引用文獻記錄擷取方法，其中每個該樣式表示一個物件，而且一群的物件都有相同的樣式。

9.如申請專利範圍第 6 項所述之引用文獻記錄擷取方

法，依據該重覆的樣式查詢一文獻紀錄資料庫，以確認其是否為引用文獻紀錄。

10.如申請專利範圍第6項所述之引用文獻記錄擷取方法，更將得到的該引用文獻記錄儲存於一文獻紀錄資料庫中。

11.一種程式產品，用以被一機器載入且執行一引用文獻記錄擷取方法，該程式產品包括：

一第一程式碼，接收一個人著述列表網頁，分析該個人著述列表網頁，以取得該個人著述列表網頁頁面的樣式資訊；

一第二程式碼，依據該樣式資訊，將該個人著述列表網頁表示成多種記號的特徵序列；

一第三程式碼，分析該特徵序列中重覆的樣式，並過濾掉非引用文獻紀錄，以取得該個人著述列表網頁的一引用文獻記錄，該第三程式碼載入該機器時更執行下列步驟：

接收一連結量，並據以標示出表示交換、配對及裂縫成本的分數矩陣；

接收該網頁特徵序列，找出該個人著述列表網頁中重覆出現的樣式；

依據該重覆的樣式確認其是否為引用文獻紀錄，並依據引用文獻紀錄的文字特徵給決定各樣式集的分數，以求出最高分數的樣式集；及

以一新進來的引用文獻紀錄與正確的引用文獻紀錄進行排序，以決定該新進來的引用文獻紀錄是否為正確的引用文獻紀錄，進而從該個人著述列表網頁中擷取到該引用文獻紀錄。

12.如申請專利範圍第 11 項所述之程式產品，該第二程式碼載入該機器時更執行下列步驟：

利用該樣式資訊中的多個屬性來表示該個人著述列表網頁中的物件；

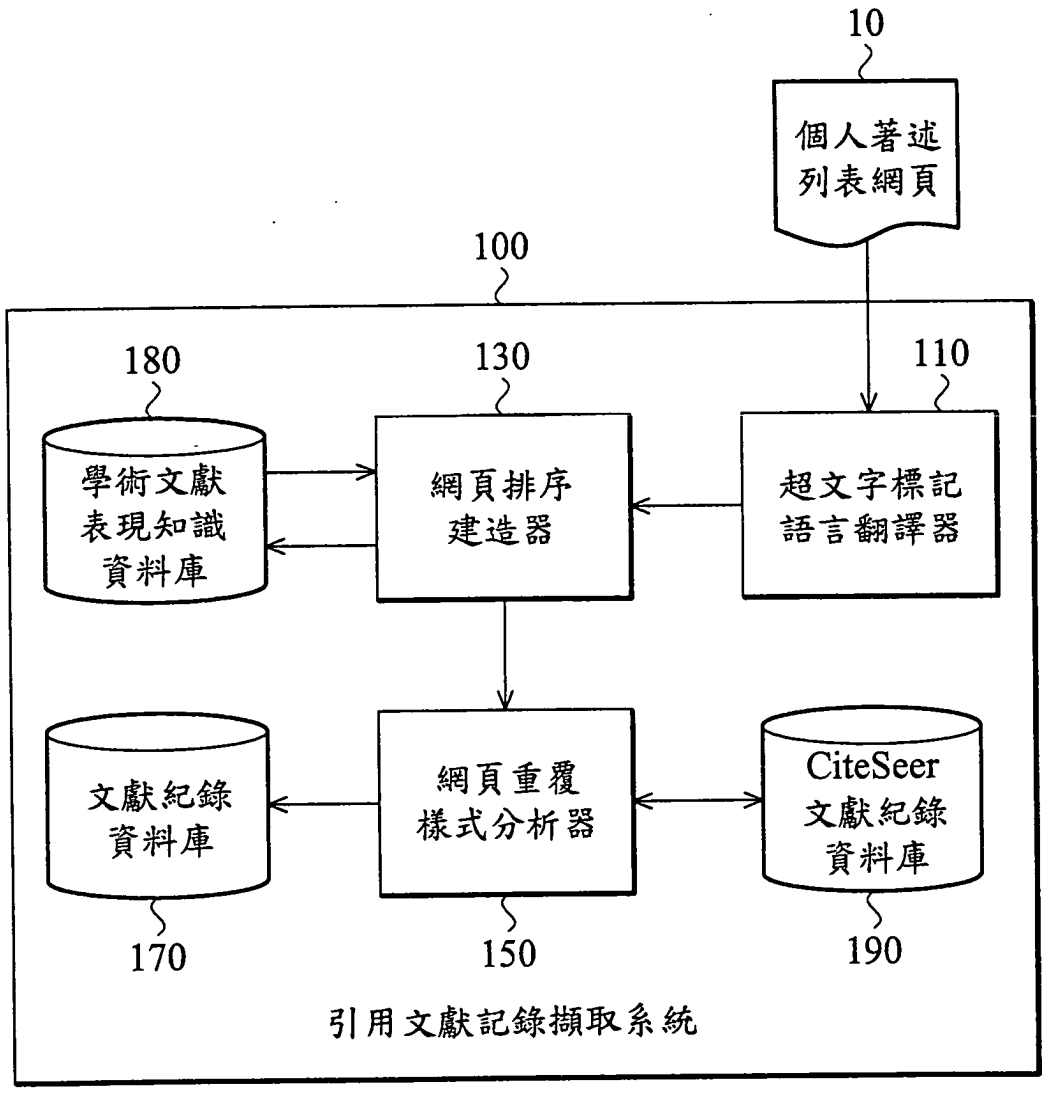
依據上述屬性，把相似度高的網頁特徵向量聚集成一個聚類，以建造多個範圍聚類；

整合該範圍聚類的結果；

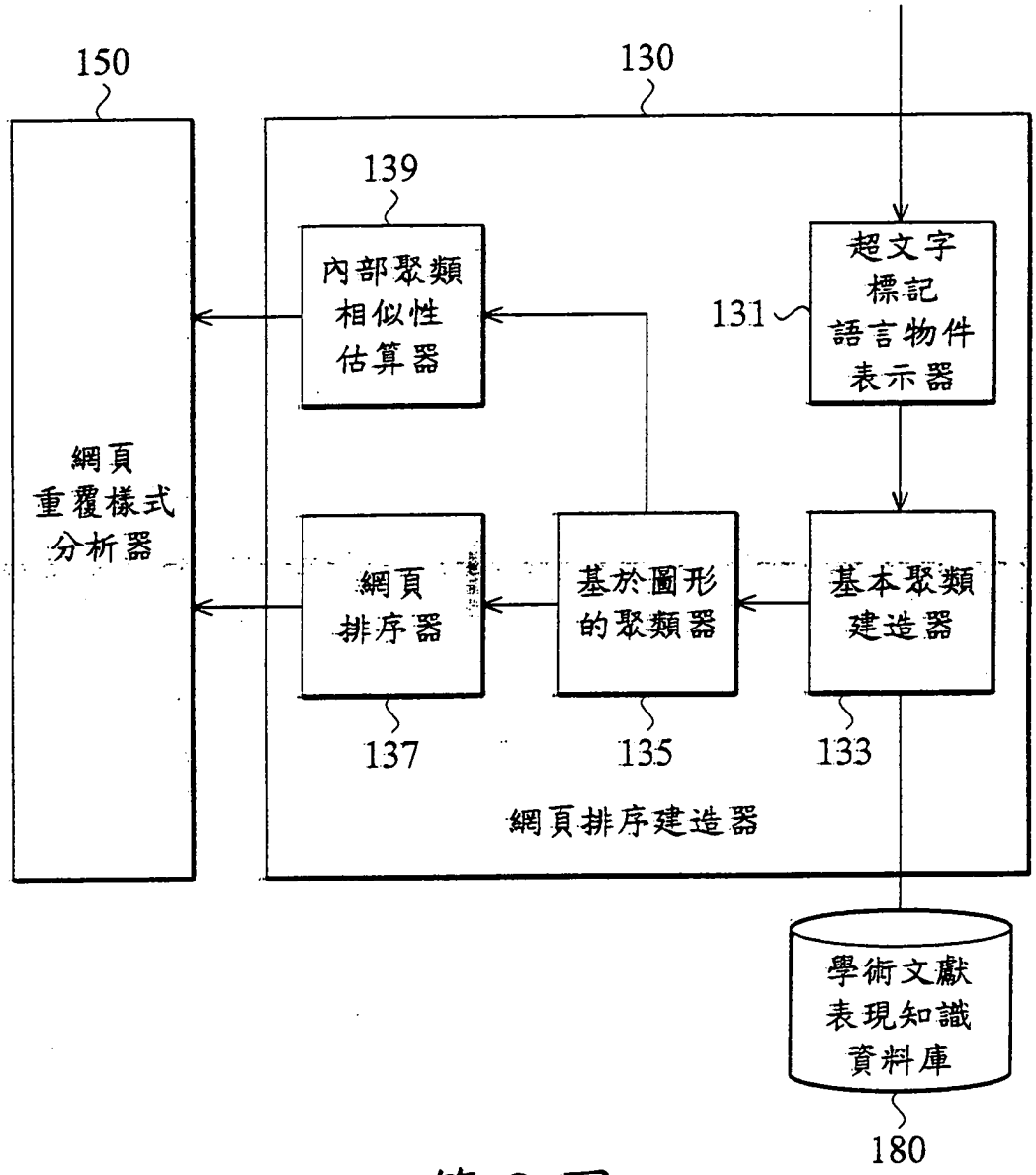
依據該整合範圍聚類的結果，把該個人著述列表網頁中的特徵表示成一序列的樣式；及

依據該整合範圍聚類的結果，計算出兩聚類之間的該連結量。

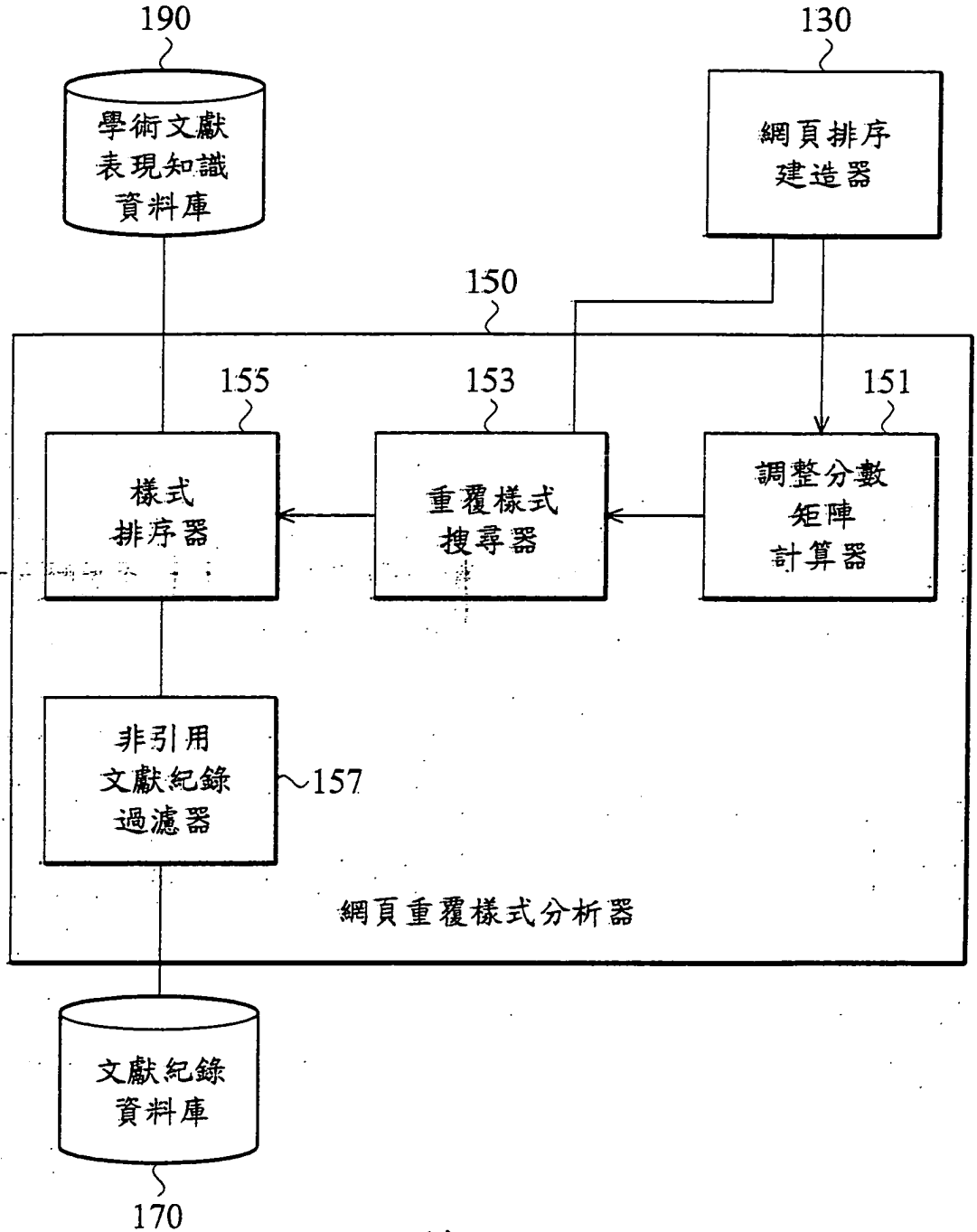
13.如申請專利範圍第 11 項所述之程式產品，該第三程式碼載入該機器時更依據該重覆的樣式查詢一文獻紀錄資料庫，以確認其是否為引用文獻紀錄。



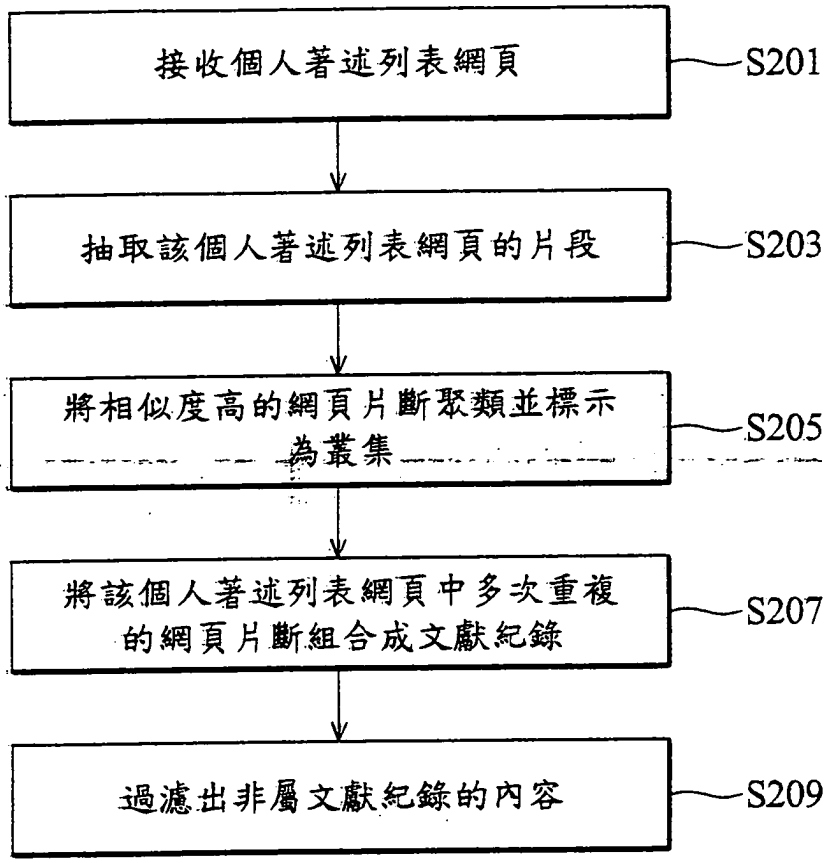
第 1 圖


150
130
139
內部聚類
相似性
估算器
超文字
標記
語言物件
表示器
131
網頁
重覆樣式
分析器
網頁
排序器
基於圖形
的聚類器
基本聚類
建造器
137
135
133
網頁排序建造器
學術文獻
表現知識
資料庫
180

第 2 圖


190
學術文獻
表現知識
資料庫
130
網頁排序
建造器
150
155
153
151
樣式
排序器
重覆樣式
搜尋器
調整分數
矩陣
計算器
非引用
文獻紀錄
過濾器
157
網頁重覆樣式分析器
文獻紀錄
資料庫
170

第 3 圖


接收個人著述列表網頁
S201
抽取該個人著述列表網頁的片段
S203
將相似度高的網頁片斷聚類並標示為叢集
S205
將該個人著述列表網頁中多次重複的網頁片斷組合成文獻紀錄
S207
過濾出非屬文獻紀錄的內容
S209

第 4 圖