

發明專利說明書

(本說明書格式、順序及粗體字，請勿任意更動，※記號部分請勿填寫)

※申請案號：**P5 136094**

※申請日期：**P5. P. 28**

※IPC 分類：**H03M¹³/₃₇ · G06F¹²/₃₀**

一、發明名稱：(中文/英文)

資料搜尋方法及其數值資料編碼方法 / DATA QUERY
METHOD AND DATA CODING METHOD THEREOF

二、申請人：(共 1 人)

姓名或名稱：(中文/英文)(簽章) **ID : 46804706**

國立交通大學 / NATIONAL CHIAO TUNG UNIVERSITY

指定 為應受送達人

代表人：(中文/英文)(簽章) 黃威 / WEI HWANG

住居所或營業所地址：(中文/英文)

新竹市大學路 1001 號 / NO. 1001 DASYUE Road, Hsinchu CITY 300-10,
Taiwan(R. O. C)

國 籍：(中文/英文) 中華民國 / ROC

電話/傳真/手機：(02)8227-8658

E-MAIL :

三、發明人：(共 3 人)

姓 名：(中文/英文)

ID :

1. 鄭哲聖 / CHER-SHENG CHENG J120318619

2. 單智君 / JYH-JIUN SHANN M220999752

3. 鍾崇斌 / CHUNG-PING CHUNG B101081298

國 籍：(中文/英文)

1. 中華民國 / ROC 2. 中華民國 / ROC 3. 中華民國 / ROC

四、聲明事項：

主張專利法第二十二條第二項第一款或第二款規定之事實，其事實發生日期為： 年 月 日。

申請前已向下列國家（地區）申請專利：

【格式請依：受理國家（地區）、申請日、申請案號 順序註記】

有主張專利法第二十七條第一項國際優先權：

無主張專利法第二十七條第一項國際優先權：

主張專利法第二十九條第一項國內優先權：

【格式請依：申請日、申請案號 順序註記】

主張專利法第三十條生物材料：

須寄存生物材料者：

國內生物材料 【格式請依：寄存機構、日期、號碼 順序註記】

國外生物材料 【格式請依：寄存國家、機構、日期、號碼 順序註記】

不須寄存生物材料者：

所屬技術領域中具有通常知識者易於獲得時，不須寄存。

五、中文發明摘要：

本發明係揭露一種資料搜尋方法及其數值資料編碼方法，此方法將已排序之複數個數值分成複數個群組，每一群組之第一個數值係代表為跳躍點，而其他數值係代表為內部點，接著，使用一第一編碼演譯程序對該些群組之跳躍點進行編碼，以產生複數個跳躍點編碼值，並使用一空間預估程序，預估此些內部點以一第二編碼演譯程序進行編碼時所需之位元保留空間，繼之，根據該第二編碼演譯程序對該些群組之該些內部點進行編碼，以產生複數個內部點編碼值。最後，根據群組之排列順序及位元保留空間，將跳躍點編碼值及內部點編碼值依序排列，以產生一輸出編碼資料。

六、英文發明摘要：

A data query method and a data coding method are disclosed. First, the method includes the step of dividing a plurality of data into a plurality of groups, wherein a first data of each group represents a skip pointer and the other data of each group represent inner pointers. Then, the skip pointers are coded with a first coding algorithm process to generate a plurality of coded skip pointer data. A coding required bits space of the inner pointers is predicted by using a bits space prediction process. Then, the inner pointers of each group are coded respectively with a second coding algorithm process to generate a plurality of coded inner pointer data. Finally, the coded skip pointer data and

200816653

the inner pointer data are posted sequentially to generate an output coded data based on the group post sequence and the coding required bits space.

七、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：

20~24：步驟流程；

八、本案若有化學式時，請揭示最能顯示發明特徵的化學式：

九、發明說明：

【發明所屬之技術領域】

本發明是有關於一種資料搜尋方法及其數值資料編碼方法，特別是有關於一種內建有可跳躍式解碼能力之轉置檔案以實現快速查詢之數值資料編碼方法。

七、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：

20~24：步驟流程；

八、本案若有化學式時，請揭示最能顯示發明特徵的化學式：

九、發明說明：

【發明所屬之技術領域】

本發明是有關於一種資料搜尋方法及其數值資料編碼方法，特別是有關於一種內建有可跳躍式解碼能力之轉置檔案以實現快速查詢之數值資料編碼方法。

【先前技術】

透過網際網路上進行資料檢索已成為目前最常用的資訊檢索，使用者可透過資料搜尋網站，如 Google 或 Yahoo，以關鍵字尋找所有具有相同關鍵字之網頁資料。一個資訊檢索系統會儲存有許多的文件，少則數千個，多則數十億個。在一個大型的資訊檢索系統中，文件的數目正以每兩年增加一倍的速度成長。為了讓使用者可以最短的時間內找到他們所需的資料，資訊檢索系統會建立一個索引結構。

請參閱第 1 圖，其係繪示一索引結構之示意圖。傳統上最常使用的索引結構是轉置索引 (inverted index)，主要兩個檔案組成：一個是索引檔案 (index file)，一個是轉置檔案 (inverted file)。所有出現在資料所儲存之文件中的關鍵字 (terms) 都會被記錄在索引檔案，例如圖中關鍵字 “computer” 及 “architecture”，並利用一個指標指到轉置檔案中相對應的轉置串列 (inverted list)，例如關鍵字 “computer” 係對應於轉置檔案中位址 355 之轉置串列 11，而關鍵字 “architecture” 係對應於轉置檔案中位址 252 之轉置串列 10。

轉置串列係紀錄此關鍵字出現於哪些文件中。例如，在資料庫中，文件 1、文件 2、文件 11、文件 12、文件 20、文件 72 及文件 80 之內容具有關鍵字 “architecture”，因此轉置串列 10 紀錄 1、2、11、12、20、72 及 80 之數值。同理，轉置串列 11 紀錄 1、3、12、13、20、73 及 80 之數值。一般而言，索引檔案所需的空間比較小，可以儲存在記憶體中，而轉置檔案比較大，必須被壓縮並儲存於磁碟中。

當使用者使用轉置索引處理一道檢索命令 (query)

時，資訊檢索系統會先到索引檔案搜尋檢索命令中提及的關鍵字，並取得關鍵字相對應的轉置串列，然後將各相關轉置串列中的文件編號依邏輯運算子(AND、OR 等)予以比對，以找出符合檢索命令的所有文件編號，讓使用者可藉由這些文件編號取得真正的文件。例如，當使用者輸入關鍵字為” architecture <and> computer”時，則資訊檢索系統係對轉置串列 10 及轉置串列 11 進行 AND 運算，找出同時紀錄於轉置串列 10 及轉置串列 11 之數值，以第 1 圖為例，對應關鍵字為” architecture <and> computer”之文件為文件 1、文件 12、文件 20 及文件 80。而關鍵字為” architecture <or> computer”之文件為文件 1、文件 2、文件 3、文件 11、文件 12、文件 13、文件 20、文件 72、文件 73 及文件 80。藉由上述流程，使用者便可於文件資料庫找出欲得到之資料。

由於轉置檔案比較大，必須被壓縮並儲存於磁碟，因此，若將轉置串列的數值進行編碼成較節省儲存空間之不定長度的位元串列流，可以減少磁碟讀取所需之頻寬以加快查詢速度。然而，這也導致轉置串列的數值在解碼過程必須一個數值一個數值的循序解碼，降低比對轉置串列數值的效率。例如當使用者輸入關鍵字為” architecture <and> interface”時，雖然於轉置串列 12 中可得知” interface”於文件 71 之後才會出現，但是資訊檢索系統仍必須將轉置串列 10 之數值一一循序解碼，一直解碼到數值 72 方能進行比對。

為了讓轉置串列編碼後的位元串列流在數值比對過程中，減少多餘的解碼動作，習知技藝會藉由添加額外的資訊或設計特殊的資料結構，在位元串列流加入一

些跳躍點，賦予其可跳躍式解碼的能力。然而習知技藝所提出實現跳躍點的各種方法，所需增加的儲存空間都相當地大，致使無法在位元串列流中插入太多的跳躍點。

有鑑於習知技藝之各項問題，為了能夠兼顧解決之，本發明人基於多年研究開發與諸多實務經驗，提出一種資料搜尋方法及其數值資料編碼方法，可以使用較少的儲存空間，於轉置串列編碼後的位元串列流密集地插入跳躍點，並搭配所提出的資料搜尋方法，以作為改善上述習知技藝缺點之實現方式與依據。

【發明內容】

有鑑於此，本發明之目的就是在提供一種資料搜尋方法及其數值資料編碼方法，以兼具可跳躍式解碼及節省儲存空間之功效。

根據本發明之目的，提出一種數值資料編碼方法，用以對複數個已排序之數值進行編碼，該方法包含下列步驟：設定一分群參數，並根據該分群參數將該些已排序之數值分成複數個群組，每一該些群組之數值數目等於該分群參數，每一該些群組之第一個數值係代表為跳躍點，而其他數值係代表為內部點；根據一第一編碼演譯程序對該些群組之跳躍點進行編碼，以產生複數個跳躍點編碼值；根據一空間預估程序，預估該些群組之該些內部點以一第二編碼演譯程序進行編碼所需之位元保留空間；根據該第二編碼演譯程序對該些群組之該些內部點進行編碼，以產生複數個內部點編碼值；根據該些群組之排列順序及該些位元保留空間，將該些

跳躍點編碼值及該些內部點編碼值依序排列，以產生一輸出編碼值。

此外，本發明更提出一種資料搜尋方法，用以搜尋一編碼數值資料中是否包含一目標數值，該編碼數值資料係以上述之數值資料編碼方法進行編碼，該編碼數值資料包含複數個跳躍點編碼值及複數個內部點編碼值，該方法包含下列步驟：(a)根據第一編碼演譯程序，於該編碼數值資料中取得兩跳躍點編碼值，並解碼出一第一比對跳躍點及一第二比對跳躍點；(b)判斷該目標數值是否介於該第一比對跳躍點及該第二比對跳躍點之間，若是，則執行步驟(c)，若否，則執行步驟(d)；(c)取得位於該第二比對跳躍點隨後之內部點編碼值，並根據一第二編碼演譯程序對該內部點編碼值進行解碼，以取得至少一內部點數值，並判斷該些內部點是否包含該目標數值；(d)根據該第一比對跳躍點、該第二比對跳躍點及一空間預估程序，估算一位元保留空間，並間隔該位元保留空間之後取得另一跳躍點編碼值，並解碼出一另一跳躍點，將該第二比對跳躍點設為該第一比對跳躍點，並將該另一跳躍點設為該第二比對跳躍點，執行步驟(b)。

茲為使 貴審查委員對本發明之技術特徵及所達到之功效有更進一步之瞭解與認識，謹佐以較佳之實施例及配合詳細之說明如後。

【實施方式】

以下將參照相關圖式，說明依本發明較佳實施例之資料搜尋方法及其數值資料編碼方法，為使便於理解，下

述實施例中之相同元件係以相同之符號標示來說明。

請參閱第 2 圖至第 4 圖，第 2 圖為本發明之數值資料編碼方法之步驟流程圖，第 3 圖為本發明之數值序列之分群示意圖，而第 4 圖為本發明之輸出編碼資料之位元串列流示意圖。在第 2 圖中，數值資料編碼方法係用以對複數個已排序之數值進行編碼，此方法包含下列步驟：

步驟 20：設定一分群參數 n ，並根據分群參數 n 將這些已排序之數值分成 m 個群組，如第 3 圖所示，每一群組包含 n 個數值，每一群組之第一個數值係代表為跳躍點，而其他數值係代表為內部點；

步驟 21：根據一第一編碼演譯程序對第 3 圖所示之跳躍點 1 至跳躍點 m 進行編碼，以產生 m 個跳躍點編碼值，如第 4 圖所示之 $E1$ 至 E_m ，其中第 4 圖所示之 $E1$ 至 E_m 分別表示第 3 圖所示之跳躍點 1 至跳躍點 m 根據一第一編碼演譯程序產生之編碼值；

步驟 22：根據一空間預估程序，預估每一第 3 圖所示之內部點群以一第二編碼演譯程序進行編碼所需之位元保留空間，如第 4 圖所示之 $S1$ 及 $S2$ ，其中第 4 圖所示之 $S1$ 及 $S2$ 分別表示第 3 圖所示之內部點群 1 及內部點群 2 根據一空間預估程序，預估以一第二編碼演譯程序進行編碼所需之位元保留空間；

步驟 23：根據一第二編碼演譯程序對每一內部點群進行

編碼，以產生複數個內部點編碼值，如第 4 圖所示之 P_{11} 至 P_{1x} 及 P_{21} 至 P_{2y} ，其中第 4 圖所示之 P_{11} 至 P_{1x} 及 P_{21} 至 P_{2y} 分別表示第 3 圖所示之內部點 D_2 至 D_n 及 D_{n+2} 至 D_{2n} 以一第二編碼演譯程序進行編碼產生之編碼值；

步驟 24：根據這些群組之排列順序及這些位元保留空間，將這些跳躍點編碼值及這些內部點編碼值依序排列，以產生一輸出編碼資料。

其中，若這些群組之最後群組之數值數目小於分群參數時，最後群組之數值係代表殘餘點，以一第一編碼演譯程序進行編碼，如第 4 圖所示之 E_{m+1} 及 E_{m+2} ，其中第 4 圖所示之 E_{m+1} 及 E_{m+2} 分別表示第 3 圖所示之殘餘點 D_{m+2} 及 D_{m+3} 根據一第一編碼演譯程序產生之編碼值。上述空間預估程序係用以預估內部點群以一第二編碼演譯程序進行編碼所需之最大位元保留空間，因此內部點編碼值之位元長度係小於或等於位元保留空間。

如第 4 圖所示，輸出編碼資料之資料結構係先配置前兩個跳躍點編碼值 E_1 及 E_2 ，接著配置內部點編碼值 P_{11} 至 P_{1x} 。若內部點編碼值 P_{11} 至 P_{1x} 所需之空間小於位元保留空間 S_1 ，則保留剩餘空間 T 。續之，於位元保留空間 S_1 之後配置跳躍點編碼值 E_3 及內部點編碼值 P_{11} 至 P_{1y} ，並接著依序配置跳躍點編碼值及內部點編碼值。若內部點編碼值之位元長度小於位元保留空間，則保留剩餘空間。

上述第一編碼演譯程序較佳的是由 d -gap 編碼程序及

prefix-free coding 編碼程序所組成，而第二編碼演譯程序較佳的是 interpolative coding 編碼程序。

請參閱第 5 圖，其係為本發明之數值資料編碼方法之實施例之步驟流程圖。圖中，數值資料編碼方法用以對 11 個已排序之數值 $\langle 5, 8, 12, 13, 15, 18, 23, 28, 29, 32, 33 \rangle$ 進行編碼，此實施例包含下列步驟：

步驟 50：設定一分群參數為 4，將 11 個數值成群為 $\langle 5, 8, 12, 13 \rangle$ 、 $\langle 15, 18, 23, 28 \rangle$ 及 $\langle 29, 32, 33 \rangle$ 共 3 群組，其中 5、15 及 29 為跳躍點，8、12 及 13 是介於 5 與 15 之間的內部點，而 18、23 及 28 是介於 15 與 29 之間的內部點，而 32 及 33 為殘餘點；

步驟 51：使用 d-gap 編碼程序與 prefix-free coding 編碼程序對跳躍點 (5、15、29) 與殘餘點 (32、33) 進行編碼，並分別以 PF(5)、PF (15-5)、PF (29-15)、PF (32-29) 及 PF (33-32) 表示編碼結果，其中，d-gap 編碼程序與 prefix-free coding 編碼程序為熟悉此領域者所熟知，在此不再贅述；

步驟 52：使用 interpolative coding 編碼程序對內部點群 (8、12、13) 及內部點群 (18、23、28) 進行編碼，並分別以 IP(8、12、13) 及 IP(18、23、28) 表示編碼結果，其中，interpolative coding 編碼程序為熟悉此領域者所熟知，在此不再贅述；

步驟 53：使用表 1 所示之空間預估表估算內部點群 (8、

12、13)及內部點群(18、23、28)以 interpolative coding 編碼程序進行編碼所需之較大位元保留空間，其中 g 為分群參數， D 為包圍內部點群之兩端跳躍點之差值再減 1， h 為 $\lceil \log_2(D-2) \rceil - 2$ ，而兩內部點群之位元保留空間之估算結果分別為 8 位元及 10 位元；

$$\text{位元保留空間}(D, g=4) = \begin{cases} 0 & \text{if } D=3 \\ 2 & \text{if } D=4 \\ 3(h+1)+1 & \text{if } 4 < D < 3 \times 2^h + 3 \\ 3(h+1)+2 & \text{if } 3 \times 2^h + 3 \leq D \end{cases}$$

表 1

步驟 54：將跳躍點編碼值及內部點編碼值依序排列，以產生一編碼位元串列流，其中因為 IP(8、12、13)之位元長度為 7 位元，其小於位元保留空間之估算結果，因此需保留 1 位元(8-7=1)，而 IP(18、23、28) 之位元長度為 10 位元，與位元保留空間之估算結果相同，編碼後產生之位元串列流依序為 PF(5)、PF(15-5)、IP(8, 12, 13)、保留位元、PF(29-15)、IP(18, 23, 28)、PF(32-29)及 PF(33-32)。

請參閱第 6 圖，其係為本發明之資料搜尋方法之步驟流程圖。此方法用以搜尋一編碼數值資料中是否包含一目標數值，此編碼數值資料係以上述之數值資料編碼方法進行編碼，此編碼數值資料包含複數個跳躍點編碼值及複數個內部點編碼值，該方法包含下列步驟：

步驟 60：根據第一解碼演譯程序，於該編碼數值資料

中取得兩跳躍點編碼值，並解碼出一第一比對跳躍點及一第二比對跳躍點；

步驟 61：判斷該目標數值是否介於該第一比對跳躍點及該第二比對跳躍點之間，若是，則執行步驟 62，若否，則執行步驟 63；

步驟 62：取得位於該第二比對跳躍點隨後之內部點編碼值，並根據一第二解碼演譯程序對該內部點編碼值進行解碼，以取得至少一內部點數值，並判斷該些內部點是否包含該目標數值；

步驟 63：根據該第一比對跳躍點、該第二比對跳躍點及一空間預估程序，估算一位元保留空間，並間隔該位元保留空間之後取得另一跳躍點編碼值，並解碼出一另一跳躍點，將該第二比對跳躍點設為該第一比對跳躍點，並將該另一跳躍點設為該第二比對跳躍點，執行步驟 61。

上述第一解碼演譯程序較佳的是由 d-gap 解碼程序及 prefix-free coding 解碼程序所組成，而第二解碼演譯程序較佳的是 interpolative coding 解碼程序。

以第 5 圖所示之編碼位元串列流為例，欲搜尋此編碼位元串列流中是否包含一目標數值 23，首先，使用 d-gap 解碼程序與 prefix-free coding 解碼程序，於編碼位元串列流中取得 PF(5)及 PF(15-5)，並解碼出跳躍點 5 及跳躍點 15。因為 23 不介於 5 及 15 之間，因此，接著根據上述表 1，估算

出跳躍點 5 及跳躍點 15 之間之位元保留空間為 8 位元，因此自 PF(15-5)隨後間隔 8 位元，取得 PF(29-15) 並解碼出跳躍點 29。因為 23 介於 15 及 29 之間，因此接著取出 IP(18, 23, 28)，並解碼出內部點 18、23 及 28，並判斷出此編碼位元串列流包含目標數值 23。

以上所述僅為舉例性，而非為限制性者。任何未脫離本發明之精神與範疇，而對其進行之等效修改或變更，均應包含於後附之申請專利範圍中。

【圖式簡單說明】

- 第 1 圖 係繪示一索引結構之示意圖；
- 第 2 圖 係為本發明之數值資料編碼方法之步驟流程圖；
- 第 3 圖 係為本發明之數值序列之分群示意圖；
- 第 4 圖 係為本發明之輸出編碼資料之位元串列流示意圖；
- 第 5 圖 係為本發明之數值資料編碼方法之實施例之步驟流程圖；以及
- 第 6 圖 係為本發明之資料搜尋方法之步驟流程圖。

【主要元件符號說明】

10, 11, 12：轉置串列；

20~24：步驟流程；

S1, S2：位元保留空間；

T：保留位元；

D1, D2, Dn, Dn+1, Dn+2, D2n, Dmn+1, Dmn+2, Dmn+3：跳躍點編碼值；

$E_1, E_2, E_3, E_m, E_{m+1}, E_{m+2}$: 跳躍點編碼值 ;

$P_{11}, P_{1x}, P_{21}, P_{2y}$: 內部點編碼值 ;

50~54 : 步驟流程 ; 以及

60~63 : 步驟流程。

十、申請專利範圍：

- 1、一種數值資料編碼方法，用以對複數個已排序之數值進行編碼，該方法包含下列步驟：

設定一分群參數，並根據該分群參數將該些已排序之數值分成複數個群組，每一該些群組之數值數目等於該分群參數，每一該些群組之第一個數值係代表為跳躍點，而其他數值係代表為內部點；

根據一第一編碼演譯程序對該些群組之跳躍點進行編碼，以產生複數個跳躍點編碼值；

根據一空間預估程序，預估該些群組之該些內部點以一第二編碼演譯程序進行編碼所需之位元保留空間；

根據該第二編碼演譯程序對該些群組之該些內部點進行編碼，以產生複數個內部點編碼值；以及

根據該些群組之排列順序及該些位元保留空間，將該些跳躍點編碼值及該些內部點編碼值依序排列，以產生一輸出編碼值。

- 2、如申請專利範圍第1項所述之數值資料編碼方法，其中該第一編碼演譯程序係為 d-gap 編碼程序及 prefix-free coding 編碼程序所組成。
- 3、如申請專利範圍第1項所述之數值資料編碼方法，其中該第二編碼演譯程序係為 interpolative coding 編碼程序。
- 4、如申請專利範圍第1項所述之數值資料編碼方法，其中當該些群組之最後群組之數值數目小於該分群參數時，該最後群組之數值係以該第一編碼演譯程序進行編碼。
- 5、一種資料搜尋方法，用以搜尋一編碼數值資料中是否包含一目標數值，該編碼數值資料係以申請專利範圍第1項所述之數值資料編碼方法進行編碼，該編碼數值資料包含複

數個跳躍點編碼值及複數個內部點編碼值，該方法包含下列步驟：

(a)根據第一解碼演譯程序，於該編碼數值資料中取得兩跳躍點編碼值，並解碼出一第一比對跳躍點及一第二比對跳躍點；

(b)判斷該目標數值是否介於該第一比對跳躍點及該第二比對跳躍點之間，若是，則執行步驟(c)，若否，則執行步驟(d)；

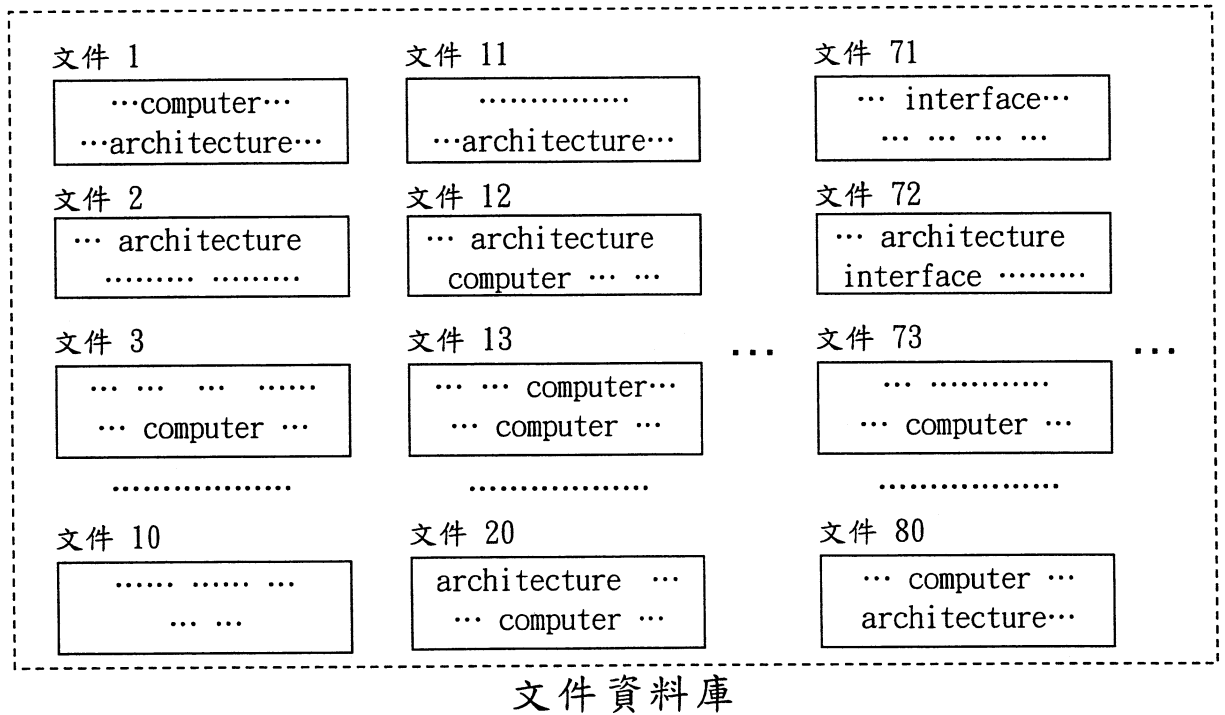
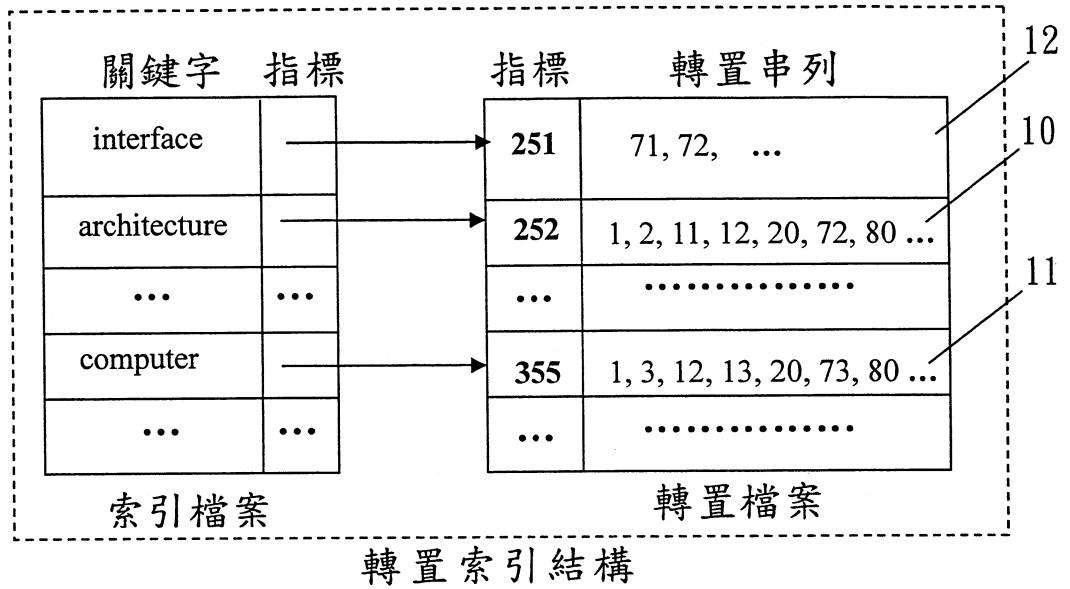
(c)取得位於該第二比對跳躍點隨後之內部點編碼值，並根據一第二解碼演譯程序對該內部點編碼值進行解碼，以取得至少一內部點數值，並判斷該些內部點是否包含該目標數值；以及

(d)根據該第一比對跳躍點、該第二比對跳躍點及一空間預估程序，估算一位元保留空間，並間隔該位元保留空間之後取得另一跳躍點編碼值，並解碼出一另一跳躍點，將該第二比對跳躍點設為該第一比對跳躍點，並將該另一跳躍點設為該第二比對跳躍點，執行步驟(b)。

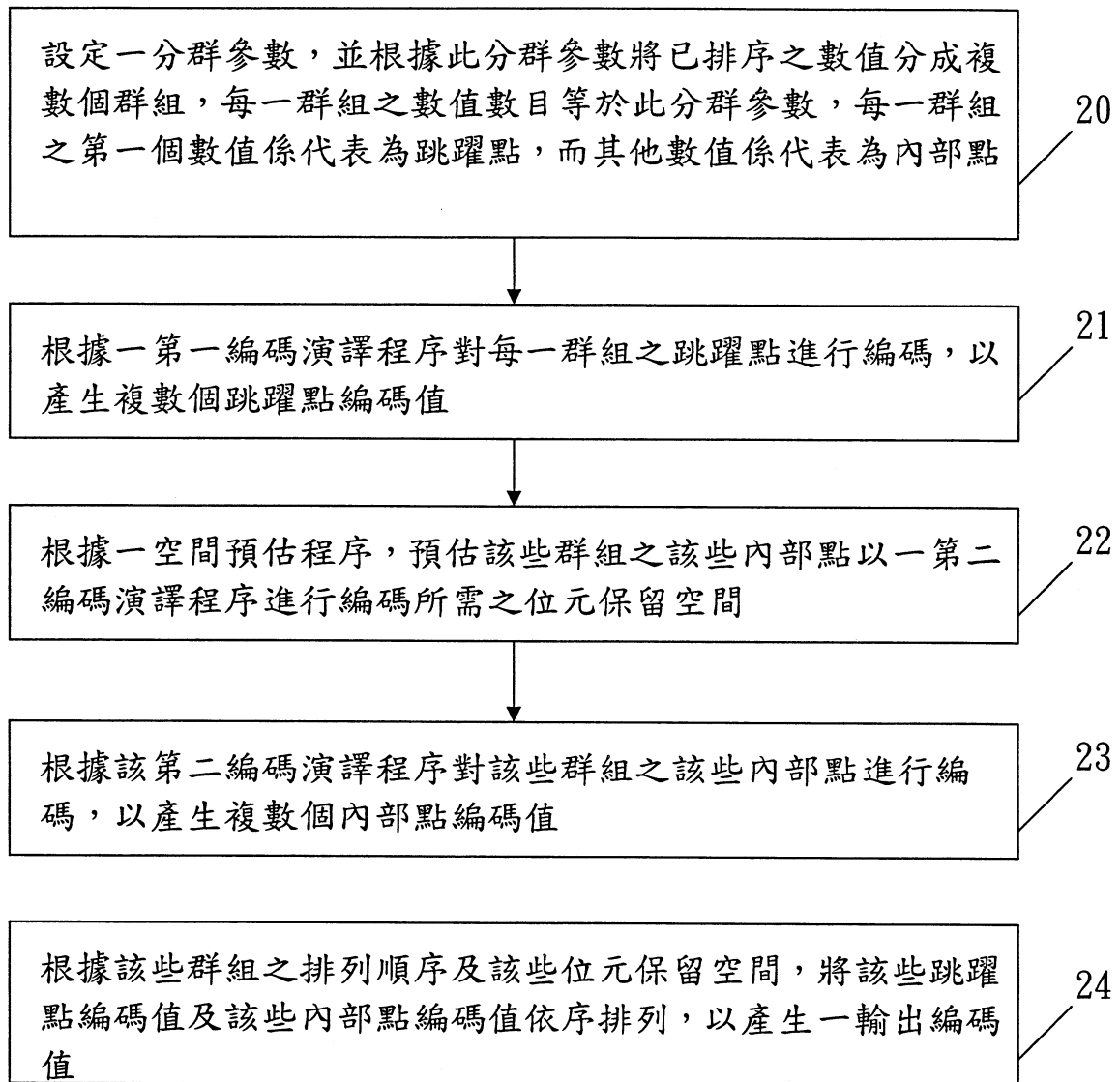
6、如申請專利範圍第5項所述之資料搜尋方法，其中該第一解碼演譯程序係為 d-gap 解碼程序及 prefix-free coding 解碼程序所組成。

7、如申請專利範圍第5項所述之資料搜尋方法，其中該第二解碼演譯程序係為 interpolative coding 解碼程序。

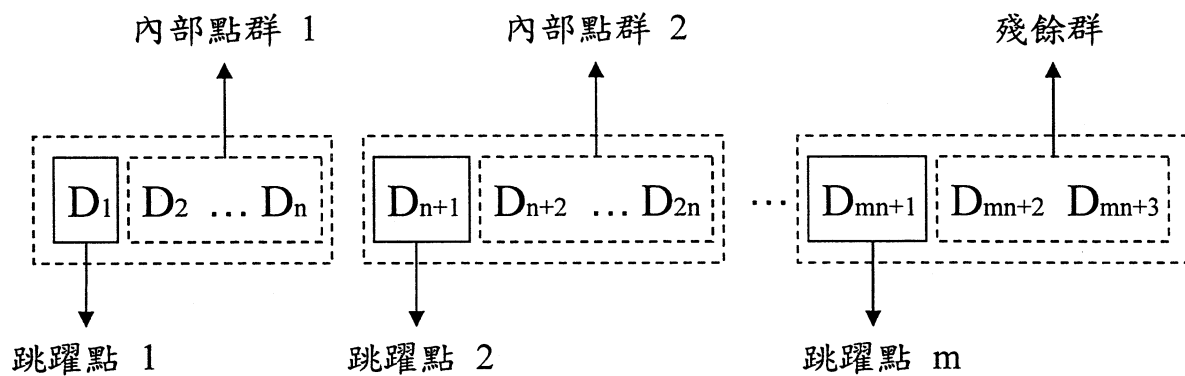
十一、圖式：



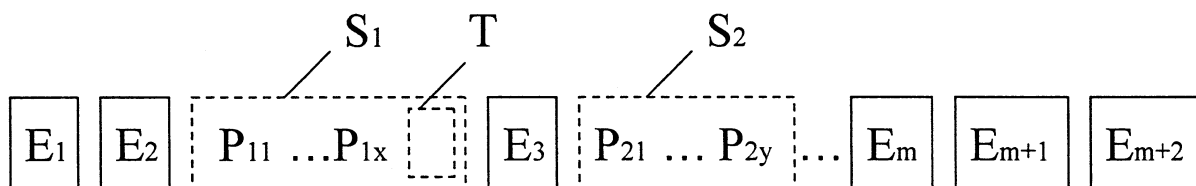
第 1 圖



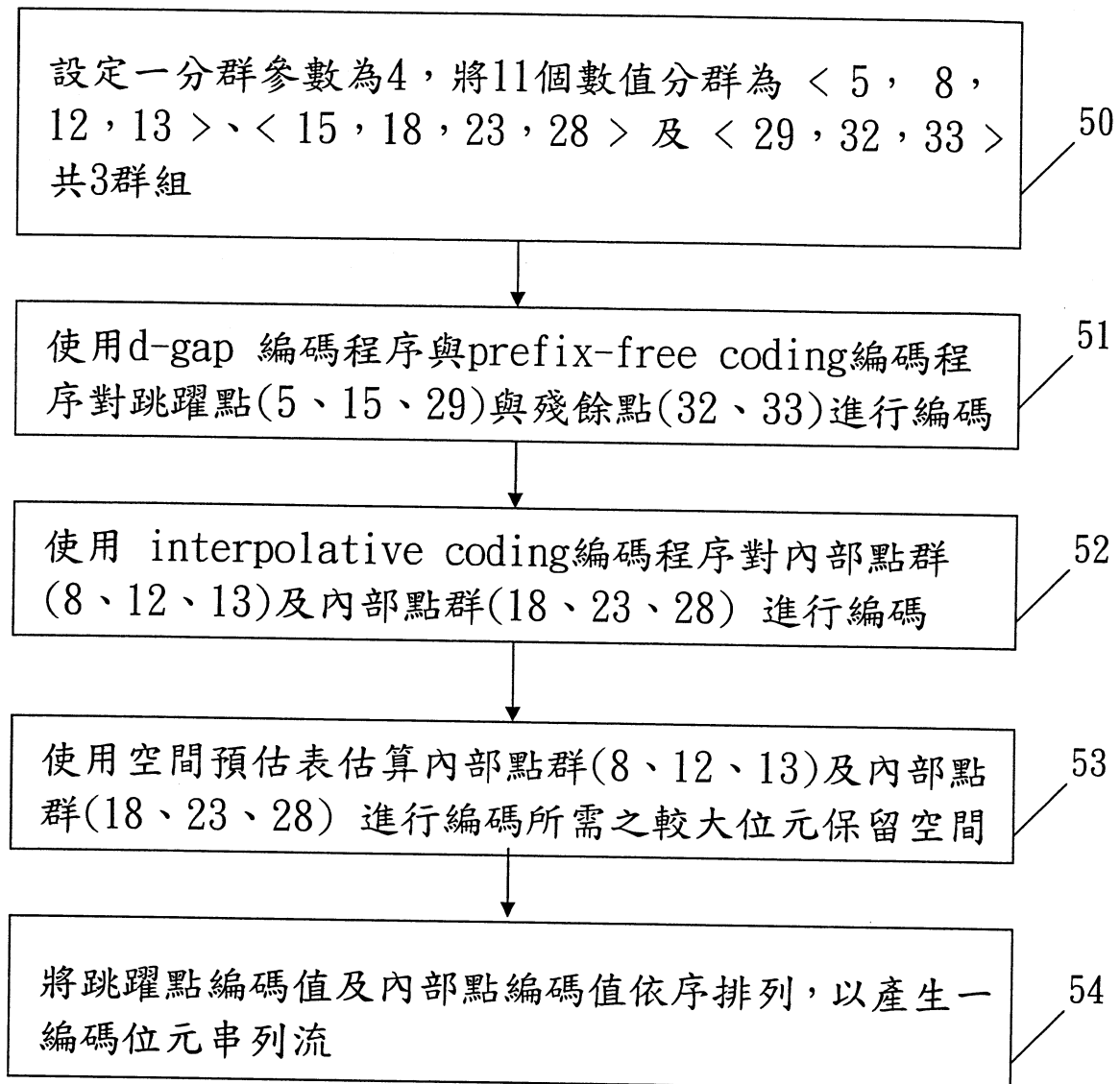
第 2 圖



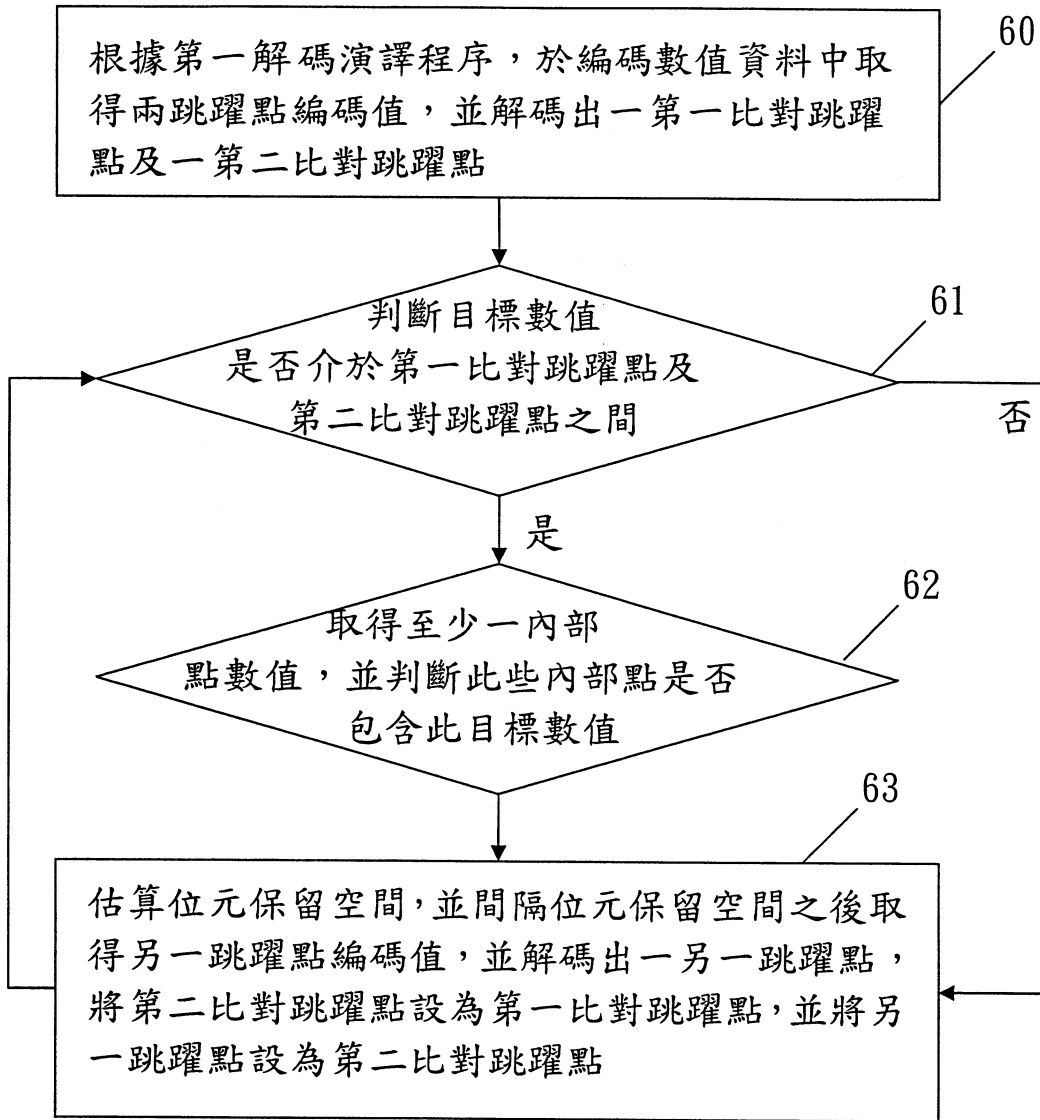
第 3 圖



第 4 圖



第 5 圖



第 6 圖