I486800

(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 I486800 B**

**(45)公告日： 中華民國 104 (2015) 年 06 月 01 日**

(21)申請案號：098106721 (22)申請日： 中華民國 98 (2009) 年 03 月 02 日

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

(30)優先權：2008/04/11 美國 12/101,951

(71)申請人：微軟公司 (美國) MICROSOFT CORPORATION (US)
美國

(72)發明人：坦克維拉第米爾 TANKOVICH, VLADIMIR (BY) ；李航 LI, HANG (CN) ；米耶戎狄米崔 MEYERZON, DMITRIY (US) ；盧禎 XU, JUN (CN)

(74)代理人：蔡坤財；李世章

(56)參考文獻：

| | | | |
|---|---|---|---|
| TW | 530224 | TW | 575813 |
| TW | I227976 | TW | I284818 |
| US | 2003/0195882A1 | US | 2004/0141354A1 |

審查人員：何旭智

申請專利範圍項數：20 項 圖式數：12 共 51 頁

(54)名稱

用於使用編輯距離以及文件資訊的搜尋結果排序之系統與方法

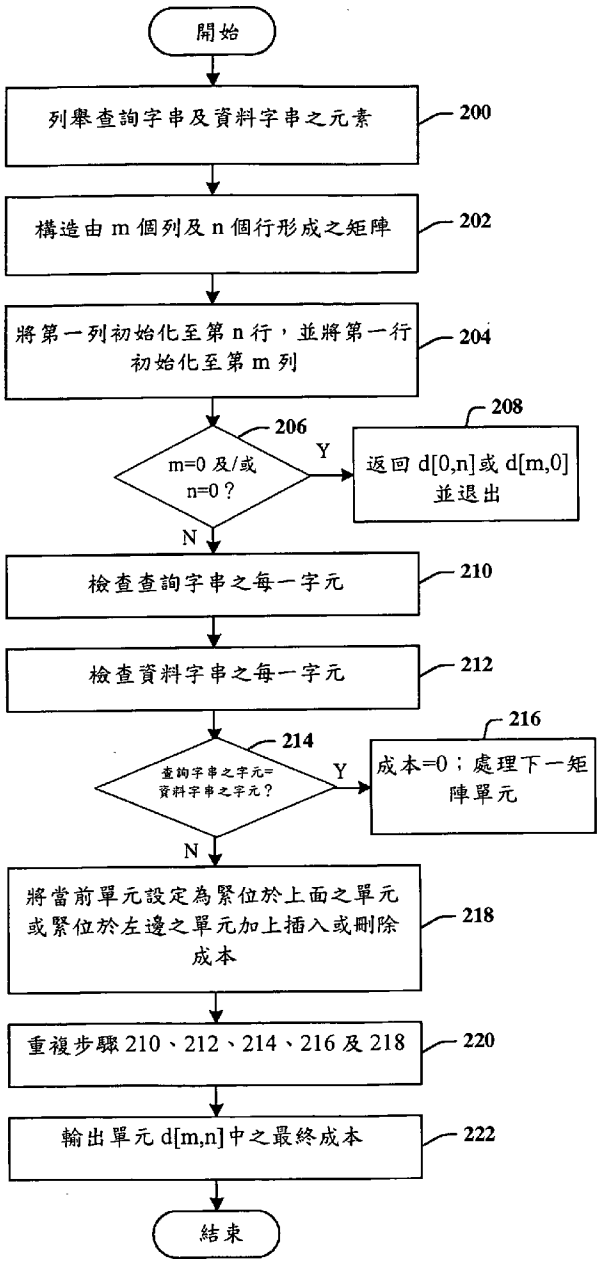
SYSTEM AND METHOD FOR SEARCH RESULTS RANKING USING EDITING DISTANCE AND DOCUMENT INFORMATION

(57)摘要

提供一種架構，用於根據一查詢字串從作為搜尋結果接收之文件中提取文件資訊，並計算資料字串與該查詢字串間之一編輯距離。該編輯距離用於藉由偵測整個查詢或該查詢之一部分之接近匹配項，確定文件之相關性以作為結果排序之一部分。該編輯距離評價該查詢字串與包含例如 TAUC (標題、錨文本、URL、點擊)資訊等文件資訊之一既定資料串流之接近程度。該架構包括對 URL 中之複合詞進行索引時間分離，以便能夠更有效地發現查詢詞。此外，利用對錨文本之索引時間過濾，找到該等文件結果中一或多者之頂部 N 個錨。TAUC 資訊可輸入至一神經網路(例如 2 層式)，以改良用於對搜尋結果排序之相關性量度。

Architecture for extracting document information from documents received as search results based on a query string, and computing an edit distance between the data string and the query string. The edit distance is employed in determining relevance of the document as part of result ranking by detecting near-matches of a whole query or part of the query. The edit distance evaluates how close the query string is to a given data stream that includes document information such as TAUC (title, anchor text, URL, clicks) information, etc. The architecture includes the index-time splitting of compound terms in the URL to allow the more effective discovery of query terms. Additionally, index-time filtering of anchor text is utilized to find the top N anchors of one or more of the document results. The TAUC information can be input to a neural network (e.g., 2-layer) to improve relevance metrics for ranking the search results.

200~222・・・步驟流程




第 2 圖

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；惟已有申請案號者請填寫）

※ 申請案號：98106721

※ 申請日期：2009 年 3 月 2 日　　※IPC 分類：G06F 17/30 (2006.01)

## 一、發明名稱：（中文/英文）

用於使用編輯距離以及文件資訊的搜尋結果排序之系統與方法
SYSTEM AND METHOD FOR SEARCH RESULTS RANKING USING EDITING DISTANCE AND DOCUMENT INFORMATION

## 二、中文發明摘要：

提供一種架構，用於根據一查詢字串從作為搜尋結果接收之文件中提取文件資訊，並計算資料字串與該查詢字串間之一編輯距離。該編輯距離用於藉由偵測整個查詢或該查詢之一部分之接近匹配項，確定文件之相關性以作為結果排序之一部分。該編輯距離評價該查詢字串與包含例如 TAUC（標題、錨文本、URL、點擊）資訊等文件資訊之一既定資料串流之接近程度。該架構包括對 URL 中之複合詞進行索引時間分離，以便能夠更有效地發現查詢詞。此外，利用對錨文本之索引時間過濾，找到該等文件結果中一或多者之頂部 N 個錨。TAUC 資訊可輸入至一神經網路（例如 2 層式），以改良用於對搜尋結果排序之相關性量度。

## 三、英文發明摘要：

Architecture for extracting document information from documents received as search results based on a query string, and computing an edit distance between the data string and the query string. The edit distance is employed in determining relevance of the document as part of result ranking by detecting near-matches of a whole query or part of the query. The edit distance evaluates

how close the query string is to a given data stream that includes document information such as TAUC (title, anchor text, URL, clicks) information, etc. The architecture includes the index-time splitting of compound terms in the URL to allow the more effective discovery of query terms. Additionally, index-time filtering of anchor text is utilized to find the top N anchors of one or more of the document results. The TAUC information can be input to a neural network (e.g., 2-layer) to improve relevance metrics for ranking the search results.

四、指定代表圖：

(一)本案指定代表圖為：第（ 2 ）圖。

(二)本代表圖之元件符號簡單說明：

200~222 步驟流程

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】

本發明係關於使用編輯距離以及文件資訊的搜尋結果排序。

【先前技術】

在一典型之搜尋引擎服務中，使用者可藉由自一經索引之URL（全球資源定位器）集合中選擇與一查詢相匹配之最頂層相關之文件，以進入該查詢。為快速地服務於該等查詢，搜尋引擎利用一或多種方法（例如，一種反轉之索引資料結構）將關鍵字映射對映至文件。舉例而言，引擎所執行之一第一步驟可係為識別一候選文件集合，其中包含有該用戶查詢所指定之關鍵字之候選文件集合。該等關鍵字可位於文件主體或中繼資料，或者實際儲存於其它文件或資料庫中的關於此文件之其它中繼資料（例如錨文本）中。

在大的索引集合中，端視查詢詞之常見性而定，候選文件集合之基數可很大（例如，可能上百萬個）。搜尋引擎並非返回整個候選文件集合，而是執行一第二步驟以根據相關性對候選文件進行排序。通常，搜尋引擎利用一排序函數來預測一文件與一特定查詢之相關程度。該排序函數自該文件中取多個特徵作為輸入，並計算一數字，該數字使搜尋引擎能夠按所預測之相關性對該等文件進行排序。

就該排序函數預測一文件相關性之準確程度而言，該排序函數之品質最終取決於使用者對搜尋結果之滿意度或者平均而言使用者在多少次中找到了對於所提問題之答案。對系統之總體使用者滿意度可由一單一數字（或量度）近似表示，乃因可藉由改變排序函數而使該數字最佳化。通常，針對代表性的一組查詢來計算該等量度，該等查詢係藉由對查詢日誌實施隨機取樣而被選出，且該等量度涉及到對該引擎為各該評價查詢所返回之每一結果分配相關性標記。然而，該等用於文件排序及相關性之方法仍不足以提供所期望之結果。

【發明內容】

以下提供簡要概述，以提供對本文所述某些新穎實施例之基本理解。該概述既非係為詳盡綜述，亦非旨在標識關鍵/緊要元件或限定其範圍。其只是為了以簡要形式提供某些概念以作為下文所作更詳細說明之前序。

該架構提供一種用於根據一查詢字串從作為搜尋結果接收之文件中提取文件資訊並計算一資料字串與該查詢字串間之一編輯距離之機制。該資料字串可係對該文件之簡短、準確描述，該描述係自例如 TAUC（標題、錨文本、URL（全球資源定位器）以及點擊）等文件資訊中獲得。編輯距離用於作為結果排序之一部分來確定文件相關性。該機制藉由採用與接近性相關之特徵集合來偵測整個查詢

或該查詢之一部分之接近匹配項，進而提高了搜尋結果排序之相關性。

處理該編輯距離，以評價該查詢字串與包含文件資訊之一既定資料串流之接近程度。該架構包括對 URL 中之複合詞進行索引時間分離，以便能夠更有效地發現查詢詞。此外，利用對錨文本之索引時間過濾，找到該等文件結果中一或多者之前 N 個錨。利用 TAUC，可將資訊輸入至一神經網路（例如 2 層式），以改良用於對搜尋結果排序之相關性量度。

為達成上述及相關目的，本文結合下文說明及附圖來描述某些例示性態樣。然而，該等態樣係用以說明本文所揭示原理之各種應用方式中之僅僅幾種，並意欲包括所有此等態樣及等價態樣。結合附圖閱讀下文詳細說明，本發明之其它優點及新穎特徵將變得顯而易見。

【實施方式】

本文所揭示之架構藉由實施一與接近性相關之特徵集合以偵測整個查詢之接近匹配項或文件之相關精確中繼資料（例如標題、錨、URL 或點擊）之匹配項，來提高搜尋結果之相關性。舉例而言，考量一查詢“company store”、一第一文件之文件標題“company store online”及一第二文件之文件標題“new NEC LCD monitors in company store”。假設對於第一文件及第二文件而言其它特性皆相同。為使

一所選串流與該查詢匹配，需要進行一些編輯工作，該架構即根據該編輯工作量之大小為文件指派一得分。在本實例中，選擇文件標題進行評價。第一文件之標題僅需要一個刪除操作（刪除“online”一詞）便可實現完全匹配，而第二文件之標題則需要五次刪除（刪除詞“new”、“NEC”、“LCD”、“monitors”及“in”）。因此，計算出第一文件更具相關性。

標題係為 TAUC（標題、錨、URL 及點擊）文件資訊之一元素，為此，可對某些資料串流（例如一 URL）應用處理以便可從複合詞中找到查詢詞。舉例而言，再次考量查詢“company store”，且 URL 為“www.company store.com”。結果，URL 被分成四個部分（或詞）：“www”、“company”、“store”及“com”。

現在參見附圖，其中在所有附圖中，相同參考編號用於指代相同元件。在下文說明中，為便於解釋起見，述及諸多具體細節以提供對本文之透徹理解。然而，顯然，無需該等具體細節亦可實施本發明。在其它情形中，以方塊圖形式顯示眾所習知之結構及裝置，以利於對其進行說明。

第 1 圖例示一由電腦實施之相關性系統 100。系統 100 包括一處理組件 102，用於根據一查詢字串 110 從一作為搜尋結果 108 接收之文件 106 中提取文件資訊 104。系統 100 亦可包括一接近性組件 112，用於計算一從文件資訊 104 導出之資料字串 116 與查詢字串 110 間之編輯距離 114。編輯距離 114 用於確定文件 106 之相關性，以作為搜

尋結果 108 之一部分。

用於產生資料字串 116 之文件資訊 104 可包括例如標題資訊（或字元）、鏈接資訊（例如 URL 字元）、點擊串流資訊、及/或錨文本（或字元）。處理組件 102 在索引時間中將文件資訊 104 之複合詞分離，以計算編輯距離 114。處理組件 102 亦在索引時間中過濾例如錨文本等文件資訊，以計算排在頂部之一錨文本集合。

編輯距離 114 之計算係基於為增加資料字串 116 與查詢字串 110 間之接近性（使其更接近）而進行之詞插入及刪除。編輯距離 114 之計算亦可基於與為增加資料字串 116 與查詢字串 110 間之接近性（使其更接近）而進行之詞插入及刪除相關聯之成本。

考量根據在查詢字串 110 中進行詞插入及/或刪除來產生一資料字串 116（例如 TAUC）之情形。此種詞處理可根據四種操作來執行：將一非查詢字插入查詢字串 110 中；將一查詢詞插入查詢字串 110 中；從查詢字串 110 中刪除一 TAUC 詞；及/或從查詢字串 110 中刪除一非 TAUC 詞。

編輯距離 114 係基於插入及刪除操作，但不基於替換。可存在兩種為插入所定義之成本類型。考量根據查詢字串 110 產生資料字串 116 之情景。在該產生操作中，可將一字插入查詢字串 110 中，若該字存在於原始查詢字串 110 中，則成本被定義為 1；否則，成本被定義為 $w1(\geq 1)$。此處，w1 係為一所調整之加權參數。舉例而言，若查詢字串 110 為 AB，則產生資料字串 ABC 之成本高於產生資料

字串 ABA 之成本。直觀上，在資料字串 116 中插入「不相關之字」會使整個資料字串 116（例如 TAUC）更不相關。

可存在兩種類型之刪除成本。同樣，考量根據查詢字串 110 產生資料字串 116 之情形。當在查詢字串 110 中刪除一詞時，若該詞存在於原始資料字串 116 中，則成本被定義為 1；否則，成本被定義為 w2(≥1)。

另一種類型之成本係位置成本。若在資料字串 116 之第一位置處發生一刪除或插入，則存在一額外成本(+w3)。直觀上，在該等兩字串（查詢字串 110 及資料字串 116）起始處之匹配，其被賦予之重要性高於在該等字串靠後處之匹配。考量以下實例，其中查詢字串 110 係為“cnn”且資料字串 116 係為標題＝“cnn.com–blur blur”。若在第一位置處發生插入及刪除，則其可明顯降低該解決方案之有效性。

第 2 圖例示一種用於計算編輯距離之實例性改良匹配演算法 200 之流程圖。儘管為使解釋簡明起見，將本文例如以流程圖形式顯示之一或多種方法顯示及描述成一系列動作，然而應理解及瞭解，該等方法並不受限於動作之次序，乃因根據該等方法，某些動作之執行次序可不同於本文所示及所述次序及/或與其它動作同時進行。舉例而言，熟習此項技術者將理解及瞭解，一種方法亦可被表示成一系列相互關聯之狀態或事件，例如狀態圖形式。而且，並非在一種方法中所示之所有動作皆係為一種新穎實施方案所必需的。

在 200 處，列舉查詢字串及資料（或目標）字串之元

素。此係藉由將 n 設定為查詢字串（其中查詢字串中每一詞係為 s[i]）之長度並將 m 設定為目標（或資料）字串（其中目標字串中每一詞被標記為 t[j]）之長度來達成。在 202 中，構造一矩陣，該矩陣包含列 0...m 及行 0...n（其中矩陣中每一詞被標記為 d[j,i]）。在 204 處，以一取決於不同刪除成本之值將第一列初始化，並以一取決於不同插入成本之值將第一行初始化。在 206 處，若 n=0，則返回 d[m,0]並退出；若 m=0，則返回 d[0,n]並退出，如在 208 處所示。在 210 處，檢查查詢字串之每一字元（i 從 1 到 n）。在 212 處，檢查目標資料字串之每一字元（j 從 1 到 m）。在 214 處，若查詢字串中之字元字串等於資料字串中之字元，則流程轉至 216，在 216 中，成本為 0 並處理下一矩陣單元。換言之，若 s[i]等於 t[j]，則成本為 0 且 d[j,i]=d[j-1,i-1]。

若查詢字串單元中之字元不等於資料字串單元中之字元，則流程從 214 轉至 218，在 218 中將當前單元設定為緊位於上面之單元或緊位於左邊之單元加上插入或刪除成本。換言之，若 s[i]不等於 t[j]，則將矩陣之單元 d[j,i]設定成等於以下中之最小者：緊位於上面之單元加上相應插入成本（被表示為 d[j-1,i]+cost_insertion）或緊位於左邊之單元加上相應刪除成本（被表示為 d[j,i-1]+cost_deletion）。在 220 處，重複步驟 210、212、214、216 及 218 直到完成。在 222 處，輸出在單元 d[m,n]中所見之最終成本。應注意，該實例中 cost_insertion 及 cost_deletion 二者皆具有兩種值；舉例而言，對於插入成

本，w1=1，w3=4，而對於刪除成本，w2=1，w4=26。

換言之，d[j,i]包含字串 s[0...i]與 t[0...j]間之編輯距離。按定義，d[0,0]=0（不需要進行編輯便會使一空字串等於空字串）。d[0,y]=d[0,y-1]+(w2 或 w4)。若已知形成字串 d[0,y-1]使用了多少次編輯，則可將 d[0,y]計算為 d[0,y-1]+從目標字串刪除當前字元之成本，該成本可為 w2 或 w4。若在 s[0...n]與 t[0...m]二者中皆存在當前字元，則使用成本 w2，否則使用 w4。d[x,0]=d[x-1,0]+（w1 或 w3）。若已知形成字串 d[x-1,0]使用了多少次編輯，則可將 d[x,0]計算為 d[x-1,0]+將當前字元從 s 插入 t 中之成本，該成本可為 w1 或 w3。若在 s[0...n]與 t[0...m]二者中皆存在當前字元，則使用成本 w1，否則使用 w3。

對於每一(j,i)，若 s[i]=t[j]，則 d[j,i]可等於 d[j-1,i-1]。可計算字串 t[j-1]與 s[i-1]間之編輯距離，且若 s[i]=t[j]，則可對該等兩字串附加一共用字元以使該等字串相等而不造成編輯。因此，有三次移動被採用，其中選取為當前 d[j,i]提供最小編輯距離之移動。換言之，

d[j,i] = min(
d[j-1,i-1]，if s[i]=t[j];
d[j-1,i] + (w1，if s[j]同時存在於該等兩字串中；else，w3);
d[j,i-1] + (w2，if t[i]同時存在於該等兩字串中；else，w4)
)

第 3 圖例示利用改良之編輯距離及匹配演算法，根據一查詢字串及資料字串處理及產生編輯距離。該過程涉及左右計算、上下計算及對角線計算中之一或多種。對照一

目標詞資料字串“CBAX”(其中X表示一詞不存在於查詢字串中)處理一查詢詞字串“ABC”。計算一編輯距離之過程可按不同方式執行；然而，用於執行編輯距離之改良版本之具體細節係不同的，如根據所揭示之架構所計算。根據n×m構造一4×5矩陣300，其中查詢字串之n=3且資料字串之m=4。查詢字串302沿矩陣300之橫軸放置，目標資料字串304則沿矩陣300之豎軸放置。

本說明將使用被標記為具有四個行(0-3)及五個列(0-4)之矩陣300。藉由從列0、行0開始從左到右地應用第2圖所述之編輯距離匹配演算法，交叉單元d[0,0]接收到“0”，乃因將查詢字串ABC之空單元與目標資料字串CBAX之空單元相比較並不會導致為使查詢字串與目標資料字串相同而造成詞之插入或刪除。該等「詞」係為相同的，故編輯距離為0。

向右移動以將查詢字串302之詞A與列0之空單元相比較係使用一次刪除來使該等字串相同；因此，單元d[0,1]接收到值“1”。再次向右移動至行2，現在在查詢字串302之詞AB與目標資料字串行之空單元之間進行比較。因此，在查詢字串302中使用兩次刪除來使該等字串相同，從而使編輯距離“2”被置入單元d[0,2]。相同之過程亦適用於行3，其中將查詢字串302之詞ABC與目標字串行中之空單元相比較，利用三次刪除來使該等字串相同，從而使編輯距離“3”被置入單元d[0,3]。

向下到達列1並繼續從左到右，將查詢字串列之空單

元與目標資料字串 304 之第一詞 C 相比較。使用一次刪除來使該等字串相同，故 d[1,0]中之編輯距離為“1”。向右移動至行 1，在查詢字串 302 之詞 A 與目標資料字串 304 之詞 C 之間進行比較。使用一次刪除及一次插入來使該等字串相同，故值“2”被插入單元 d[1,1]中。跳至最末單元 d[1,3]，使 ABC 與 C 相匹配之匹配過程結果係使用兩次刪除，使單元 d[1,3]中之編輯距離為“2”。為簡明起見並且為找到總編輯距離，移動至列 4 及行 3，使詞 ABC 與詞 CBAX 匹配會在單元 d[4,3]中得到一編輯距離“8”：在目標字串之第一詞 C 中使用插入/刪除而得到值“2”，詞 B 之間相匹配而得到值“0”，為使詞 C 與 A 相匹配而進行插入/刪除，從而得到值“2”，插入詞 X 而得到值“1”，以及因位置成本而得到值“3”，結果在單元 d[4,3]中得到一最終編輯距離值“8”。

第 4 圖例示利用改良之編輯距離及匹配演算法，根據一查詢字串及目標資料字串處理及產生編輯距離值之另一實例。此處，產生一矩陣 400，以根據以下權重將一查詢字串 402(ABC)與一目標資料字串 404(AB)相比較：插入成本 cost_insertion 之權重為 w1=1 及 w3=4，刪除成本之權重為 w2=1 及 w4=26。換言之，在列 0 中從左到右，使查詢字串之詞 A 與目標字串 404 前面之空單元相匹配之結果係在目標字串 404 中進行一次插入（詞 A），從而得到單元 d[0,1]之值為“1”。將查詢字串 402 之詞 AB 與目標字串 404 前面之空單元相匹配之結果係在目標字串 404 中進行兩次

插入（詞 AB），從而得到單元 d[0,2]之值為“2”，且將查詢字串 402 之詞 ABC 與目標字串 404 前面之空單元相匹配之結果係在目標字串 404 中進行兩次插入（詞 AB）加上詞 C 之值 w4=26，從而得到單元 d[0,3]之值為“28”，乃因詞 C 不同時在該等兩字串中。

在列 1 中從左到右（應理解，d[1,0]=1），使查詢字串 402 之詞 A 與目標字串 404 之詞 A 相匹配之結果係目標字串 404 與查詢字串 402 相等，從而得到單元 d[1,1]之值為“0”，此係藉由自 d[j-1,i-1]=d[0,0]=“0”取值而得到。使查詢字串 402 之詞 AB 與目標字串 404 之詞 A 相匹配之結果係在目標字串 404 中插入一次詞 B 而得到單元 d[1,2]中之最小值“1”。對於單元 d[1,3]，使查詢字串 402 之詞 ABC 與目標字串 404 之詞 A 相匹配之結果得到一最小值，該最小值係與 d[j-1,i]=d[0,3]之值加上 w3 而在 d[1,3]中得到之值“28”和 d[j,i-1]=d[1,2]之值 1 加上 26 所得到之值 27 進行比較相關聯，因詞 C 不同時處於該等兩字串中，故在 d[1,3]中得到最小值“27”。

在列 2 中從左到右，使查詢字串 402 之詞 A 與目標字串 404 之詞 AB 相匹配之結果係在目標字串 404 中進行一次刪除而在單元 d[2,1]中得到值“1”。對於單元 d[2,2]中之距離，使查詢字串 402 之詞 AB 與目標字串 404 之詞 AB 相匹配之結果係相等，從而從 d[j-1,i-1]=d[1,1]得到該值作為單元 d[2,2]之值“0”。對於單元 d[2,3]，使查詢字串 402 之詞 ABC 與目標字串 404 之詞 AB 相匹配之結果得到一最小

值，該最小值係與 d[j-1,i]=d[1,3]=27 之值加上 w3=1 而得到之值“28”和(因詞 C 並非在目標字串內，亦基於)d[i,j-1]=d[2,2]=0 之值加上 26 所得到之值 26 進行比較相關聯，因詞 C 不同時處於該等兩字串中，故在 d[2,3]中得到最小值“26”。

第 5 圖例示一由電腦實施之相關性系統 500，其採用一神經網路 502 來幫助產生文件 106 之相關性得分 504。系統 500 包括處理組件 102 和接近性組件 112，處理組件 102 用於根據查詢字串 110，從作為搜尋結果 108 接收之文件 106 中提取文件資訊 104，接近性組件 112 則用於計算從文件資訊 104 導出之資料字串 116 與查詢字串 110 間之編輯距離 114。編輯距離 114 用於確定文件 106 之相關性，以作為搜尋結果 108 之一部分。

神經網路 502 可用於接收文件資訊 104 作為一輸入，以用於計算文件 106 之一相關性得分。可僅根據或部分地根據某些或所有搜尋結果 108 之相關性得分，對搜尋結果 108 中之文件排序。系統 500 採用神經網路 502 及代碼庫來產生用於在搜尋結果 108 中對相關文件進行排序之相關性得分。

以下係對該編輯距離演算法之說明，該編輯距離演算法用於計算查詢字串與各該資料字串間之編輯距離以獲得每一對之 TAUC 得分。

因在一文件中僅存在一個標題，故可按下式根據標題來計算 TAUC 得分：

$$TAUC(Title)=ED(Title)$$

其中 *TAUC(Title)* 在應用一變換函數後隨後用作神經網路之一輸入，且 *ED(Title)* 係為該標題之編輯距離。

可存在一文件之多個錨文本的情況，以及多個 URL 及點擊（其中點擊係為使該文件得到點擊的一先前執行之查詢）的情況。想法係為，對於類似之查詢，該文件更為相關。在索引時間中，選取具有最高頻率之 N 個錨文本。然後，針對每一所選錨計算 ED 得分。最後，按下式確定錨之 TAUC 得分：

$$TAUC\ (Anchor) = Min\{ED(Anchor_i)\} \quad i\text{：頂部的 N 個錨；}$$

直觀地，若對於其中一個錨存在良好之匹配，便足以滿足要求。*TAUC (Anchor)* 係在應用一變換函數後用作一神經網路輸入。

在計算 ED 之前，對 URL 字串利用特殊處理。在索引時間中，使用一字元集合作為分離符將 URL 字串分離成多個部分。然後，在每一部分中找到來自一標題及錨詞字典之詞。將字典中之詞之每一次出現儲存於一索引中，該索引之位置係以距 URL 字串起始處之字元數進行量測。

在查詢時間中，從在索引時間中所儲存之索引讀取查詢詞之每一次出現，並以「非查詢」詞填充斷點。在該處理之後，計算 ED。在應用一變換函數之後，ED 處理之結果係為一神經網路輸入。

可被處理之另一特性係為使用者針對一既定文件內容所

輸入之「點擊」次數。每當一使用者點擊該文件時，皆將一串流輸入一資料庫並使其與該文件相關聯。該過程亦可應用於文件資訊文本中之串流資料，例如短的資料串流。

索引時間 URL 處理演算法利用一字元集合作為分離符，將整個 URL 分離成多個部分。該分離功能亦將 urlpart.startpos 設定為源字串中一部分之位置。該分離功能執行對 URL 各無用部分之過濾。

舉例而言，“http://www.companymeeting.com/index.html”被過濾成“companymeeting/index”並被分離成“companymeeting”和“index”。

Startpos: 0

```
Urlparts = split(url, dictionary)

// find terms in different url parts.
For each (term in dictionary)
{
    Int pos = 0;
    For each (urlpart in urlparts)
    {
        pos = urlpart.Find(term, pos);
        while (pos >= 0)
        {
                // parts_separator is used to distinguish different parts at query
time
            storeOccurrence(term, pos +
urlpart.startpos*parts_separator);
            pos = url.Find(term, pos + term.length);
        }
    }
    setIndexStreamLength(parts_separator * urlparts.Count);
```

}

假設字典包含“company meeting comp”，可產生以下關鍵字：Company: 0；Meeting: 7；及 Comp: 0。該字串之總長度係為 parts_separator*2。

對於在 ED 之前之查詢時間處理，在查詢時間中，讀取查詢詞之出現，按在源 URL 中之出現次序來構造一查詢詞字串，並以「非查詢」字標記填充各詞間之空間。舉例而言，考量一查詢字串“company policy”及一結果字串“company”「非查詢詞」「非查詢詞」。

確定一 parts_separator、多個查詢詞位置、及串流長度，以得知在原始 URL 字串中具有多少個部分以及哪一部分包含一既定查詢。每一無查詢詞之部分皆被視為包含一「非查詢詞」。若某一部分不以一查詢詞開頭，則在該詞之前插入一「非查詢詞」。各查詢詞間之所有空間皆被填充以「非查詢詞」。

第 6 圖例示可在文件資訊 104 中用於確定查詢字串與資料字串間之編輯距離之資料類型。文件資訊 104 可包括 TAUC 資料 602，例如標題文本 604、錨文本 606、URL 608 文本或字元、以及點擊資訊 610，以供處理組件 102 處理並用於產生資料（或目標）字串 116。文件資訊 104 亦可包括與一使用者點擊文件內容之次數、使用者（藉由點擊）所選之內容類型、點擊內容之次數、文件總體等相關之點擊資訊 610。

第 7 圖例示一索引時間處理資料流 700。在頂部，根

據文件分析及提取，接收呈標題 604、文件錨 606、點擊資訊 610 等形式之文件資訊。藉由一詞分離演算法 704 處理標題 604，然後將標題 604 提供至一字典 706。字典 706 係對在標題 604、錨 606、點擊資訊 610 等中所見之不同詞之暫時儲存。字典 706 用以藉由一 URL 分離演算法 708 分離 URL 608。URL 分離演算法 708 之輸出被發送至一索引過程 710 以用於相關性及排序處理。文件錨 606 亦可經一用於頂部 N 個錨之過濾器 712 處理。點擊資訊 610 可直接透過該索引過程 710 處理。可相應地處理其它文件資訊（例如詞分離、過濾等）。

第 8 圖例示一方塊圖 800，其顯示來自第 7 圖所示索引過程 710 的用於結果排序之神經網路輸入。索引過程 710 可用於計算相對於查詢字串 110 之一 URL 編輯距離(ED)802、相對於查詢字串 110 之一頂部 N 個錨之 ED、相對於查詢字串 110 之一標題 ED 806、相對於查詢字串 110 之一點擊 ED 808、以及其它與編輯距離無關之特徵 810，其中之某些或全部（URL ED 802、頂部 N 個錨 ED 804、標題 ED 806、點擊 ED 808 及其它特徵 810）可用作神經網路 502 之輸入，以最終找到該相關聯文件之相關性得分並隨後得到該文件在其它文件搜尋結果中之排序。神經網路 502 可係為一 2 層式模型，其接收至少該等 TAUC 特徵作為原始輸入特徵來用於識別文件之相關性。該神經網路確定如何將該等特徵組合成一單一數字，該單一數字可供搜索引擎用於分類。

應瞭解，神經網路 502 僅係為可用於相關性及排序處理之數學或計算模型之一實例。亦可採用其它形式之統計回歸（例如 naïve Bayes、Bayesian 網路、決策樹、模糊邏輯模型），並可採用其它代表不同獨立性圖案之統計分類模型，其中分類係包含用於分配等級及/或優先權之方法。

第 9 圖例示神經網路 502、編輯距離輸入及用於計算和產生搜尋結果之原始特徵輸入之實例性系統 900 實施方式。神經網路 502 之該（等）輸入上之原始排序特徵集合 810 可包括一 BM25 函數 902（例如 BM25F）、點擊距離 904、URL 深度 906、文件類型 908，及語言匹配 910。BM25 組件可例如包括主體、標題、作者、錨文本、URL 顯示器名稱，及所提取之標題。

第 10 圖例示一種用於確定相關性之方法。在 1000 處，接收一查詢字串，該查詢字串係作為一搜尋過程之一部分。在 1002 處，從在搜尋過程期間返回之一文件中提取文件資訊。在 1004 處，根據文件資訊產生一資料字串。在 1006 處，計算資料字串與查詢字串間之編輯距離。在 1008 處，根據編輯距離計算一相關性得分。

該方法之其它態樣可包括採用詞插入作為計算編輯距離之一部分，以及評估為產生資料字串而將一詞插入查詢字串之插入成本，該成本被表示為一加權參數。該方法可更包括採用詞刪除作為計算編輯距離之一部分，以及評估為產生資料字串而刪除查詢字串中之一詞之刪除成本，該成本被表示為一加權參數。可計算一位置成本，該位置成

本作為計算編輯距離之一部分，該位置成本與資料字串中一詞位置之詞插入及/或詞刪除相關聯。另外，在資料字串之字元與查詢字串之字元之間執行一匹配過程，以計算編輯距離之總體計算成本。

資料字串之URL之複合詞分離可發生於索引時間中。該方法可更包括根據在文件中之出現頻率，過濾資料字串之錨文本以找到排在頂部之一錨文本集合，並計算該集合中錨文本之編輯距離得分。在應用一變換函數之後，藉由計算編輯距離所導出之編輯距離得分可被輸入至一兩層式神經網路，該得分係根據計算與以下中至少一者相關聯之編輯距離而產生：標題資訊，錨資訊，點擊資訊，或URL資訊。

第11圖例示一種計算文件相關性之方法。在1100處，處理一查詢字串，該查詢字串作為一搜尋過程之一部分，以返回一文件結果集合。在1102處，根據從該結果集合中之一文件所提取之文件資訊，產生一資料字串，該文件資訊包括以下中之一或多者：來自該文件之標題資訊、錨文本資訊、點擊資訊、及URL資訊。在1104處，根據詞插入、詞刪除及詞位置，計算資料字串與查詢字串間之編輯距離。在1106處，根據編輯距離計算一相關性得分，該相關性得分用於在結果集合中對該文件進行排序。

該方法可更包括計算與詞插入、詞刪除及詞位置中之每一者相關聯之成本、將該成本之因素納入相關性得分之計算、在索引時間中分離URL資訊之複合詞，以及在索引

時間中根據錨文本在文件中出現之頻率過濾錨文本資訊以找到排在頂部之錨文本集合。可執行對查詢字串中各詞之出現之讀取，以按在一源 URL 字串中之出現次序構造一查詢詞字串並以字標記填充該等詞間之空間。

本申請案中所用之措詞「組件」及「系統」旨在指代一與計算機相關之實體，其可係為硬體、硬體與軟體之組合、軟體，或執行中之軟體。舉例而言，一組件可係為但不限於：在一處理器上運行之一過程、一處理器、一影碟機、多個（光學及/或磁性儲存媒體之）儲存碟、一對象、一可執行檔、一執行緒、一程式、及/或一電腦。作為例示，在一伺服器上運行之程式與該伺服器二者皆可係為一組件。一或多個組件可駐存於一過程及/或執行緒內，並且一組件可侷限於一個電腦上及/或分佈於二或更多個電腦上。

現在參見第 12 圖，其例示一計算系統 1200 之方塊圖，該計算系統可操作以根據所揭示之架構，利用 TAUC 特徵執行編輯距離處理以進行搜尋結果排序。為對其不同態樣提供額外之語境，第 12 圖及下文之論述旨在對一適宜計算系統 1200 提供簡要之大體說明，在該計算系統 1200 中可執行該等不同態樣。儘管以上說明係在可在一或多個電腦上運行之電腦可執行指令之大體語境中進行，然而熟習此項技術者將知，亦可與其它程式模組相組合地實施一新穎實施例及/或將該新穎實施例實施為硬體與軟體之組合。

一般而言，程式模組包括用於執行特定任務或實施特定抽象資料類型之常式、程式、組件、資料結構等等。而

且，熟習此項技術者將瞭解，本發明之方法亦可以其它電腦系統組態實施，包括單一處理器或多處理器電腦系統、微電腦、主機電腦，以及個人電腦、手持式計算裝置、基於微處理器之或可程式化之消費電子產品等等，其分別可操作地耦合至一或多個關聯裝置。

所示各態樣亦可實施於分佈式計算環境中，其中某些任務係由透過一通信網路進行鏈接之遠端處理裝置執行。在一分佈式計算環境中，程式模組可位於本地記憶體儲存裝置與遠端記憶體儲存裝置二者上。

一電腦通常包括各種各樣之電腦可讀媒體。電腦可讀媒體可係為任何可由電腦存取之可用媒體，並包括揮發性及非揮發性媒體、可移除式及不可移除式媒體。例如但不限於，電腦可讀媒體可包括電腦儲存媒體及通信媒體。電腦儲存媒體包括以任何方法或技術實施之揮發性及非揮發性、可移除式及不可移除式媒體，用於儲存例如電腦可讀指令、資料結構、程式模組或其它資料等資訊。電腦儲存媒體包括但不限於 RAM、ROM、EEPROM、快閃記憶體或其它記憶體技術、CD-ROM、數位視訊光碟(DVD)或其它光碟儲存器、磁性卡匣、磁帶、磁碟儲存器或其它磁性儲存裝置，抑或任何其它可用於儲存所需資訊並可由電腦存取之媒體。

再次參見第 12 圖，用於實施各種態樣之實例性計算系統 1200 包括一電腦 1202，電腦 1202 具有一處理單元 1204、一系統記憶體 1206 及一系統匯流排 1208。系統匯

流排 1208 為包括但不限於系統記憶體 1206 之系統組件提供通往處理單元 1204 之介面。處理單元 1204 可係為各種市售處理器其中之任一種。亦可採用雙重微處理器及其它多處理器架構作為處理單元 1204。

系統匯流排 1208 可係為多種匯流排結構類型中之任一種，該匯流排結構可更使用各種市售匯流排架構中之任一種互連至一記憶體匯流排（具有或不具有一記憶體控制器）、一周邊匯流排、以及一本地匯流排。系統記憶體 1206 可包括非揮發性記憶體(NON-VOL)1210 及/或揮發性記憶體 1212（例如隨機存取記憶體(RAM)）。一基本輸入/輸出系統(BIOS)可儲存於非揮發性記憶體 1210（例如 ROM、EPROM、EEPROM 等等）中，該 BIOS 係為基本常式，有助於在電腦 1202 內之各元件之間傳送資訊，例如在啟動期間。揮發性記憶體 1212 亦可包括一用於快取資料之高速 RAM，例如靜態 RAM。

電腦 1202 更包括一內部硬碟機(HDD)1214（例如 EIDE、SATA）、一軟磁碟驅動機(FDD)1216（例如用於對一可移除式磁碟 1218 進行讀取或寫入）及一光碟機 1220（例如讀取一 CD-ROM 碟 1222 或對例如 DVD 等其它高容量光學媒體進行讀取或寫入），其中該內部 HDD 1214 亦可被組態成外用於一適用之底盤上。HDD 1214、FDD 1216 及光碟機 1220 可分別透過一 HDD 介面 1224、一 FDD 介面 1226 及一光碟機介面 1228 連接至系統匯流排 1208。用於外部驅動實施方案之 HDD 介面 1224 可包括通用串列匯流排

(USB)與 IEEE 1394 介面技術至少其中之一或同時包括二者。

該等驅動機及相關聯之電腦可讀媒體提供對資料、資料結構、電腦可執行指令等等之非揮發性儲存。對於電腦1202，該等驅動機及媒體容許以一適宜之數位格式儲存任何資料。儘管以上對電腦可讀媒體之描述係指一 HDD、一可移除式磁碟（例如 FDD）以及例如 CD 或 DVD 等可移除式光學媒體，然而熟習此項技術者應瞭解，可由電腦讀取之其它類型之媒體（例如 zip 驅動機、磁性卡匣、快閃記憶卡、匣式磁帶等等）亦可用於該實例性作業環境中，且進一步，任何此等媒體皆可包含電腦可執行指令以用於執行所揭示架構之新穎方法。

該等驅動機及揮發性記憶體 1212 中可儲存若干程式模組，包括一作業系統 1230、一或多個應用程式 1232、其它程式模組 1234，及程式資料 1236。該一或多個應用程式 1232、其它程式模組 1234，及程式資料 1236 可包括系統 100 及關聯區塊、系統 500 及關聯區塊、文件資訊 104、TAUC 資料 602、點擊資訊 610、資料流 700（及演算法），以及方塊圖 800（及關聯區塊）。

作業系統、應用程式、模組，及/或資料之全部或某些部分亦可被快取於揮發性記憶體 1212 中。應瞭解，可使用不同市售作業系統或作業系統之組合來實施所揭示架構。

使用者可透過一或多個有線/無線輸入裝置（例如，鍵盤 1238 及指標裝置，例如滑鼠 1240）輸入命令及資訊至

電腦 1202 中。其它輸入裝置（圖未示）可包括麥克風、IR 遠端控制器、操縱桿、遊戲搖桿、指示筆、觸控螢幕等等。該等及其它輸入裝置常常透過一與系統匯流排 1208 相耦合之輸入裝置介面 1242 連接至處理單元 1204，但亦可藉由例如並列埠、IEEE 1394 串列埠、遊戲埠、USB 埠、IR 介面等其它介面相連。

一監視器 1244 或其它類型之顯示裝置亦透過一介面（例如一視訊配接器 1246）連接至系統匯流排 1208。除監視器 1244 外，電腦還通常包括其它周邊輸出裝置（圖未示），例如揚聲器、列印機等等。

電腦 1202 可利用邏輯連接，藉由與一或多個遠端電腦（例如一或多個遠端電腦 1248）之有線及/或無線通信，在一聯網環境中運作。該（等）遠端電腦 1248 可係為一工作站、一伺服器電腦、一選路器、一個人電腦、可攜式電腦、基於微處理器之娛樂裝置、一對等裝置或其它常用網路節點，且通常包括上文關於電腦 1202 所述之許多或所有元件，儘管為簡明起見，圖中僅示出一記憶體/儲存裝置 1250。所描繪之邏輯連接包括與一局部區域網路(LAN)1252 及/或更大網路（例如一廣域網路(WAN)1254）之有線/無線連接性。此等 LAN 及 WAN 聯網環境在辦公室及公司中很常見，並有利於達成企業範圍之電腦網路，例如內部網路，所有該等網路皆可連接至一全球通信網路，例如網際網路。

當用於一 LAN 聯網環境時，電腦 1202 透過一有線及/或無線通信網路介面或配接器 1256 連接至 LAN 1252。配

接器 1256 可有利於與 LAN 1252 之有線及/或無線通信，LAN 1252 亦可包括一設置於其上之無線存取點，以用於與配接器 1256 之無線功能進行通信。

當用於一 WAN 聯網環境時，電腦 1202 可包括一數據機 1258，或者連接至 WAN 1254 上之一通信伺服器，抑或具有其它用於在 WAN 1254 上建立通信之手段（例如藉由網際網路）。數據機 1258 可處於內部或外部且係為一有線及/或無線裝置，其透過輸入裝置介面 1242 連接至系統匯流排 1208。在一聯網環境中，關於電腦 1202 所描繪之程式模組或其某些部分可儲存於遠端記憶體/儲存裝置 1250 中。應瞭解，所示網路連接係為實例性的，亦可使用其它用於在各電腦之間建立通信鏈接之裝置。

電腦 1202 可運作以與利用 IEEE 802 標準家族之有線及無線裝置或實體進行通信，諸如被可操作地設置成與例如以下裝置進行無線通信（例如，IEEE 802.11 空中調變技術）之無線裝置：列印機、掃描儀、桌上型及/或可攜式電腦、個人數位助理(PDA)、通信衛星、任何與一可無線偵測之標簽相關聯之設備或位置（例如售貨亭、報攤、休息室）以及電話。此包括至少 Wi-Fi( 或 Wireless Fidelity)、WiMax 及 Bluetooth™ 無線技術。因此，通信可如在一習知網路中一樣為一預定義結構，或者簡單地為至少兩個裝置間之專門通信。Wi-Fi 網路利用被稱為 IEEE 802.11x（a、b、g 等等）之無線電技術提供安全、可靠、快速之無線連接性。一 Wi-Fi 網路可用於將各電腦相互連接、連接至網際網路，

以及連接至有線網路（其使用 IEEE 802.3 相關媒體及功能）。

上文說明包括所揭示架構之實例。當然，不可能描述組件及/或方法之每一可設想之組合，但此項技術中之通常知識者可知，亦可存在諸多進一步之組合及排列。因此，該新穎架構旨在囊括仍歸屬於隨附申請專利範圍之精神及範圍內之所有此等改動、修改及變化。此外，對於在本詳細說明或申請專利範圍中所用之措詞「包括(include)」而言，該措詞旨在具有與措詞「包括(comprising)」在申請專利範圍項中用作轉接詞時相類似之包括方式。

【圖式簡單說明】

第 1 圖例示一由電腦實施之相關性系統。

第 2 圖例示一種用於計算編輯距離之實例性匹配演算法之流程圖。

第 3 圖例示利用改良之編輯距離及匹配演算法，根據一查詢字串及資料字串處理及產生編輯距離值。

第 4 圖例示利用改良之編輯距離及匹配演算法，根據一查詢字串及資料字串處理及產生編輯距離值之另一實例。

第 5 圖例示一由電腦實施之相關性系統，其採用一神經網路來幫助產生文件之相關性得分。

第 6 圖例示可在文件資訊中用於確定查詢字串與資料

字串間之編輯距離之資料類型。

第 7 圖例示一索引時間處理資料流。

第 8 圖例示一方塊圖，其顯示來自第 7 圖所示索引過程的用於結果排序之神經網路輸入。

第 9 圖例示一神經網路、編輯距離輸入及用於計算和產生搜尋結果之原始特徵輸入之實例性系統實施方式。

第 10 圖例示一種用於確定一文件結果集合之文件相關性之方法。

第 11 圖例示一種計算文件相關性之方法。

第 12 圖例示一計算系統之方塊圖，該計算系統可操作以根據本文所揭示之架構，利用 TAUC 特徵執行編輯距離處理以進行搜尋結果排序。

【主要元件符號說明】

| | |
|---|---|
| 100 | 由電腦實施之相關性系統 |
| 102 | 處理組件 |
| 104 | 文件資訊 |
| 106 | 文件 |
| 108 | 搜尋結果 |
| 110 | 查詢字串 |
| 112 | 接近性組件 |
| 114 | 編輯距離 |
| 116 | 資料字串 |

| | |
|---|---|
| 300 | 矩陣 |
| 302 | 查詢字串 |
| 304 | 目標資料字串 |
| 400 | 矩陣 |
| 402 | 查詢字串 |
| 404 | 目標資料字串 |
| 500 | 由電腦實施之相關性系統 |
| 502 | 神經網路 |
| 504 | 相關性得分 |
| 602 | TAUC 資料 |
| 604 | 標題文本 |
| 606 | 錨文本 |
| 608 | URL 文本或字元 |
| 610 | 點擊資訊 |
| 700 | 索引時間處理資料流 |
| 704 | 詞分離演算法 |
| 706 | 字典 |
| 708 | URL 分離演算法 |
| 710 | 索引過程 |
| 712 | 過濾器 |
| 800 | 方塊圖 |
| 802 | URL 編輯距離(ED) |
| 804 | 頂部 N 個錨 ED |
| 806 | 標題 ED |

| | |
|---|---|
| 808 | 點擊 ED |
| 810 | 其它特徵 |
| 900 | 實例性系統 |
| 902 | BM25 函數 |
| 904 | 點擊距離 |
| 906 | URL 深度 |
| 908 | 文件類型 |
| 910 | 語言匹配 |
| 1200 | 實例性計算系統 |
| 1202 | 電腦 |
| 1204 | 處理單元 |
| 1206 | 系統記憶體 |
| 1208 | 系統匯流排 |
| 1210 | 非揮發性記憶體(NON-VOL) |
| 1212 | 揮發性記憶體 |
| 1214 | 內部硬碟機(HDD) |
| 1216 | 軟磁碟驅動機(FDD) |
| 1218 | 可移除式磁碟 |
| 1220 | 光碟機 |
| 1222 | CD-ROM 碟 |
| 1224 | HDD 介面 |
| 1226 | FDD 介面 |
| 1228 | 光碟機介面 |
| 1230 | 作業系統 |

| | |
|---|---|
| 1232 | 應用程式 |
| 1234 | 程式模組 |
| 1236 | 程式資料 |
| 1238 | 鍵盤 |
| 1240 | 滑鼠 |
| 1244 | 監視器 |
| 1246 | 視訊配接器 |
| 1248 | 遠端電腦 |
| 1250 | 記憶體/儲存裝置 |
| 1252 | 局部區域網路(LAN) |
| 1254 | 廣域網路(WAN) |
| 1256 | 配接器 |
| 1258 | 數據機 |

## 七、申請專利範圍：

1. 一種由電腦實施之相關性系統，包含：

一或更多個處理器；以及

一記憶體，耦接至該一或更多個處理器，該記憶體儲存指令，在藉由該一或更多個處理器執行該等指令時，使該一或更多個處理器實行以下步驟：

根據一查詢字串從作為搜尋結果接收之一文件中提取文件資訊，該文件資訊包括一全球資源定位器，其中該全球資源定位器包括一複合詞；

將該複合詞分離為多個單一詞；

在詞之一字典中找到該等多個單一詞中之至少一者；

根據所提取之該文件資訊產生一目標資料字串，該目標資料字串包括在該字典中找到的該等多個單一詞中之一者；以及

計算該目標資料字串與該查詢字串間之編輯距離，該編輯距離用於確定一文件之相關性以作為結果排序之一部分。

2. 如申請專利範圍第 1 項所述之系統，其中該文件資訊包括一標題資訊、全球資源定位器資訊、點擊資訊、或錨文本。

3. 如申請專利範圍第 1 項所述之系統，其中在索引時間中將該文件資訊之該等複合詞分離，以計算相對於該全球資源定位器之該編輯距離。

4. 如申請專利範圍第 1 項所述之系統，進一步包含用於在

索引時間中過濾該文件資訊之錨文本，以計算排在頂部之錨文本之一集合之指令。

5. 如申請專利範圍第 1 項所述之系統，其中該文件資訊包括標題字元、錨字元、或點擊字元中至少之一者，且其中該系統進一步包括一神經網路，該神經網路可操作以根據該文件資訊以及一 BM25F 函數、點擊距離、文件類型、語言及全球資源定位器深度之原始輸入特徵而計算該文件之相關性。

6. 如申請專利範圍第 1 項所述之系統，其中該編輯距離係根據詞插入及刪除而計算，以增大該目標資料字串與該查詢字串間之接近性。

7. 如申請專利範圍第 1 項所述之系統，其中該編輯距離係根據詞插入及刪除相關聯之成本而計算，以增大該目標資料字串與該查詢字串間之接近性。

8. 一種由電腦執行之確定一文件之相關性之方法，包含以下步驟：

接收一查詢字串，該查詢字串作為一搜尋過程之一部分；

從包括在該搜尋過程期間返回之一文件中之文件資訊提取一全球資源定位器，其中該全球資源定位器包括一複合詞；

藉由將該全球資源定位器之該複合詞分離為多個單一詞，並在詞之一字典中找到該等多個單一詞中之至少一者，而從該全球資源定位器產生一目標資料字串；

計算該目標資料字串與該查詢字串間之編輯距離；以及

根據該編輯距離，計算一相關性得分。

9. 如申請專利範圍第 8 項所述之方法，進一步包含以下步驟：採用詞插入作為計算該編輯距離之一部分，並評估將一詞插入該查詢字串中以產生該目標資料字串之一插入成本，該成本被表示為一加權參數。

10.如申請專利範圍第 8 項所述之方法，進一步包含以下步驟：採用詞刪除作為計算該編輯距離之一部分，並評估刪除該查詢字串中之一詞以產生該目標資料字串之一刪除成本，該成本被表示為一加權參數。

11.如申請專利範圍第 8 項所述之方法，進一步包含以下步驟：計算一位置成本作為計算該編輯距離之一部分，該位置成本與該目標資料字串中一詞位置之詞插入及詞刪除之一或更多者相關聯。

12.如申請專利範圍第 8 項所述之方法，進一步包含以下步驟：在該目標資料字串之字元與該查詢字串之字元之間執行一匹配過程，以計算該編輯距離之一總體計算成本。

13.如申請專利範圍第 8 項所述之方法，進一步包含以下步驟：在索引時間中分離該全球資源定位器之複合詞。

14.如申請專利範圍第 8 項所述之方法，進一步包含以下步驟：根據在該文件中之出現頻率過濾該目標資料字串之錨文本，以找到排在頂部之錨文本之一集合。

15.如申請專利範圍第 14 項所述之方法，進一步包含以下

步驟：計算該集合中錨文本之一編輯距離得分。

16.如申請專利範圍第 8 項所述之方法，進一步包含以下步驟：在應用一變換函數後，將從該編輯距離之計算中導出之一得分輸入一兩層式神經網路，該得分係根據計算與標題資訊、錨資訊、點擊資訊、或全球資源定位器資訊、以及其他原始輸入特徵中之至少一者相關聯之該編輯距離而產生。

17.一種由電腦執行之計算一文件之相關性之方法，包含以下步驟：

處理一查詢字串，該查詢字串作為一搜尋過程之一部分，以返回一文件結果集合；

根據從該結果集合之一文件中提取之文件資訊，產生一目標資料字串，該文件資訊包括一全球資源定位器，其中該全球資源定位器包括一複合詞，其中產生該目標資料字串之步驟包括以下步驟：將該複合詞分離為多個單一詞，並在詞之一字典中找到該等多個單一詞中之至少一者；

根據詞插入、詞刪除、及詞位置，計算該目標資料字串與該查詢字串間之編輯距離；以及

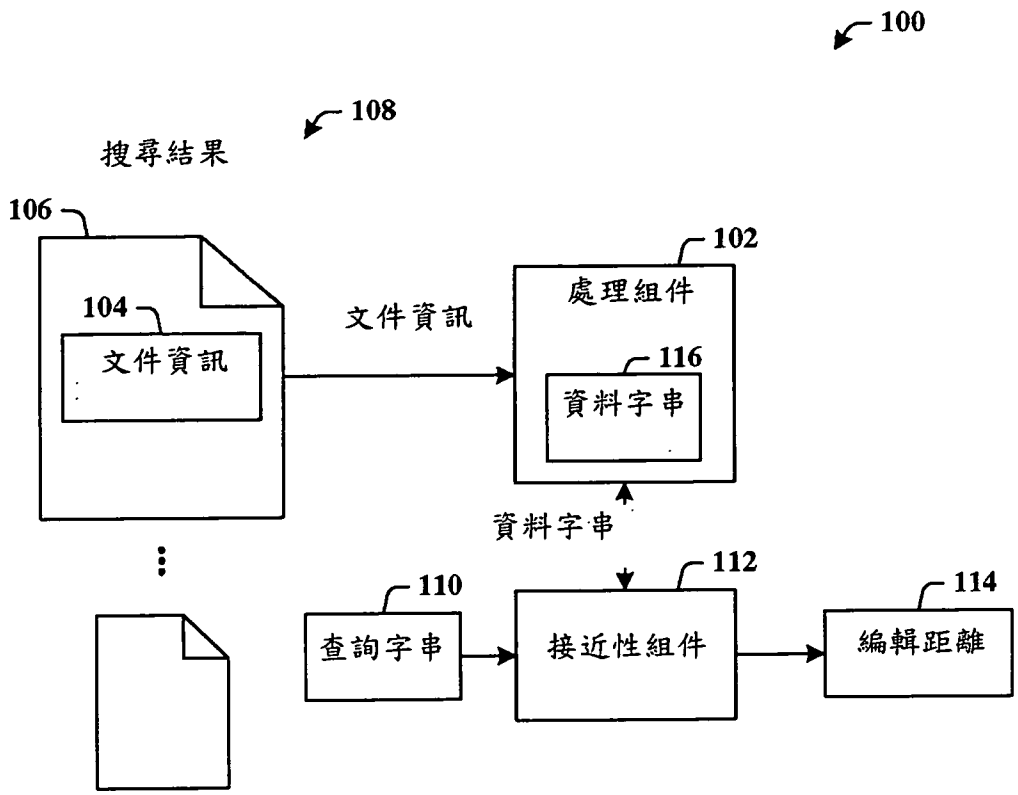
根據該編輯距離，計算一相關性得分，該相關性得分用於在該結果集合中對該文件進行排序。

18.如申請專利範圍第 17 項所述之方法，進一步包含以下步驟：計算與該詞插入、詞刪除及詞位置中之每一者相關聯之一成本，並將該成本之因素納入該相關性得分之計算中。

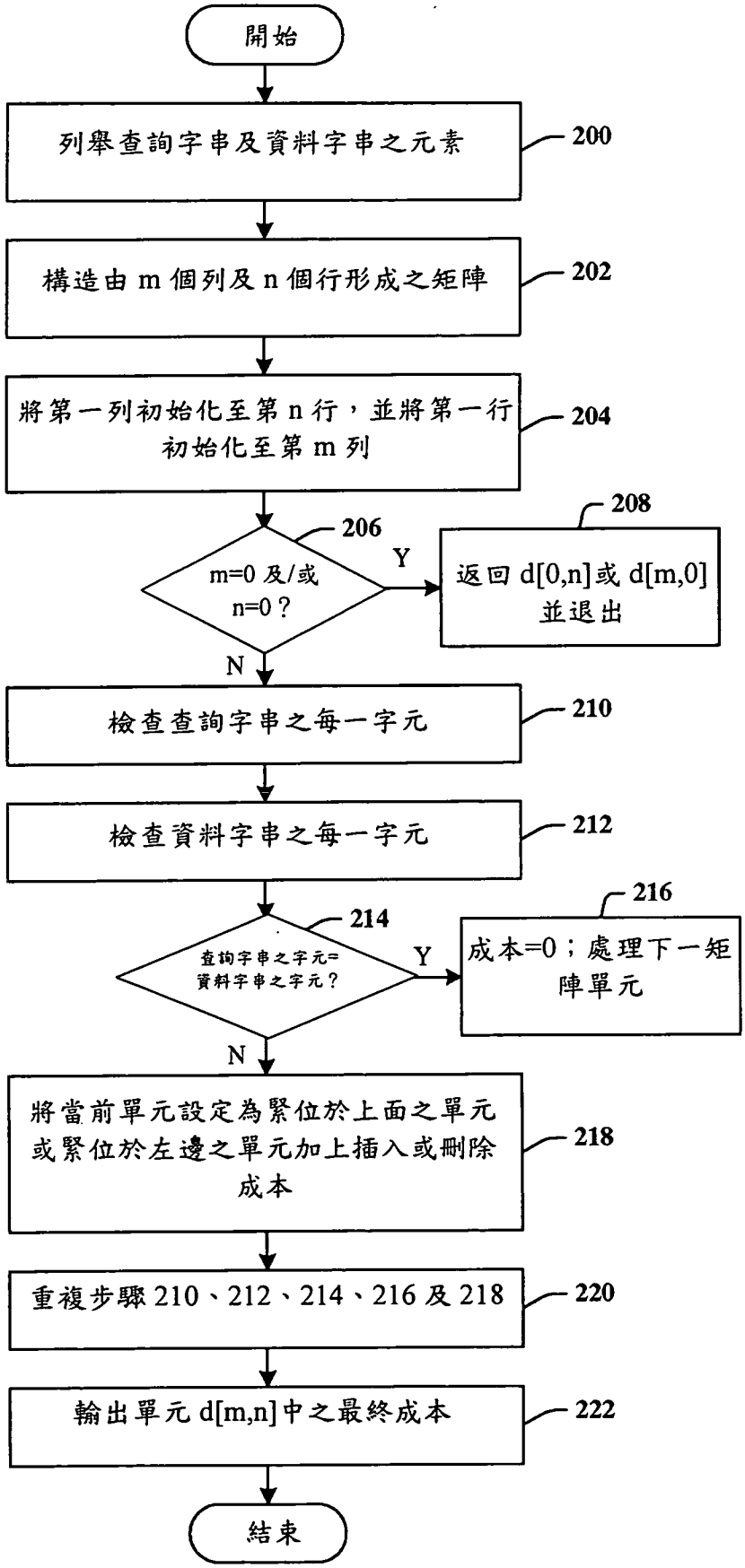
19.如申請專利範圍第 17 項所述之方法，進一步包含以下步驟：在索引時間中分離該全球資源定位器資訊之複合詞，以及在索引時間中根據該錨文本在該文件中出現之頻率過濾該錨文本資訊以找到排在頂部之錨文本之一集合。

20.如申請專利範圍第 17 項所述之方法，進一步包含以下步驟：讀取該查詢字串之詞之出現，以依照在一源全球資源定位器字串中之出現次序構造一查詢詞字串，並以字標記填充該等詞間之空間。

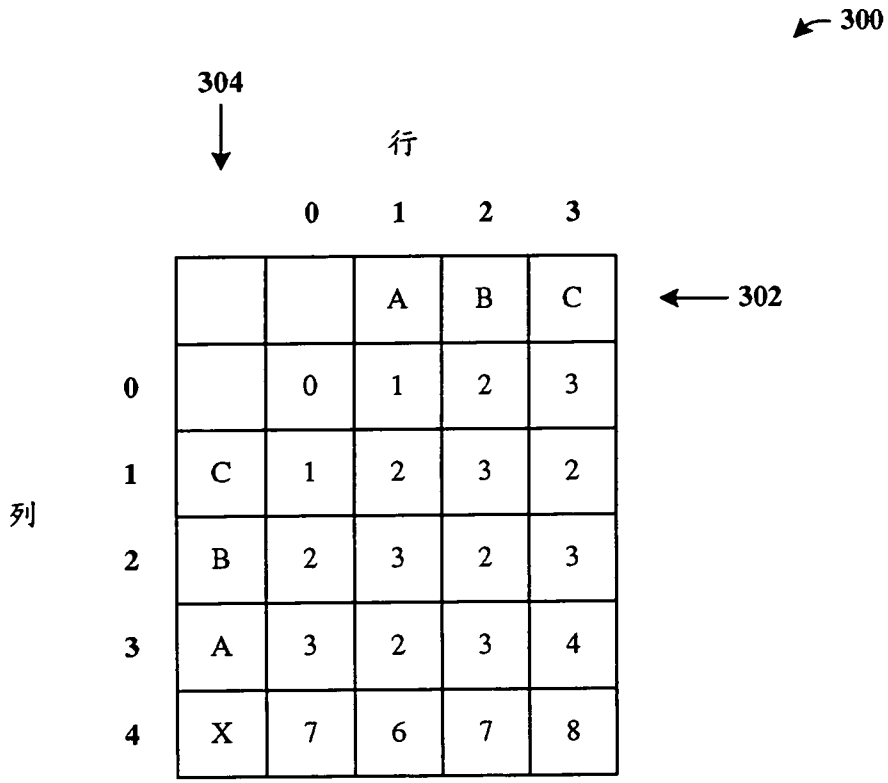
八、圖式：


100
108
搜尋結果
106
104
文件資訊
文件資訊
102
處理組件
116
資料字串
資料字串
110
查詢字串
112
接近性組件
114
編輯距離

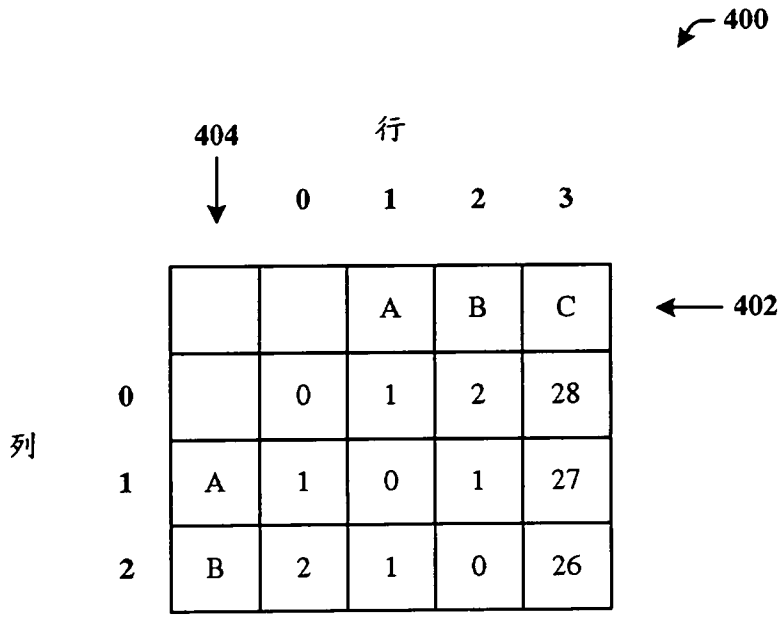
第 1 圖


開始
列舉查詢字串及資料字串之元素
200
構造由 m 個列及 n 個行形成之矩陣
202
將第一列初始化至第 n 行，並將第一行初始化至第 m 列
204
206
m=0 及/或 n=0？
Y
208
返回 d[0,n]或 d[m,0] 並退出
N
檢查查詢字串之每一字元
210
檢查資料字串之每一字元
212
214
查詢字串之字元= 資料字串之字元？
Y
216
成本=0；處理下一矩陣單元
N
將當前單元設定為緊位於上面之單元或緊位於左邊之單元加上插入或刪除成本
218
重複步驟 210、212、214、216 及 218
220
輸出單元 d[m,n]中之最終成本
222
結束

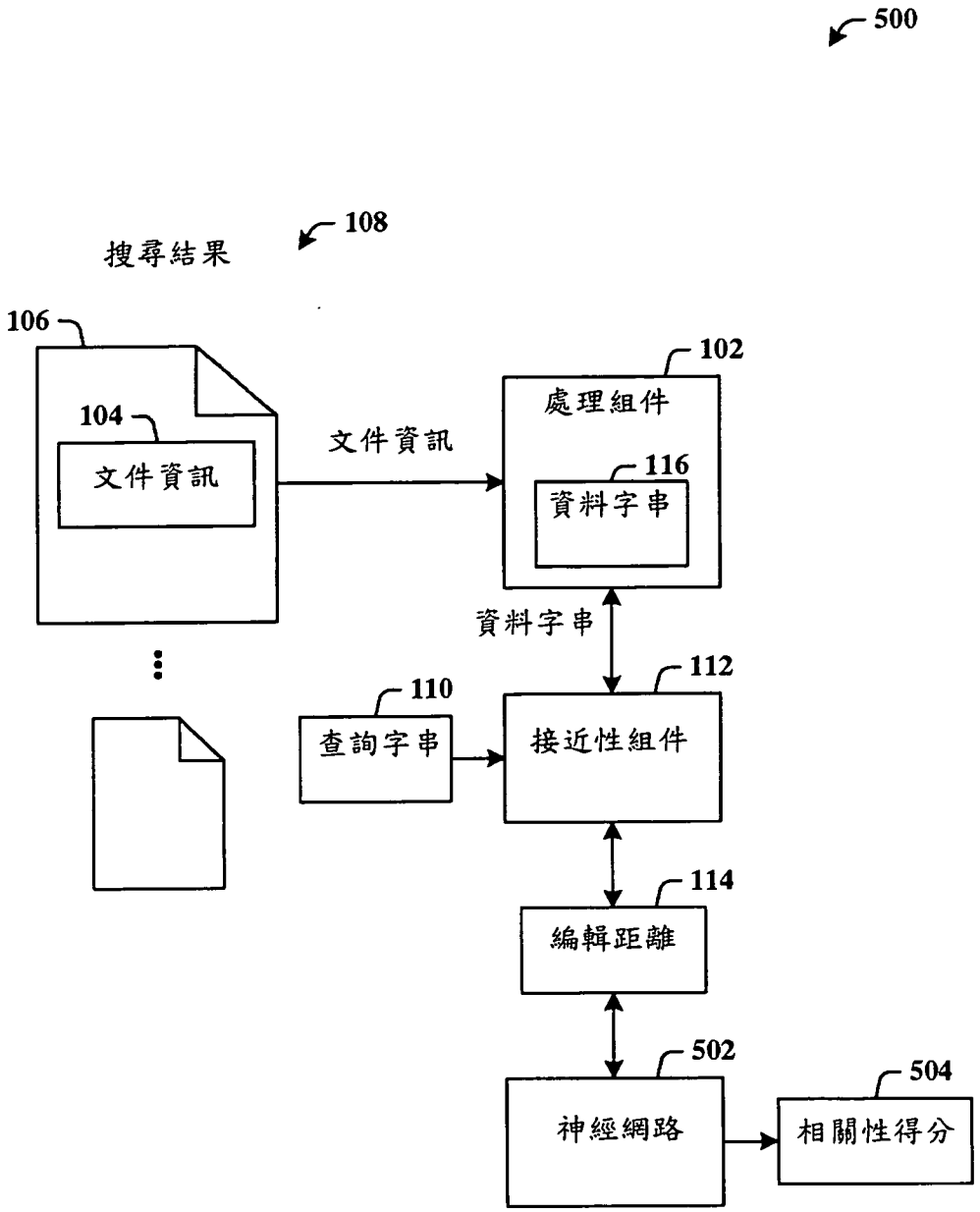
第 2 圖



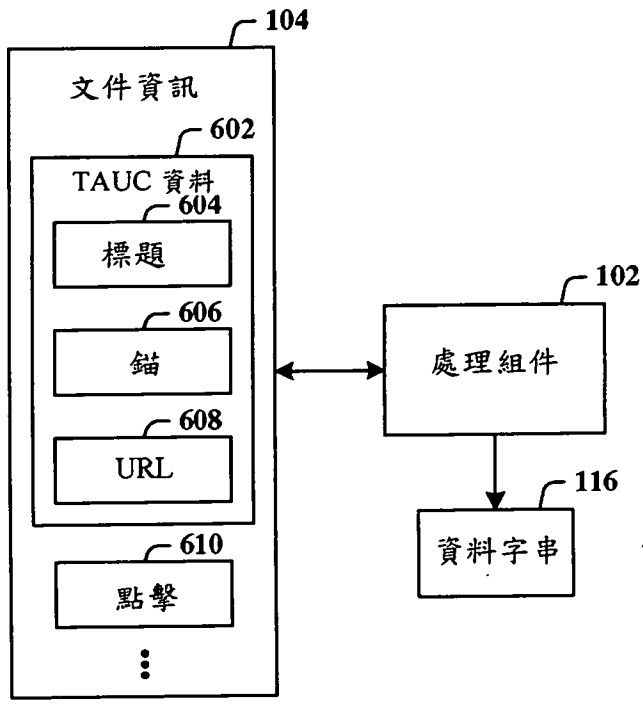

第 3 圖



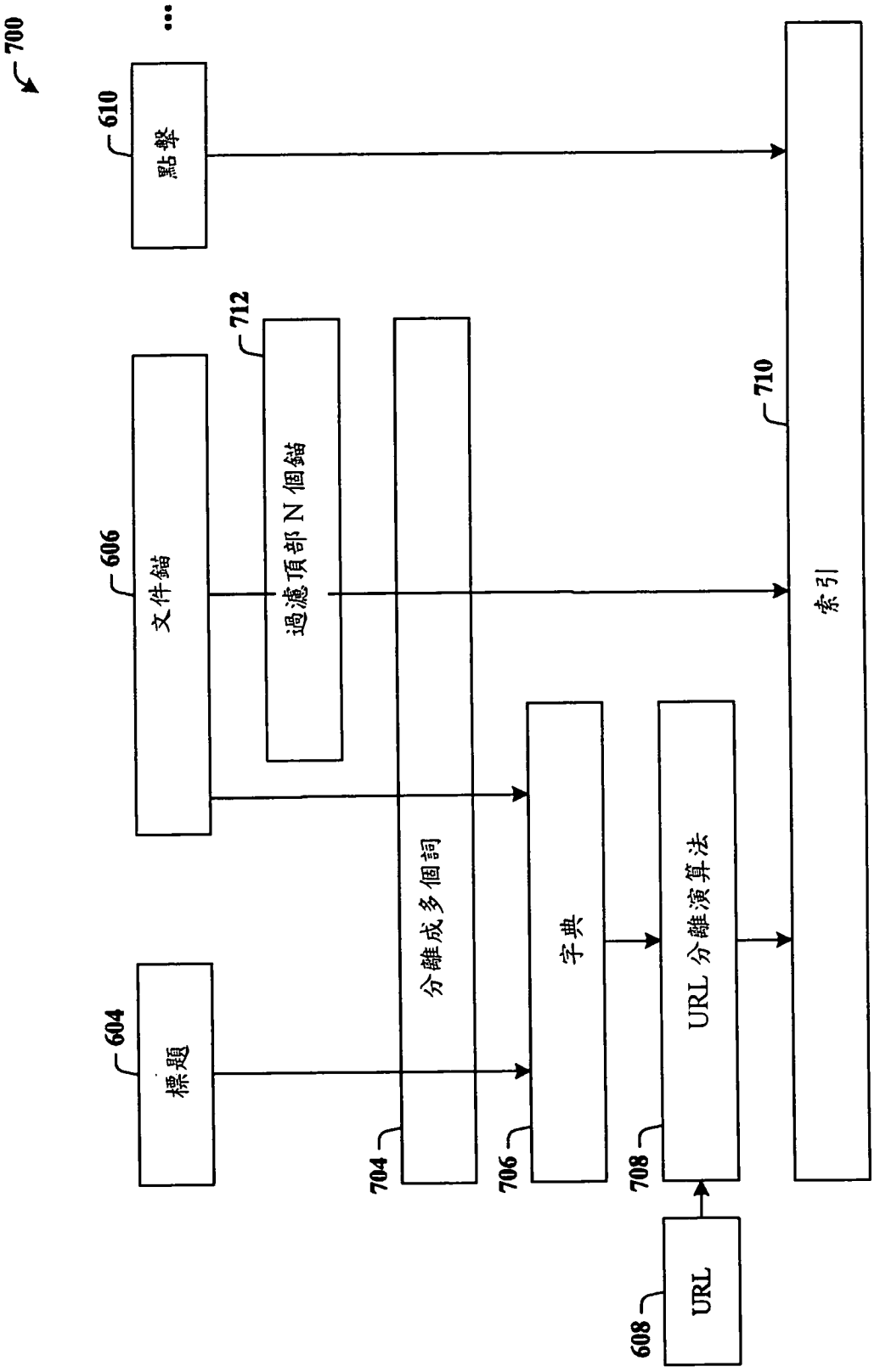

第 4 圖


500
108
搜尋結果
106
104
文件資訊
文件資訊
102
處理組件
116
資料字串
資料字串
112
110
查詢字串
接近性組件
114
編輯距離
502
504
神經網路
相關性得分

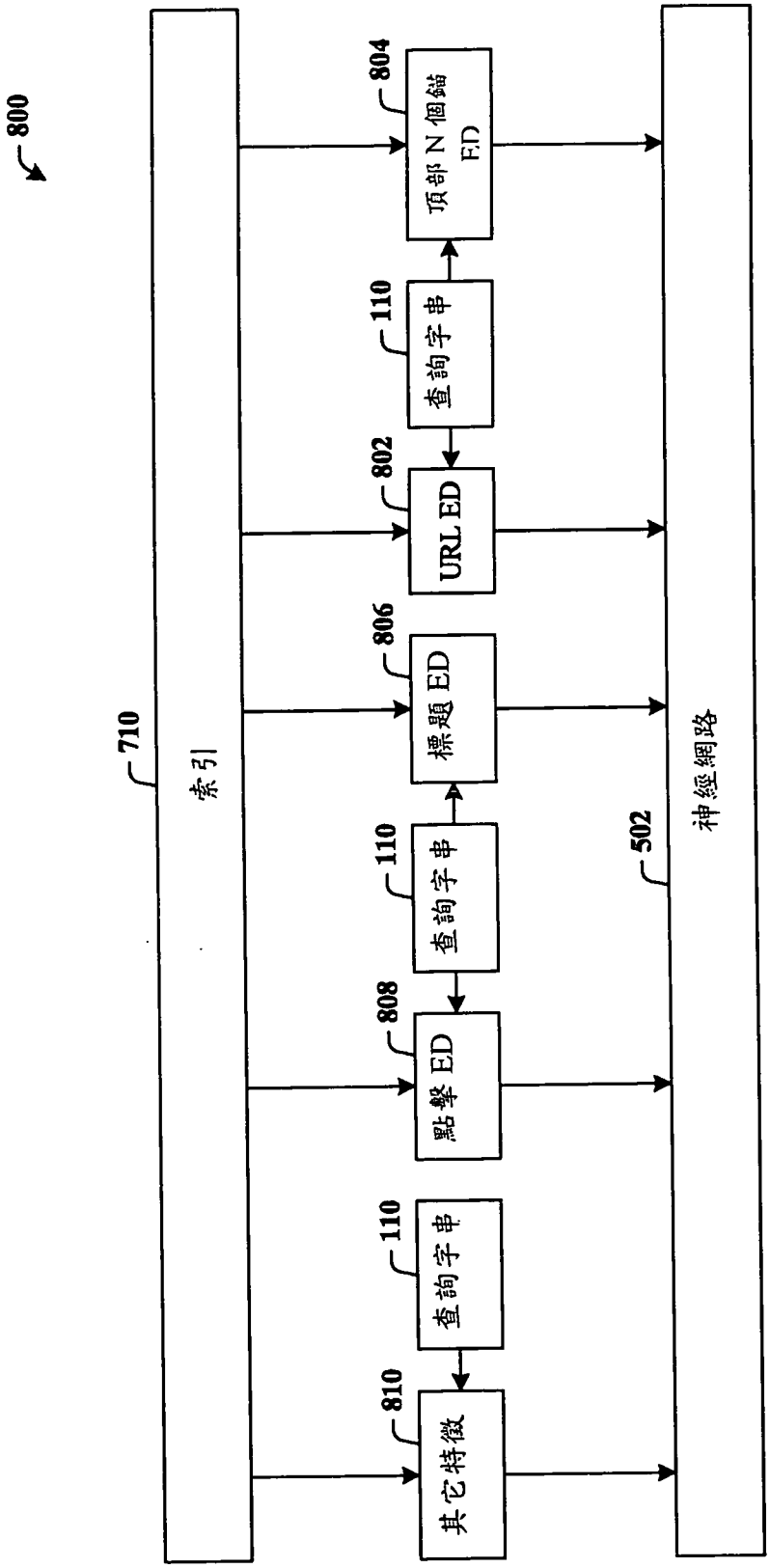
第 5 圖


104
文件資訊
602
TAUC 資料
604
標題
606
錨
608
URL
610
點擊
102
處理組件
116
資料字串

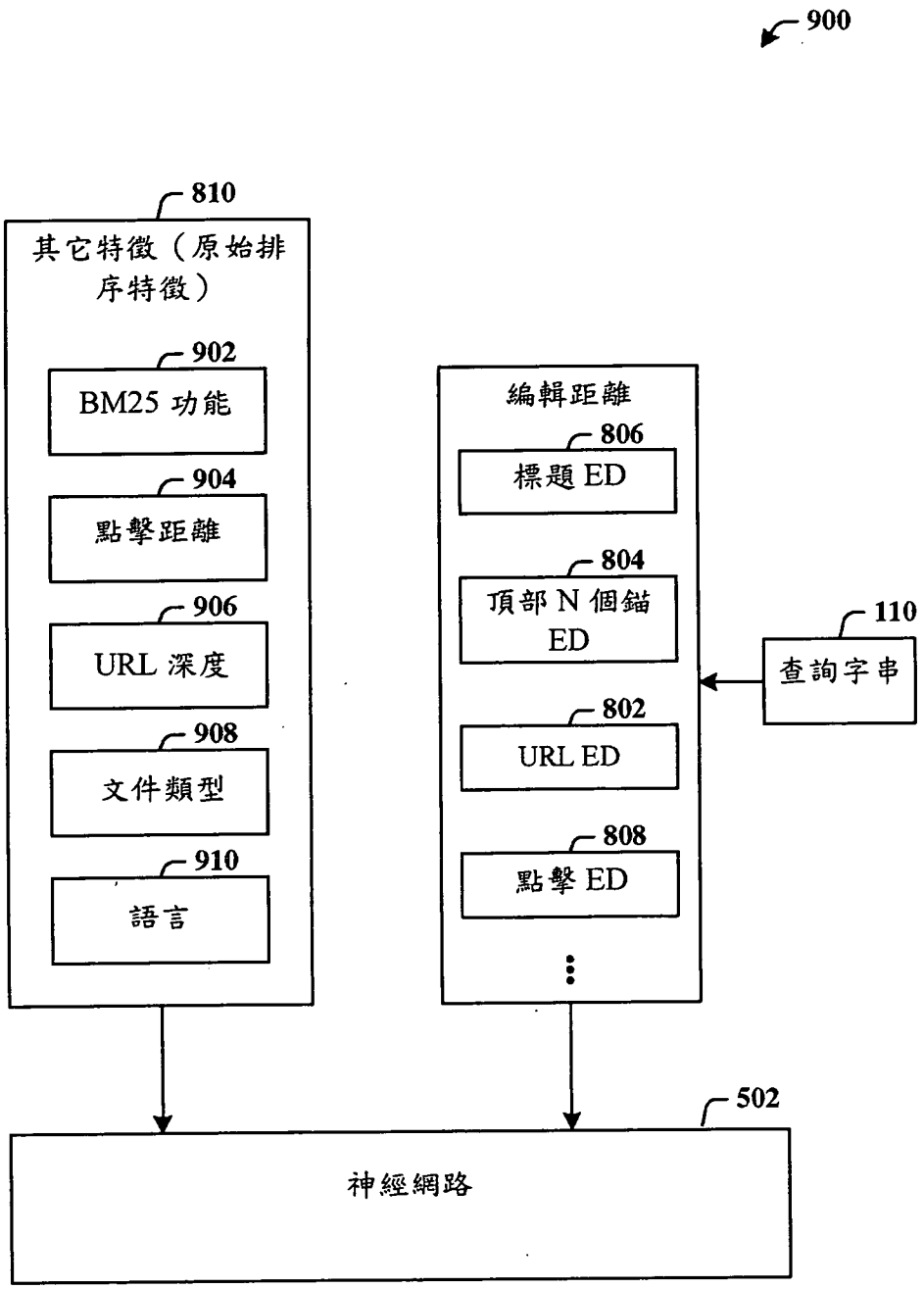
第 6 圖


700
610
點擊
606
文件錨
712
過濾頂部 N 個錨
710
索引
604
標題
分離成多個詞
704
字典
706
URL 分離演算法
708
URL
608

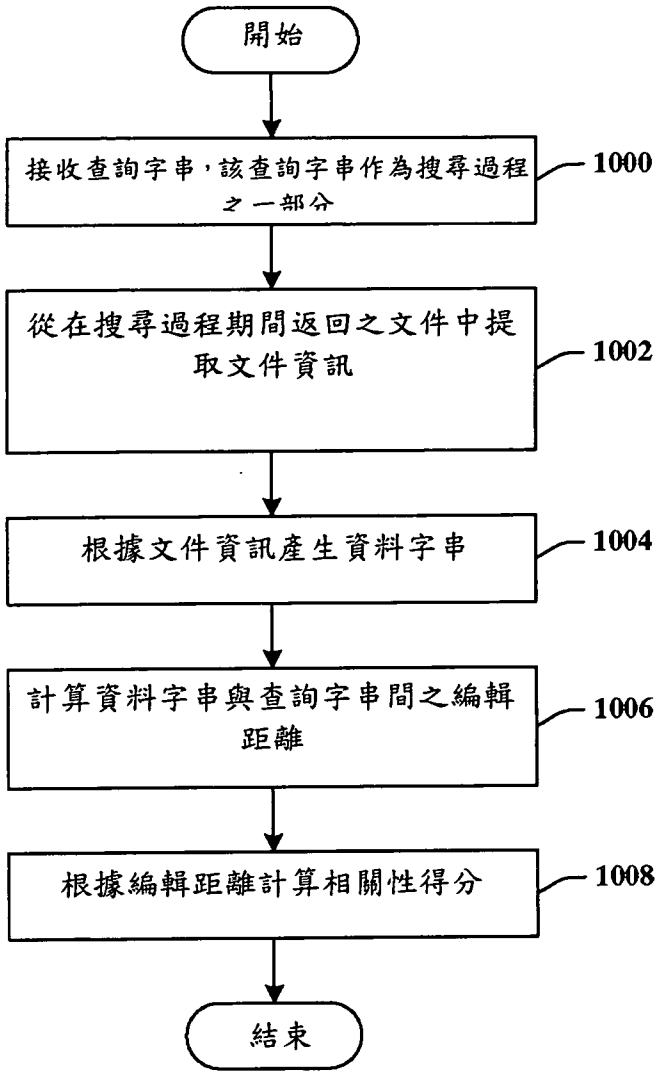
第 7 圖


800
710
索引
804
頂部 N 個錨 ED
110
查詢字串
802
URL ED
806
標題 ED
110
查詢字串
808
點擊 ED
110
查詢字串
810
其它特徵
502
神經網路

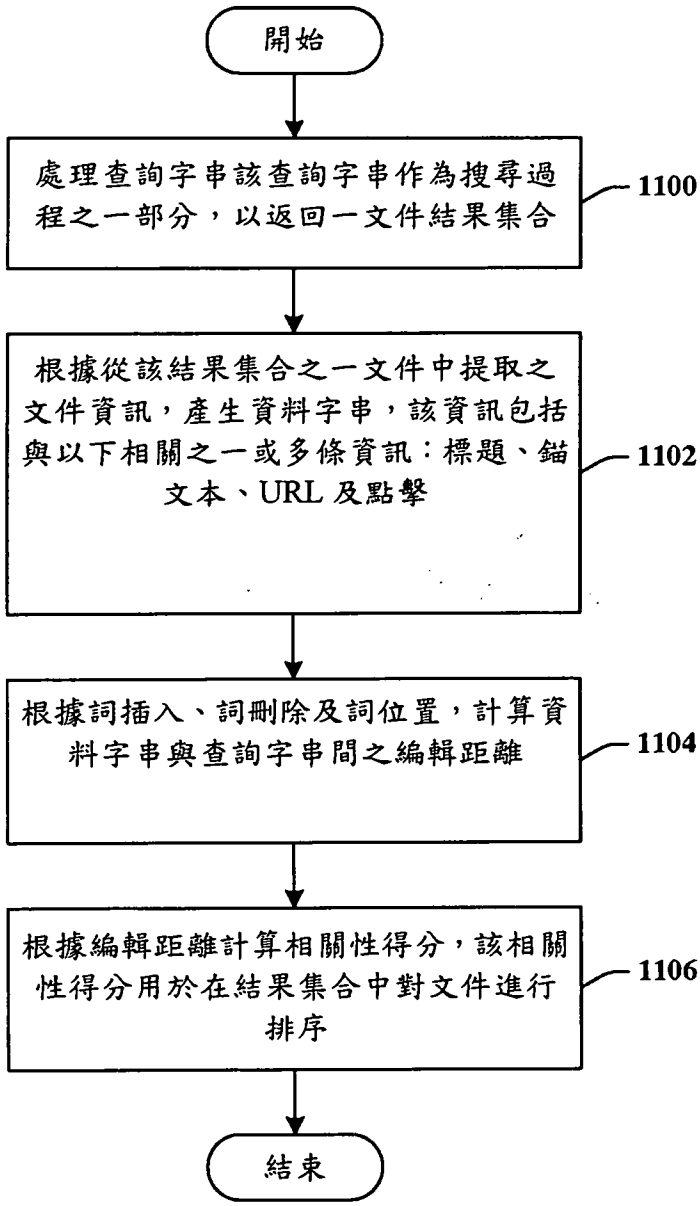
第 8 圖


900
810
其它特徵（原始排序特徵）
902
BM25 功能
904
點擊距離
906
URL 深度
908
文件類型
910
語言
編輯距離
806
標題 ED
804
頂部 N 個錨 ED
802
URL ED
808
點擊 ED
110
查詢字串
502
神經網路

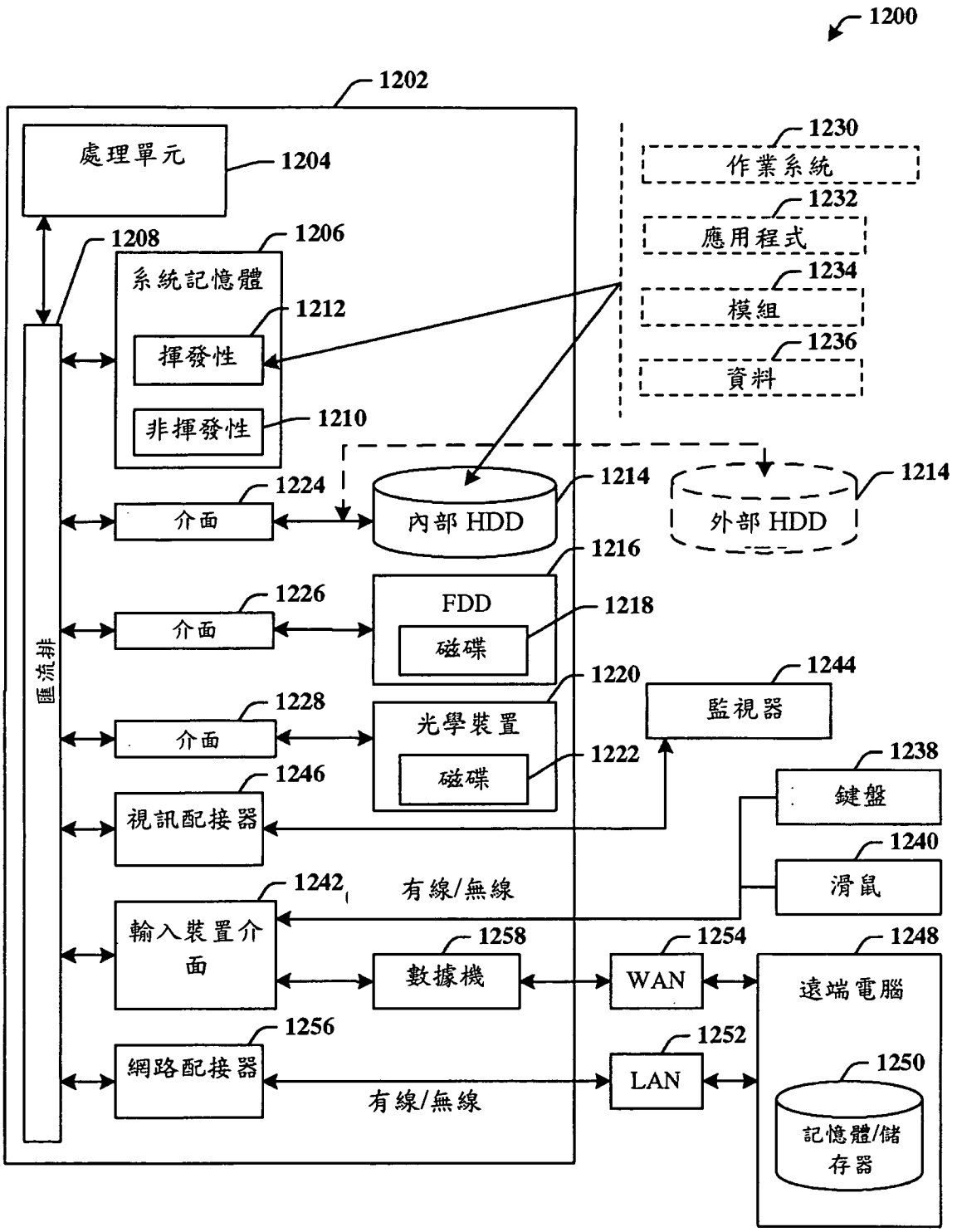
第 9 圖


開始
接收查詢字串，該查詢字串作為搜尋過程之一部分
1000
從在搜尋過程期間返回之文件中提取文件資訊
1002
根據文件資訊產生資料字串
1004
計算資料字串與查詢字串間之編輯距離
1006
根據編輯距離計算相關性得分
1008
結束

第 10 圖


開始
處理查詢字串該查詢字串作為搜尋過程之一部分，以返回一文件結果集合
1100
根據從該結果集合之一文件中提取之文件資訊，產生資料字串，該資訊包括與以下相關之一或多條資訊：標題、錨文本、URL 及點擊
1102
根據詞插入、詞刪除及詞位置，計算資料字串與查詢字串間之編輯距離
1104
根據編輯距離計算相關性得分，該相關性得分用於在結果集合中對文件進行排序
1106
結束

第 11 圖


1200
1202
處理單元
1204
1208
1206
系統記憶體
1212
揮發性
非揮發性
1210
匯流排
1230
作業系統
1232
應用程式
1234
模組
1236
資料
1224
介面
內部 HDD
1214
外部 HDD
1214
1216
1226
介面
FDD
磁碟
1218
1220
1228
介面
光學裝置
磁碟
1222
1244
監視器
1238
鍵盤
1246
視訊配接器
1240
滑鼠
1242
有線/無線
輸入裝置介面
1258
數據機
1254
WAN
1248
遠端電腦
1256
網路配接器
1252
LAN
1250
有線/無線
記憶體/儲存器

第 12 圖