

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：95142250

※ 申請日期：95.11.15

※IPC 分類：606F 7/00 (2006.01)

606F 17/30 (2006.01)

606F 17/20 (2006.01)

一、發明名稱：(中文/英文)

利用 Bloom filter 達到次線性運算時間的字串比對系統及方法

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 4 人)

姓 名：(中文/英文)

ID :

1. 林柏青/ LIN, PO-CHING

R120681373

2. 林盈達/ LIN, YING-DAR

P120502982

3. 鄭伊君/ZHENG, YI-JUN

G221432430

4. 賴源正/LAI, YUAN-CHENG

N123711133

國 籍：(中文/英文)

1. 中華民國/ ROC

2. 中華民國/ ROC

3. 中華民國/ ROC

4. 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：

本發明係揭露一種字串比對系統，其包含一待測字串、複數個特徵字串、一 M 位元組搜尋視窗及複數個過濾模組。其中，M 位元組搜尋視窗係用以由待測字串中擷取一 M 位元組字串，而每一過濾模組係包含複數個特徵字串之子字串，用以與 M 位元組字串進行比對，以輸出 M 位元組搜尋視窗之一位移長度。藉由一次移動搜尋視窗多個位元組，可達到同時比對多個位元組的加速效果。同時，利用 Bloom filter 這種較節省空間的儲存特徵字串方式，可以讓大量的特徵字串能存在單一晶片的內嵌式記憶體當中。

六、英文發明摘要：

七、指定代表圖：

(一)本案指定代表圖為：第(二)圖。

(二)本代表圖之元件符號簡單說明：

21：待測字串；

22：M位元組搜尋視窗；

221：M位元組字串；

23：過濾模組；

24：特徵字串；

241：特徵字串之子字串；以及

25：位移長度。

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】

本發明係為提供一種字串比對系統及方法，特別是一種利用複數個過濾模組來比對字串，並決定搜尋視窗位移長度，以達到在次線性運算時間硬體加速目的之字串比對的系統及方法。

【先前技術】

隨著網路技術的發展，人們使用網路的機會越來越多，網路上的資訊流通量也與日俱增，但是，也由於人們仰賴網路越來越深，因此網路的入侵與攻擊會嚴重影響正常社會的運作。例如，攻擊政府的網站或癱瘓各式各樣的伺服器，甚至個人的電腦也有可能遭到病毒的侵襲。

於是，近年來許多資訊內容安全設備，例如網路入侵偵測系統 (Network Intrusion Detection System, NIDS)、掃毒系統 (Antivirus system) 逐漸發展成為網路安全上的重要技術。在此技術中，網路封包的字串與特徵字串的比對效率是決定這類系統效能的關鍵。

若字串的比對速度太慢，會降低網路的流量，造成依賴網路的工作無法完成，若字串的比對不夠詳細或有所遺漏，則會使網路受到攻擊的機會大大增加。

請參閱第一圖，係顯示習知字串比對方法之步驟流程圖，其步驟如後：

步驟 S11：開始；

步驟 S12：利用搜尋視窗的最後若干個位元組構成一個區塊，比對位移距離表格，得到位移長度 N；

步驟 S13：位移長度是否為零，若是，執行步驟 S15，

若否，則執行步驟 S14；

步驟 S14：滑動窗格向後移 N 個位元組，返回上述步驟 S12；

步驟 S15：比對規則雜湊值表格；

步驟 S16：是否有相同的字串，若是，執行步驟 S17，若否，則執行步驟 S18；

步驟 S17：將對應的比對成功旗標設為真；

步驟 S18：滑動窗格向後移 1 個位元組，返回上述步驟 S12；以及

步驟 S19：輸出結果。

其中，習知作法係透過一位移距離表格，以查表的方式來決定滑動窗格每次移動的長度，此表格之建立需耗費大量的記憶體空間來儲存各種不同區塊對應到的位移長度，且透過單一區塊進行比對在特徵字串很大時，可能造成需頻繁的進行驗證，而使得比對速度拖慢，造成比對效率不佳。

為解決上述所提出的問題，本發明人基於多年從事研究與諸多實務經驗，經多方研究設計與專題探討，遂於本發明提出一種字串比對系統及方法，以作為前述期望一實現方式與依據。

【發明內容】

有鑑於上述課題，本發明之目的為提供一種字串比對系統及方法，特別是一種利用複數個過濾模組來比對字串，並決定搜尋視窗位移長度，以達到硬體加速目的之字串比對系統及方法。

緣是，為達上述目的，依本發明之字串比對系統，其包含一待測字串、複數個特徵字串、一 M 位元組搜尋視窗及複數個過濾

模組。其中，M 位元組搜尋視窗係用以由待測字串中擷取一 M 位元組字串。而每一過濾模組係包含複數個特徵字串之子字串，用以與 M 位元組字串進行比對，以輸出 M 位元組搜尋視窗之一位移長度。

承上所述，因依本發明之字串比對系統及方法，其利用演算法的啟示法則 (heuristic)，在不提高硬體複雜度的前提下，略過不可能比對成功的字元，來達到效果上同時比對多個字元的目的，且透過 Bloom filter 這類有效率的儲存方式，將大量的特徵字串儲存於內嵌式記憶體中，無需建立位移距離表格，可節省記憶體空間。

茲為使 貴審查委員對本發明之技術特徵及所達成之功效有更進一步之瞭解與認識，下文謹提供較佳之實施例及相關圖式以為輔佐之用，並以詳細之說明文字配合說明如後。

【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字串比對系統及方法，其中相同的元件將以相同的參照符號加以說明。

請參閱第二圖，係顯示本發明之字串比對系統之方塊圖，其包含一待測字串 21、複數個特徵字串 24、一 M 位元組搜尋視窗 22 及複數個過濾模組 23。其中，M 位元組搜尋視窗 22 係用以由待測字串 21 中擷取一 M 位元組字串 221，而每一過濾模組 23 係包含複數個特徵字串之子字串 241，用以與 M 位元組字串 221 進行比對，以輸出 M 位元組搜尋視窗 22 之一位移長度 25。

上述過濾模組較佳為一 Bloom Filter，上述 M 位元組搜尋視窗之長度係等同於特徵字串中最小長度之特徵字串之字串長度，

上述每一特徵字串係以 Q 之長度，分割為 $M-Q+1$ 個子字串，並將此些子字串依其在特徵字串中的位置分群後，分別儲存在過濾模組中。此外也把上述每一特徵字串以其前 i 個字的字首進行分群為複數個子字串，其中 $i=1, \dots, (Q-1)$ ，並分別儲存在過濾模組中，上述字串比對系統更包含一優先權設定模組，當過濾模組產生兩個以上之位移長度時，優先權設定模組係用以判定位移長度之大小，並優先輸出較小之位移長度。

請參閱第三圖，係顯示本發明之字串比對系統之較佳實施例之示意圖，如圖示， P_1 、 P_2 、 P_3 為特徵字串，要搜尋的本文分批暫存於一個足夠大的本文緩衝區 31，在本文緩衝區 31 中，移動一個搜尋視窗 32，並比對視窗內的字元看看是否有出現特徵字串集當中的特徵字串，如果特徵字串集中的字串長度非完全相同，假定其最小長度為 M ，那麼在比對中只考慮每個特徵字串的字首長度為 M 的部分，並將搜尋視窗 32 的長度也設為 M ，視窗移動的位置由字元計數器 33 控制，前一批緩衝區內的字尾會成為下一批緩衝區的起頭重複比對，以避免特徵字串橫跨在前後兩批緩衝區當中，使得某些特徵字串無法必對到。在一般情形下，搜尋視窗中的字元無須逐字比對，對於不可能比對成功的字元可直接略過，並且將搜尋視窗挪後若干字元。

為便於利用演算法中的啟示法則來決定搜尋視窗 32 的移動幅度，將特徵字串集當中的每個特徵字串看成由 $M-Q+1$ 個區塊構成，其中 Q 是區塊長度，以本較佳實施例示意圖來說 $Q=4$ ，因此 $P_1=abcdefgh$ 可視為 $abcd-bcde-cdef-defg-efgh$ 共 5 個區塊，其餘特徵字串以此類推，每個區塊依照它在特徵字串中的位置分群後，分別儲存在不同的 Bloom Filter 35 當中，此外也把每個特徵字串中的前 i 個字的字首分群， $i=1..(Q-1)$ ，也儲存於各自的 Bloom Filter 35 當中，在本實施例中，特徵字串集 $\{P_1, P_2, P_3\}$ 內

的區塊及特徵字串的字首分為 G_0 至 G_7 共 8 群，分群的結果如圖示。

在比對時從搜尋視窗 32 最右邊長度為 Q 的區塊開始找起，同時洽詢 Bloom Filter 35 比對最右邊的區塊是否出現在 G_0 至 G_4 之中，以及該區塊的字尾是否在對應的 G_5 至 G_7 之中。假定代表 G_i 的 Bloom Filter 35 洽詢成功（可能有超過一個以上的 G_i 會洽詢成功），即區塊或特徵字串字尾極可能出現在 G_i 當中。若 $i > 0$ ，則搜尋視窗 32 可移動幅度為 i 的最小值；若沒有任何一個 G_i 洽詢成功，則搜尋視窗 32 移動幅度為 M 個字元，若 $i = 0$ ，則進入下一階段的比對程序，上述啟示法則摘要如下：

$$\text{搜尋視窗移動幅度} = \begin{cases} \min\{i \mid G_i \text{ 洽詢成功}\}, i \text{ 的最小值} > 0 \\ m, \text{ 無任何 } G_i \text{ 洽詢成功} \\ \text{進入下階段比對}, i \text{ 的最小值} = 0 \end{cases}$$

在此系統下，即使 Bloom Filter 35 發生誤判，在這個啟示法則中只是讓搜尋視窗的移動幅度較小而已，但是不會因而漏掉任何一個該找到的特徵字串。以本實施例來說，搜尋視窗最右邊的“exam”區塊，既不存在特徵字串的任何區塊當中，它的字尾也不是任何特徵字串的字首，因此可以安然將搜尋視窗移動 8 個字元而不用擔心漏失任何特徵字串。

根據上述啟示法則進行洽詢之後，將各 Bloom Filter 35 的洽詢結果送往優先權設定模組 36，找出最小值後，送往位移控制器 34 計算新的字元計數器 33 的值，相當於移動搜尋視窗 32 至下一個位置。由於搜尋視窗 32 一次可略過多個字元，因此在效果上相當於同時比對處理多個字元，若搜尋視窗 32 超出本文緩衝區 31 的範圍，則載入下一批本文至緩衝區繼續搜尋，視實際應用之需

要，在本文中找出一個或多個特徵字串後（也有可能沒比對到任何特徵字串），結束搜尋工作。

若前述使 G_i 洽詢成功之 i 的最小值為 0，則進入下階段的比對程序。在此程序中從最右至左算起，逐步比對搜尋視窗之倒數第二個區塊、倒數第三個區塊…依此類推，藉由同時洽詢各 Bloom Filter 35 找出他們在各個 G_i 中的位置，比對的結果有下列兩種可能：

(1) 若由右至左倒數第 i 個區塊洽詢代表 G_{i-1} 的 Bloom Filter 成功，這表示搜尋視窗 32 的內容仍有可能是某一特徵字串的一部份，則繼續向左比對，直到搜尋視窗 32 內所有的區塊都比對完畢且洽詢 Bloom Filter 皆成功為止，此時才進入驗證程序。

(2) 若由右至左比對至倒數第 i 個區塊洽詢代表 G_{i-1} 的 Bloom Filter 失敗，表示搜尋視窗 32 的內容已經不可能存在於任一特徵字串，則無繼續比對之必要。此時，檢查該區塊在代表 G_j 中的哪一個 Bloom Filter 35 洽詢成功，其中 $j > i-1$ 。若有兩個以上的 Bloom Filter 35 洽詢成功，搜尋視窗 32 可移動的幅度為 $j-i+1$ ，反之則為 $m-i+1$ 。

若搜尋視窗 32 中的每一個區塊洽詢對應的 Bloom Filter 皆成功，表示視窗內的文字很可能是某特徵字串的一部份或全部。此時才進入驗證程序，而非在最右邊的區塊一洽詢成功就立刻驗證，因此可降低驗證的次數，減低外部攻擊者蓄意製造較差的情形使系統耗費大量的時間在驗證上。

驗證時經由一個非阻斷式 (non-blocking) 的驗證介面將這段文字交由驗證模組進行驗證，同時搜尋視窗則移動一個字元的位置繼續搜尋無須等待，因此可避免驗證程序耽誤搜尋時間。

請參閱第四圖，係顯示本發明之字串比對方法之步驟流程

圖，其步驟如後：

步驟 S41：提供複數個特徵字串；

步驟 S42：提供一待測字串；

步驟 S43：透過一 M 位元組搜尋視窗，由待測字串中擷取一 M 位元組字串；

步驟 S44：透過複數個過濾模組與 M 位元組字串進行比對，以輸出 M 位元組搜尋視窗之一位移長度，其中，每一過濾模組係包含複數個特徵字串之子字串；

步驟 S45：判斷位移長度是否為零，若是，執行步驟 S47，若否，執行步驟 S46；

步驟 S46：依照位移長度移動 M 位元組搜尋視窗；

步驟 S47：從搜尋視窗由右至左依序用區塊洽詢過濾模組做比對，並判斷是否找到移動長度大於零的區塊，若是，執行步驟 S46，若否，執行步驟 S48；

步驟 S48：待所有搜尋視窗內的區塊比對完成後，進行驗證，移動 M 位元組搜尋視窗一個字元；以及

步驟 S49：待測字串是否已全部完成比對，若是，則結束比對，若否，則重複上述步驟 S43。

其中，上述過濾模組較佳為一 Bloom Filter，上述 M 位元組搜尋視窗之長度係等同於特徵字串中最小長度之特徵字串之字串長度，上述每一特徵字串係以 Q 之長度，分割為 $M-Q+1$ 個子字串，並將此些子字串依其在特徵字串中的位置分群後，分別儲存在過濾模組中。此外也把每一特徵字串以其前 i 個字的字首進行分群為複數個子字串，其中 $i=1, \dots, (Q-1)$ ，並分別儲存在過濾模組中，上述字串比對系統更包含一優先權設定模組，當過濾模組產

生兩個以上之位移長度時，優先權設定模組係用以判定位移長度之大小，並優先輸出較小之位移長度。

請參閱第五圖，係顯示本發明之字串比對方法之較佳實施例之步驟流程圖，其步驟如後：

步驟 S51：是否已比對完待測字串所有字元或已在待測字串中找到特徵字串，若是，結束比對，若否，則執行步驟 S52；

步驟 S52：進行第一階段的比對，使用搜尋視窗最右邊的區塊同時洽詢複數個過濾模組，依啟示法則找出搜尋視窗的移動長度；

步驟 S53：移動長度是否為零，若是，執行步驟 S55，若否，執行步驟 S54；

步驟 S54：移動搜尋視窗，返回上述步驟 S51；

步驟 S55：進行第二階段的比對，從搜尋視窗由右至左依序用區塊洽詢過濾模組做比對；

步驟 S56：是否找到移動長度大於零的區塊，若是，執行步驟 S54，若否，執行步驟 S57；以及

步驟 S57：送交驗證模組進行驗證，並移動搜尋視窗一個字元，再回到上述步驟 S51。

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】

第一圖係顯示習知字串比對方法之步驟流程圖；

第二圖係顯示本發明之字串比對系統之方塊圖；

第三圖係顯示本發明之字串比對系統之較佳實施例之示意圖；

第四圖係顯示本發明之字串比對方法之步驟流程圖；以及

第五圖請參閱第五圖，係顯示本發明之字串比對方法之較佳實施例之步驟流程圖。

【主要元件符號說明】

S11~S19：步驟流程；

21：待測字串；

22：M位元組搜尋視窗；

221：M位元組字串；

23：過濾模組；

24：特徵字串；

241：特徵字串之子字串；

25：位移長度；

31：本文緩衝區；

32：搜尋視窗；

33：字元計數器；

34：位移控制器；

35：Bloom Filter；

36：優先權設定模組；

S41～S48：步驟流程；以及

S51～S57：步驟流程。

十、申請專利範圍：

1、一種字串比對系統，至少包含：

一待測字串；

複數個特徵字串；

一 M 位元組搜尋視窗，係用以由該待測字串中擷取一 M 位元組字串；以及

複數個過濾模組，每一該過濾模組係包含複數個該些特徵字串之子字串，用以與該 M 位元組字串進行比對，以輸出該 M 位元組搜尋視窗之一位移長度。

2、如申請專利範圍第 1 項所述之字串比對系統，其中該過濾模組係為一 Bloom Filter。

3、如申請專利範圍第 1 項所述之字串比對系統，其中該 M 位元組搜尋視窗之長度係等同於最小長度之該特徵字串之字串長度。

4、如申請專利範圍第 1 項所述之字串比對系統，其中當該位移長度為 N 時，該 M 位元組搜尋視窗係於該待測字串向後移動 N 個位元組，以擷取下一個 M 位元組字串。

5、如申請專利範圍第 1 項所述之字串比對系統，其中每一該特徵字串係以 Q 之長度，分割為 $M-Q+1$ 個子字串，並將該些子字串依其在該特徵字串中的位置分群後，分別儲存在該些過濾模組中。

6、如申請專利範圍第 5 項所述之字串比對系統，其中每一該特徵字串係以其前 i 個字的字首進行分群為複數個子字串，其中 $i=1, \dots, (Q-1)$ ，並分別儲存在該些過濾模組中。

7、如申請專利範圍第 1 項所述之字串比對系統，其中更包含一優先權設定模組，當該些過濾模組產生兩個以上之位移長度時，該優先權設定模組係用以判定位移長度之大小，

並優先輸出較小之位移長度。

8、一種字串比對方法，至少包含：

(a)提供複數個特徵字串；

(b)提供一待測字串；

(c)透過一 M 位元組搜尋視窗，由該待測字串中擷取一 M 位元組字串；

(d)透過複數個過濾模組與該 M 位元組字串進行比對，以輸出該 M 位元組搜尋視窗之一位移長度，其中，每一該過濾模組係包含複數個該些特徵字串之子字串；以及

(e)若該位移長度為零，則從該 M 位元組搜尋視窗由右至左依序用區塊洽詢該些過濾模組做比對，當一移動長度未大於零時，則待該些區塊比對完成後，進行驗證，並移動該 M 位元組搜尋視窗一個字元，若該位移長度為非零，或該移動長度大於零，則依照該位移長度移動該 M 位元組搜尋視窗；

(f)重複上述步驟(c)至步驟(e)，直到該待測字串已全部完成比對。

9、如申請專利範圍第 8 項所述之字串比對方法，其中該過濾模組係為一 Bloom Filter。

10、如申請專利範圍第 8 項所述之字串比對方法，其中該 M 位元組搜尋視窗之長度係等同於最小長度之該特徵字串之字串長度。

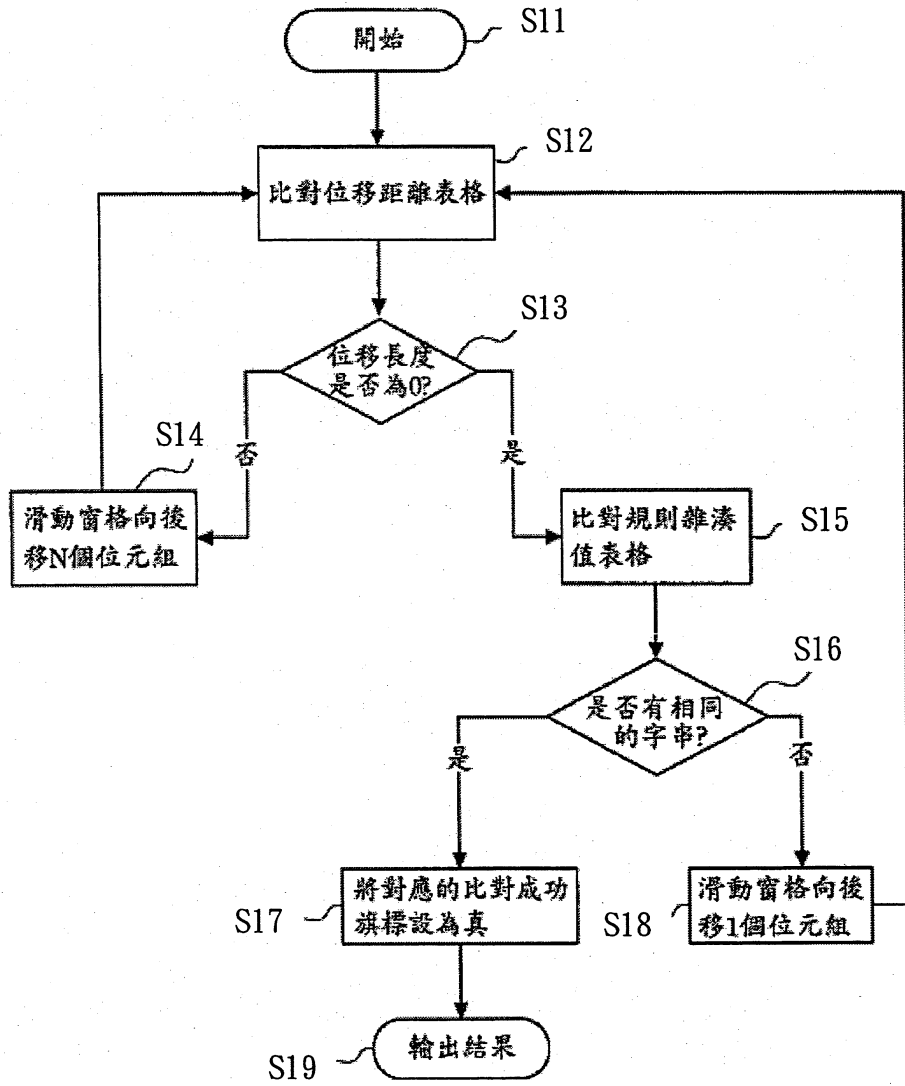
11、如申請專利範圍第 8 項所述之字串比對方法，其中當該位移長度為 N 時，該 M 位元組搜尋視窗係於該待測字串向後移動 N 個位元組，以擷取下一個 M 位元組字串。

12、如申請專利範圍第 8 項所述之字串比對方法，其中每

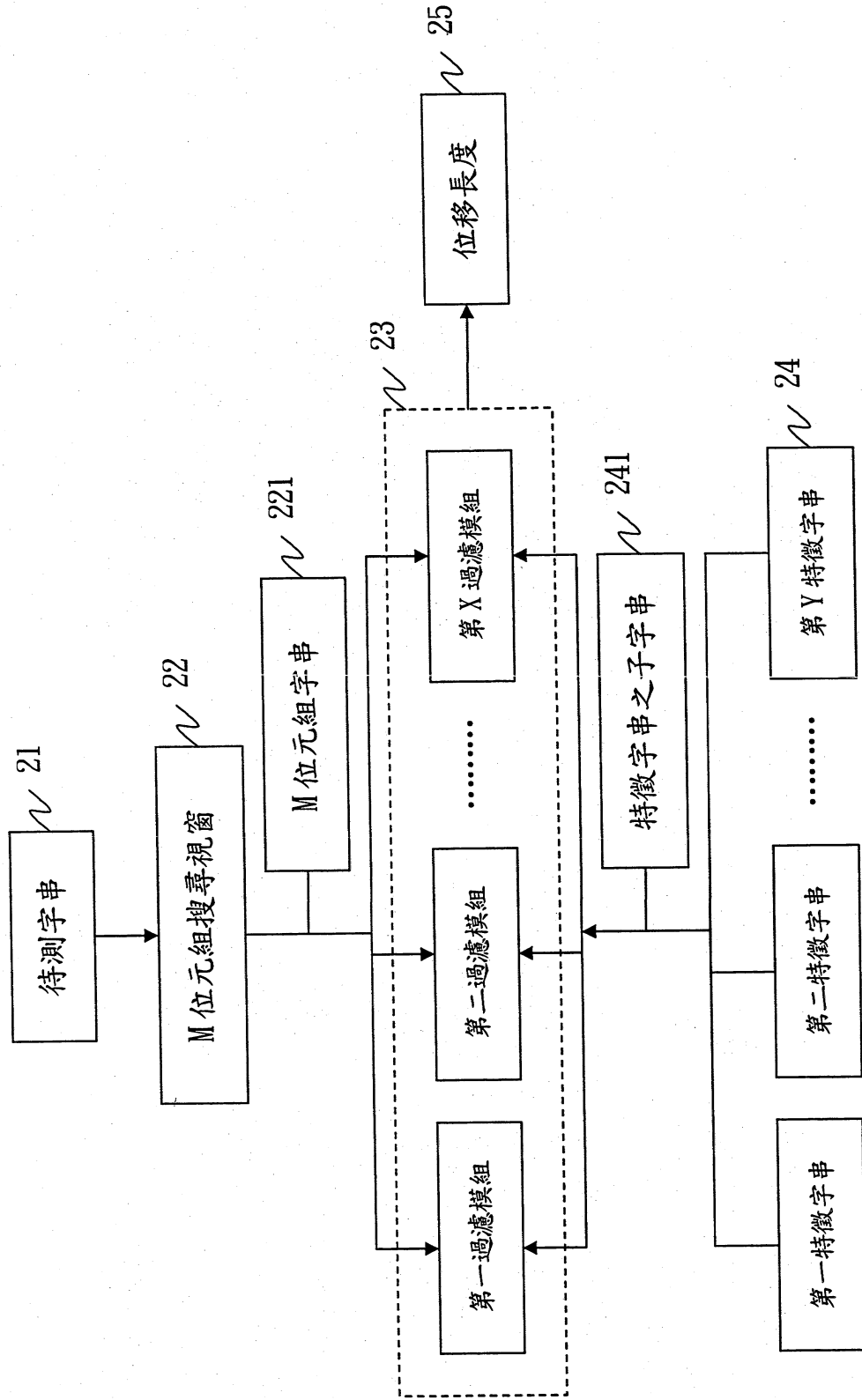
一該特徵字串係以 Q 之長度，分割為 $M-Q+1$ 個子字串，並將該些子字串依其在該特徵字串中的位置分群後，分別儲存在該些過濾模組中。

- 13、 如申請專利範圍第 12 項所述之字串比對方法，其中每一該特徵字串係以其前 i 個字的字首進行分群為複數個子字串，其中 $i=1, \dots, (Q-1)$ ，並分別儲存在該些過濾模組中。
- 14、 如申請專利範圍第 8 項所述之字串比對方法，其中更包含提供一優先權設定模組，當該些過濾模組產生兩個以上之位移長度時，該優先權設定模組係用以判定位移長度之大小，並優先輸出較小之位移長度。

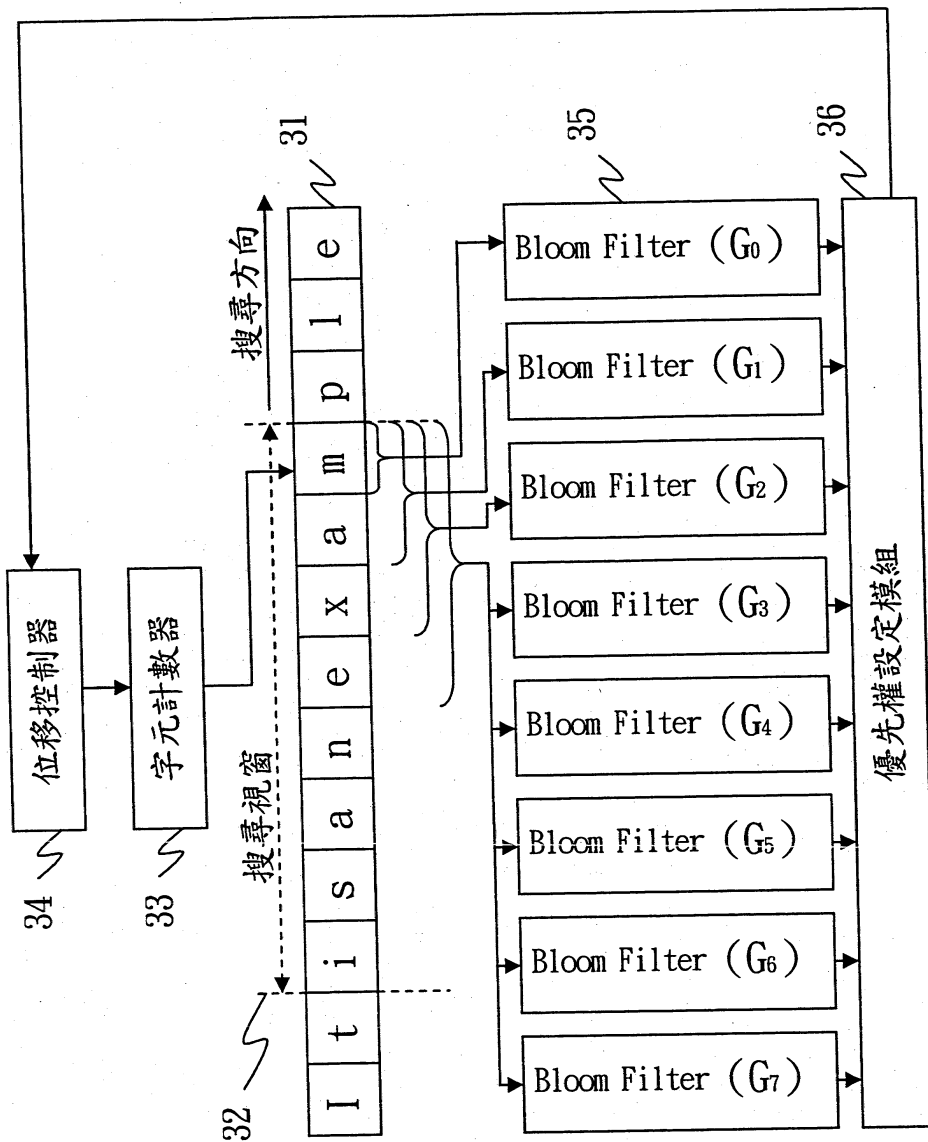
十一、圖式：



第一圖

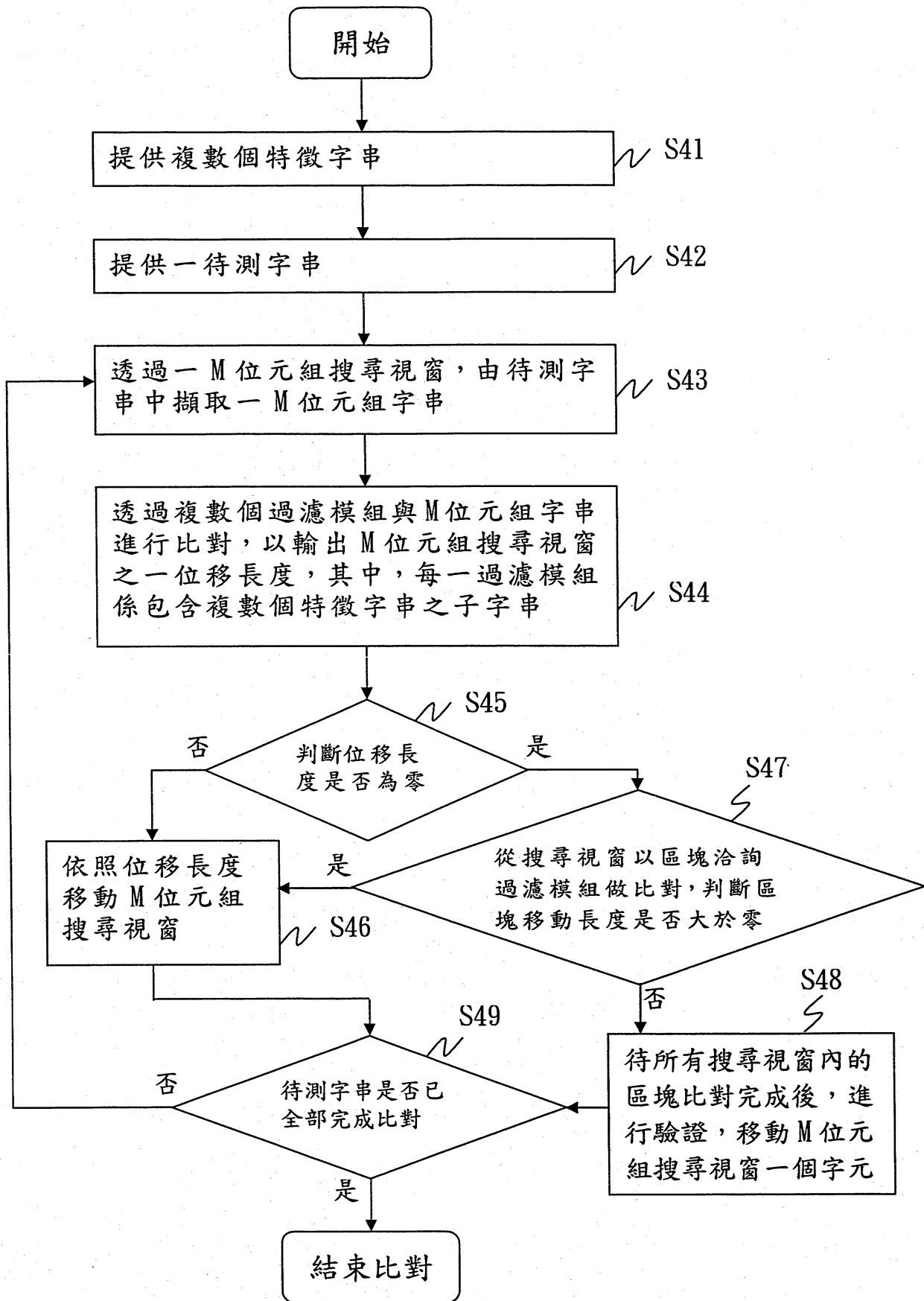


第二圖

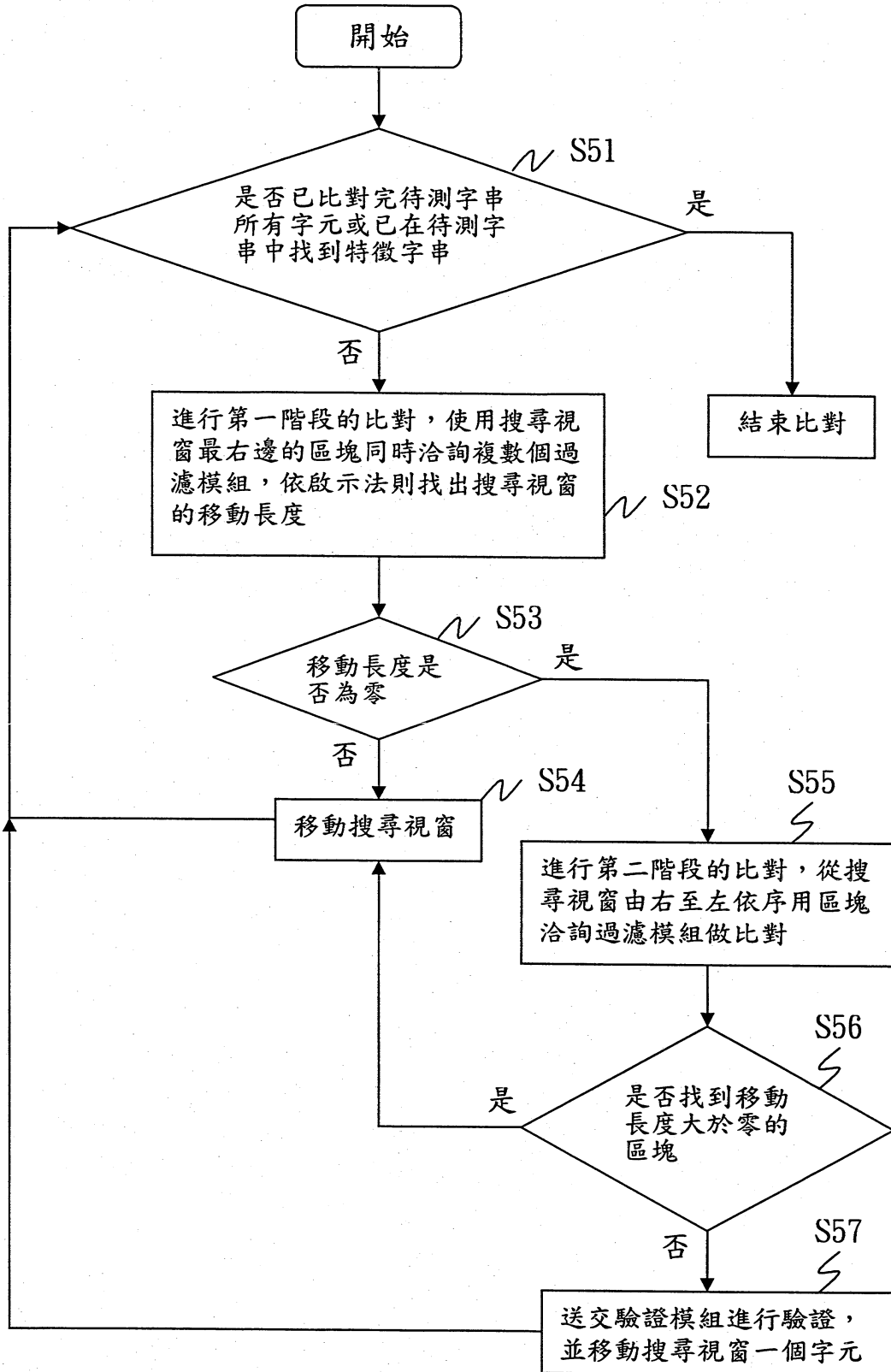


- $P_1 = abcdefgh$
- $P_2 = ijklmnop$
- $P_3 = zyxwvuts$
- $G_0 = \{efgh, mnop, vuts\}$
- $G_1 = \{defg, lmn, wvut\}$
- $G_2 = \{cdef, klmn, xwvu\}$
- $G_3 = \{bcde, jklm, yxwv\}$
- $G_4 = \{abcd, ijkl, zyxw\}$
- $G_5 = \{abc, ijk, zyx\}$
- $G_6 = \{ab, ij, zy\}$
- $G_7 = \{a, i, z\}$

第三圖



第四圖



第五圖

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：95142250

※ 申請日期：

※IPC 分類：

一、發明名稱：(中文/英文)

利用 Bloom filter 達到次線性運算時間的字串比對系統及方法

二、申請人：(共 1 人)

姓名或名稱：(中文/英文)

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) 吳重雨/WU CHUNG-YU

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

三、發明人：(共 4 人)

姓 名：(中文/英文)

1. 林柏青/ LIN, PO-CHING
2. 林盈達/ LIN, YING-DAR
3. 鄭伊君/ZHENG, YI-JUN
4. 賴源正/LAI, YUAN-CHENG

國 籍：(中文/英文)

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