公告本

I234086

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：92130036

※ 申請日期：92.10.29 ※IPC 分類：G06F17/21

壹、發明名稱：（中文/英文）

網頁內容過濾方法

貳、申 請 人：（共 1 人）

姓名或名稱：（中文/英文）

國立交通大學

代表人：（中文/英文）張俊彥

住居所或營業所地址：（中文/英文）新竹市大學路 1001 號

國　籍：（中文/英文）　中華民國

參、發明人：（共 3 人）

姓　名：（中文/英文）

1. 林柏青 / Lin Po-Ching　　ID：R120681373
2. 林盈達 / Lin Ying-Dar　　ID：P120502982
3. 劉明道 / Liu Ming-Dao　　ID：H122343904

住居所地址：（中文/英文）

1. 台北市東園街 66 巷 37 弄 58 號 4F
2. 新竹市大學路 1001 號交通大學資訊科學系
3. 桃園縣龍潭鄉三坑村 16 鄰 103 號

國　籍：（中文/英文）

1. 中華民國
2. 中華民國
3. 中華民國

肆、聲明事項：

☐ 本案係符合專利法第二十條第一項☐第一款但書或☐第二款但書規定之期間，其日期為： 年 月 日。

◎本案申請前已向下列國家（地區）申請專利 ☐ 主張國際優先權：

【格式請依：受理國家（地區）；申請日；申請案號數 順序註記】

1.

2.

3.

4.

5.

☐ 主張國內優先權(專利法第二十五條之一)：

【格式請依：申請日；申請案號數 順序註記】

1.

2.

☐ 主張專利法第二十六條微生物：

☐ 國內微生物 【格式請依：寄存機構；日期；號碼 順序註記】

☐ 國外微生物 【格式請依：寄存國名；機構；日期；號碼 順序註記】

☐ 熟習該項技術者易於獲得，不須寄存。

伍、中文發明摘要：

一種網頁內容過濾方法，通常可應用於客戶端的閘道器設備上，當一個網站的瀏覽工具發出網頁的存取要求後，藉由分析從網站傳回的網頁內容，決定該網頁之通行與否，其主要包括早期阻擋和及早期通過判斷標準，在過濾途中一但有足夠證據證明標的文件屬於某禁止類別，就可以及早期做出禁止的判斷，反之，當發現標的文件應屬於正常文件時，就會及早做出忽略的決定，因此，可有效提升網頁內容判斷的速度，減少網頁瀏覽者等待的時間，並提升網頁內容分析的容量。

陸、英文發明摘要：

柒、指定代表圖：

(一)本案指定代表圖為：第（ 1 ）圖。

(二)本代表圖之元件代表符號簡單說明:

捌、本案若有化學式時，請揭示最能顯示發明特徵的化學式：

玖、發明說明：

【發明所屬之技術領域】

一種網頁內容過濾方法，尤指一種關於網路安全管理之網頁內容過濾方法。

【先前技術】

隨著網站迅速的增加，各種資訊在網路上氾濫不已，因此，近年來市場上出現一種過濾網頁內容之軟硬體設備，一方面，可禁止未成年人利用網路尋訪色情或暴力的網站，另一方面，也利於管理員工利用上班時間進入其他網站進行工作以外的活動。為了效能考量，目前已有部分硬體產品問世，如 Allot 的 NetPure 5000 等等。不論在台灣或美國，這方面的產品在市場上扮演著不容忽視的地位。

以現有技術而言，為了判定網頁的內容是否為該被阻擋的網頁，有以下幾種方法：

1.建立一個資料庫，儲存預定阻擋之網頁的 Universal Resource Indicator (URI)，以實際網頁的要求中的 URI 與資料庫內容進行配對，若符合，則禁止存取該網頁。目前市面上絕大部分的產品皆採用此方式處理網頁過濾，然而，此方法需要維護一龐大的資料庫，而且，在網站不斷新增或更新之情況下，要維護一個最新的資料庫將付出極大成本。

2.採用關鍵字比對，檢查網頁的內容是否有某些特定的關鍵字作為篩選標準，然而，單憑網頁內容中是否出現某些關鍵字很容易造成判斷錯誤，例如用 sex 當關鍵字，可能會

把談論性別的網頁也一併過濾掉，造成篩選品質降低。

3.利用人工智慧的方式自動學習某類網頁的特徵，並對網頁做自動分類，這樣的方式雖然在判斷的精確度上沒有前述兩個方法會產生的問題，但是，此自動分類的過程往往需掃瞄整篇網頁內容，才能找出其特徵值並做出判斷，使得分類判斷的效能變差。

【發明內容】

爰是，本發明之主要目的，在於解決上述之問題，避免缺失的存在，本發明提供一可以自我訓練，具有高篩選精度及高過濾效率之網頁內容過濾方法。

為達到上述目的，本發明之網頁內容過濾方法，係在維持網頁內容判斷結果的正確性之前提下，考量 1.在不損害過濾的精確度的前提下，應儘早決定該阻擋的網頁；2.在不損害過濾的精確度的前提下，應儘早通過該通過的網頁；3.將傳統文件分類的方法應用到網頁內容過濾上；如是，藉由提早判斷的技術，可提升判斷速度，使效能較看完全部文件的方式提高到四倍以上，使得在客戶端的閘道器系統上進行網頁內容過濾的可行性大為提高。

【實施方式】

有關本創作之詳細說明及技術內容，現就配合圖式說明如下：

請審查委員參閱『第一圖』，係本發明之網頁內容過濾

流程圖，如圖所示：本創作為一種網頁內容過濾方法，係在網頁傳輸之通訊節點，由電腦執行而決定網頁之通行資訊，通常可應用於客戶端的閘道器設備上，當一個網站的瀏覽工具發出網頁的存取要求後，藉由電腦分析從網站傳回的網頁內容，決定該網頁之通行資訊，藉以判斷是否允許該網頁通過閘道器而出現客戶端的瀏覽器上，其步驟包括：

（1A）建立一網頁篩選標準，其至少包含有一關鍵字群組、一相對每一關鍵字之關聯概率表、一阻擋閥值（blocking threshold）及一通行閥值（bypassing threshold），並初始化一差異級數 SD；

（1B）取得網頁之資訊文本；

（1C）由資訊文本中取出一目標字；

（1D）判斷該目標字是否為關鍵字，若為是，則進行下一步，若為否，則跳至步驟（1H），繼續檢查文本；

（1E）依據關聯概率表重新計算目前各類別中分數最高及次高之差異級數 SD；

（1F）判斷差異級數 SD 是否超過阻擋閥值，若為是，則標記通行資訊為應阻擋之網頁，並結束步驟，若為否，則進行下一步；

（1G）判斷差異級數 SD 是否低於通行閥值，若為是，則標記通行資訊為應通過之網頁，並結束步驟，若為否，則進行下一步；

（1H）由資訊文本中取出下一目標字，判斷是否為資訊文本之資料末端，若為是，則標記通行資訊為應通過之網

頁，並結束步驟，若為否，則跳至步驟（1D）；

如是，藉由每次比較差異級數 SD 是否超過阻擋閥值，無須掃瞄全文且不損害過濾的精確度，對於關聯程度過高的網頁，即可早期阻擋預定篩選之網頁，而藉由每次比較差異級數 SD 是否低於通行閥值，同樣，無須掃瞄全文且不損害過濾的精確度，對於關聯程度過高的網頁，即可早期通過與篩選條件無關之網頁，透過此方法，可大幅提昇網頁內容過濾效率。

其中，網頁篩選標準另包含一使用者欲阻擋之文件類別群組 $C=\{c_1, c_2, \ldots, c_{|C|}\}$，如遊戲、色情、網路購物、金融投資等類別，而關聯概率表係依文件類別分別建立，且差異級數係由電腦執行下列步驟而完成，其步驟包括：

（2A）初始化一對應每一文件類別之評比積分群組 S；

（2B）透過關鍵字群組，比較目標字元與各文件類別之關聯程度，依關聯概率表計算每一評比積分 S；

$$\text{Score}(c_j \mid d_i) = \frac{P(c_j)\log(\prod_{k=1}^{di} P(w_{di,k} \mid c_j))}{P(d_i)}$$

（2C）選擇二最顯著之評比積分 S；

（2D）以此二評比初值之相對差距最為差異級數 SD。

請參閱『第二、三圖』，分別為色情網頁及一般網頁經本方法計算而獲得之評比積分，由實驗數據可以清楚看見，以色情網頁進行各文件類別之評比積分計算，其對於色情文

件類別之關聯程度明顯高於其他文件類別，而以一般網頁計算各文件類別之評比積分，則各文件類別之評比積分皆差異不多，因此，可依實驗數據制訂一阻擋閥值（blocking threshold），並透過差異級數 SD 之計算而辨認該網頁是否落入某一文件類別，進而可早期給定通行資訊，並阻擋網頁。

其中，通行閥值係為一隨關鍵字配對次數而改變之函數，此外，該步驟（1A）還包括（3A）建立一間距閥值，並初始化一間距數值 I，步驟（1E）還包括（3B）計算資訊文本中目標字元之平均間距作為間距數值，另步驟（1G）之判斷條件還包括（3C）間距數值是否大於間距閥值。

請參閱『第四圖』，係為通行閥值與一般網頁之差異級數比較表，由於差異級數隨關鍵字配對次數增加而增加，因此，通行閥值係為一隨關鍵字配對次數而改變之函數，如是可在不損害過濾的精確度的情況下，無須掃瞄全文，即可辨認那些極不可能屬於禁止文件類別之網頁，進而早期給定通行資訊，並使網頁通過。

然而，在計算前面幾個關鍵字之差異級數時，禁止網頁（如色情網頁）其差異級數可能低於通行閥值，本方法可能誤認為是一般網頁而給予通行，因此，需要其他機制來加以限制，請參閱『第五圖』，係為一般網頁與禁止網頁之關鍵字間距機率分佈圖，由圖可清楚發現，關鍵字在禁止網頁中相隔間距較小的機率高於一般網頁，因此，步驟（1G）除判斷差異級數 SD 是否低於通行閥值，另需判斷其間距數值是否大於間距閥值，若同時滿足前述二條件，即可給定通行資

訊為可通行之網頁，如是，可確保其篩選精卻度。

此外，關鍵字群組及關聯概率表係可藉由電腦執行下列步驟完成，如一般貝氏學習方法（Bayesian learning porcedure），其包括有：

（4A）提供一依文件類別群組 C 分類之調校文本群組 $D=\{d_1, d_2, \ldots, d_{|D|}\}$，每一調校文本 d_i 係由一字元群組 $V=\{w_1, w_2, \ldots, w_{|V|}\}$ 所構成，且每一文件類別 c_j 至少包含一調校文本 d_i；

（4B）透過調校文本，依文件類別分別計算每一字元 wt 在該文件類別 c_j 中出現之字元概率 $P(w_t|c_j)$；

（4C）在所有文件類別中依字元概率之顯著程度選出總字數為預定數額的一字元集為關鍵字群組；

（4D）以關鍵字及相對應之字元概率制訂關聯概率表。

根據實驗顯示，請參閱『第六圖』，係為關鍵字數量與篩選成效之比較表，如圖所示，篩選成功率隨關鍵字數量增加而提升，但是，關鍵字數量增加又造成電腦計算量暴增，為權衡篩選成效與計算量，使用者可自行決定預定數額，可採用 500 個關鍵字，同時降低計算量並兼顧一定篩選成效。

使用時，本方法可整合或搭配客戶的防火牆系統設備，透過防火牆的設定強迫員工存取網路一定得經由本系統做為代理伺服器，其間，可以依據客戶的政策提供由客戶打算阻擋的類別來訓練本系統，即是執行（4A）至（4D）步驟，制訂關鍵字群組及關聯概率表，爾後，員工存取的網頁內容會先經過本系統，根據網頁內容比對關鍵字並給定一差異級

數，藉由阻擋閥值（blocking threshold）及通行閥值（bypassing threshold）之條件比較，可在不損失篩選精確度的情況下，無須掃瞄全文即可早期提供通行資訊，並決定是否該讓員工存取。

即時性的內容分析對網頁內容過濾來說是一項非常重要的技巧。但其同時也有準確度較低及處理時間過長的問題。藉由上述方法，本發明提出兩種加速其分類過程的技術，分別為早期阻擋和及早期通過，在過濾途中一但有足夠證據證明標的文件屬於某禁止類別，就可以及早做出禁止的判斷。反之，當發現標的文件應屬於正常文件時，就會及早做出忽略的決定，因此，可提升網頁內容判斷的速度，減少網頁瀏覽者等待的時間，並使實作這套方法的網路閘道器裝置可以分析更多的網頁內容。

實驗結果顯示，在使用 Pentinum III 700 MHz CPU 及 NetBSD 1.6 的作業系統環境下，本方法與原始的 Bayesian 分類演算法相比較，本方法之傳輸效能可以提升四倍以上，同時，以 F1 評估方法顯示，本方法仍可維持相當好的判斷準確度，在禁止流量中可達 92%，在正常流量中可達 96%。

惟以上所述者，僅為本發明之較佳實施例而已，當不能以此限定本發明實施之範圍；故，凡依本發明申請專利範圍及發明說明書內容所作之簡單的等效變化與修飾，皆應仍屬本發明專利涵蓋之範圍內。

【圖式簡單說明】

第一圖，係為本發明之網頁內容過濾流程圖。

第二圖，係為色情網頁經本方法計算而獲得之評比積分。

第三圖，係為一般網頁經本方法計算而獲得之評比積分。

第四圖，係為通行閥值與一般網頁之差異級數比較表。

第五圖，係為一般網頁與禁止網頁之關鍵字間距機率分佈圖。

第六圖，係為關鍵字數量與篩選成效之比較表。

拾、申請專利範圍：

1. 一種網頁內容過濾方法，係在網頁傳輸之通訊節點，由電腦執行下列步驟而決定網頁之通行資訊，其步驟包括：

（1A）建立一網頁篩選標準，其至少包含有一關鍵字群組、一相對每一關鍵字之關聯概率表、一阻擋閥值及一通行閥值，並初始化一差異級數；

（1B）取得網頁之資訊文本；

（1C）由資訊文本中取出一目標字元；

（1D）判斷該目標字元是否為關鍵字，若為是，則進行下一步，若為否，則跳至步驟（1H），繼續檢查文本；

（1E）依據關聯概率表計算目標字元相對應之差異級數；

（1F）判斷差異級數是否超過阻擋閥值，若為是，則標記通行資訊為應阻擋之網頁，並結束步驟，若為否，則進行下一步；

（1G）判斷差異級數是否低於通行閥值，若為是，則標記通行資訊為應通過之網頁，並結束步驟，若為否，則進行下一步；

（1H）由資訊文本中取出下一目標字元，判斷是否為資訊文本之資料末端，若為是，則標記通行資訊為應通過之網頁，並結束步驟，若為否，則跳至步驟（1D）。

2. 如申請專利範圍第1項所述之網頁內容過濾方法，其中，通行閥值係為一隨關鍵字配對次數而改變之函數，且網頁篩選標準另包含一使用者欲阻擋之文件類別群組 $C=\{c_1, c_2, \ldots, c_{|C|}\}$，而關聯概率表係依文件類別分別建立，且

差異級數係由電腦執行下列步驟而完成，其步驟包括：

（2A）初始化一對應每一文件類別之評比積分群組 S；

（2B）透過關鍵字群組，比較目標字元與各文件類別之關聯程度，依關聯概率表計算每一評比積分；

（2C）選擇二最顯著之評比積分；

（2D）以此二評比初值之相對差距為差異級數。

3. 如申請專利範圍第 1 項所述之網頁內容過濾方法，其中，該步驟（1A）還包括（3A）建立一間距閥值，並初始化一間距數值，步驟（1E）還包括（3B）計算資訊文本中目標字元之平均間距作為間距數值，另步驟（1G）之判斷條件還包括（3C）間距數值是否大於間距閥值。

4. 如申請專利範圍第 1 項所述之網頁內容過濾方法，其中，關鍵字群組及關聯概率表係藉由電腦執行下列步驟完成，其包括有：

（4A)提供一依文件類別群組 C 分類之調校文本群組 $D=\{d_1, d_2, \ldots, d_{|D|}\}$，每一調校文本 d_i 係由一字元群組 $V=\{w_1, w_2, \ldots, w_{|V|}\}$所構成，且每一文件類別 c_j 至少包含一調校文本 d_i；

（4B）透過調校文本，依文件類別分別計算每一字元 wt 在該文件類別 c_j 中出現之字元概率 $P(w_t|c_j)$；

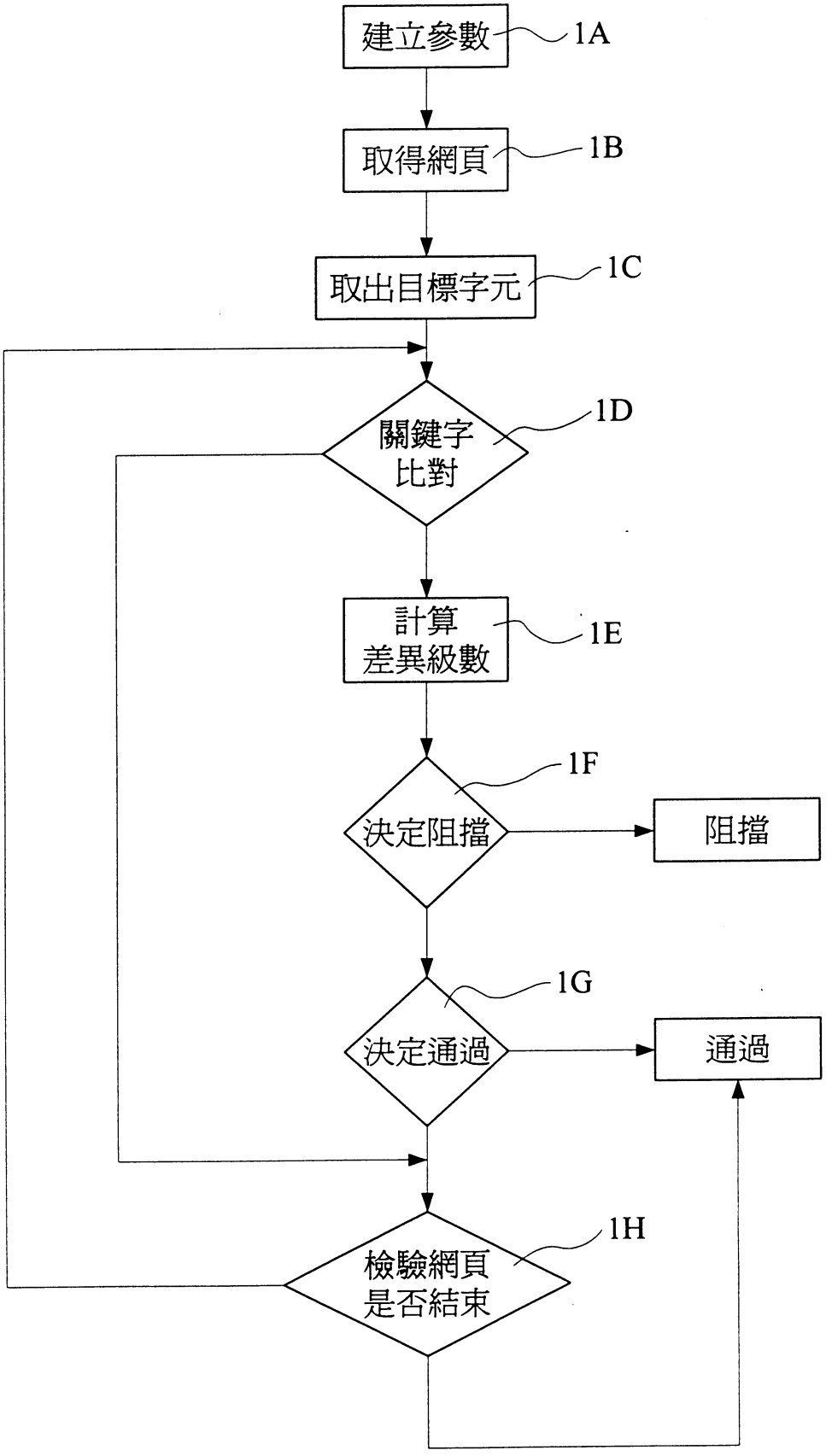
（4C）在所有文件類別中依字元概率之顯著程度選出總字數為預定數額的一字元集為關鍵字群組；

（4D）以關鍵字及相對應之字元概率制訂關聯概率表。

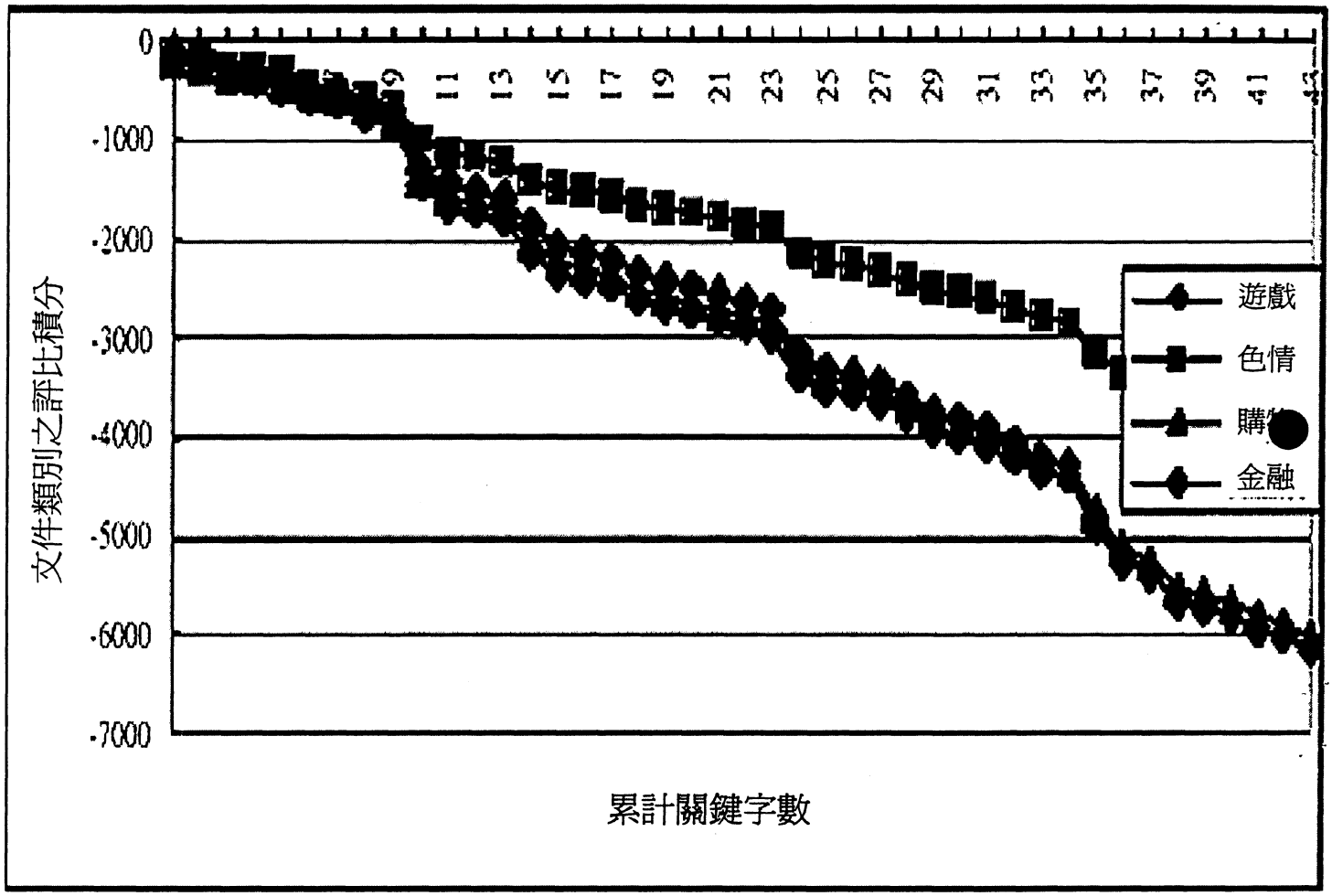
5. 如申請專利範圍第 1 項所述之網頁內容過濾方法，其中，

該通訊節點係為一區域網路之防火牆設備。

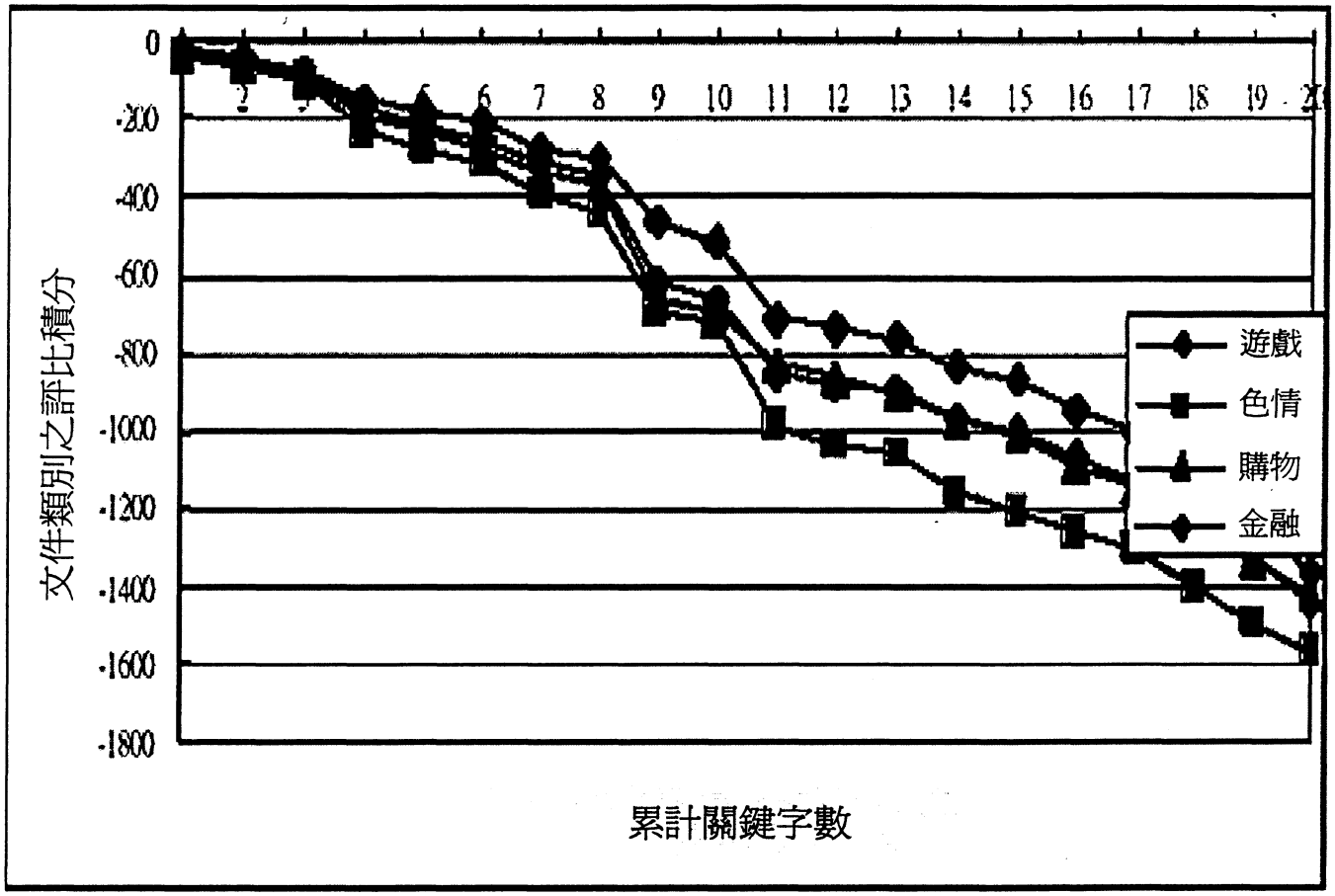
6. 如申請專利範圍第1項所述之網頁內容過濾方法，其中，該通訊節點係為一個人電腦之防火牆設備。


建立參數
1A
取得網頁
1B
取出目標字元
1C
關鍵字
比對
1D
計算
差異級數
1E
決定阻擋
1F
阻擋
決定通過
1G
通過
檢驗網頁
是否結束
1H

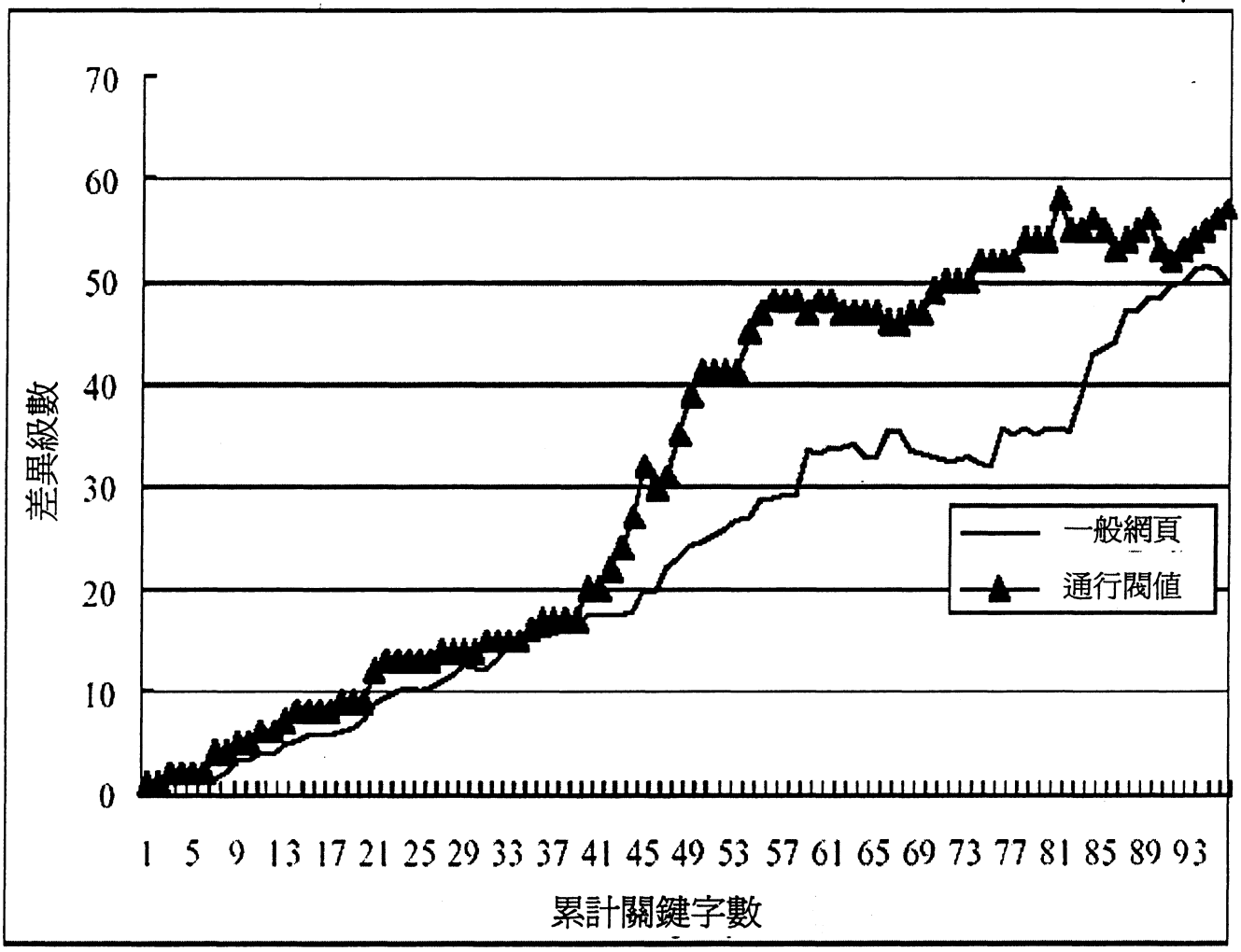
第一圖


0
-1000
-2000
-3000
-4000
-5000
-6000
-7000
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
文件類別之評比積分
累計關鍵字數
遊戲
色情
金融

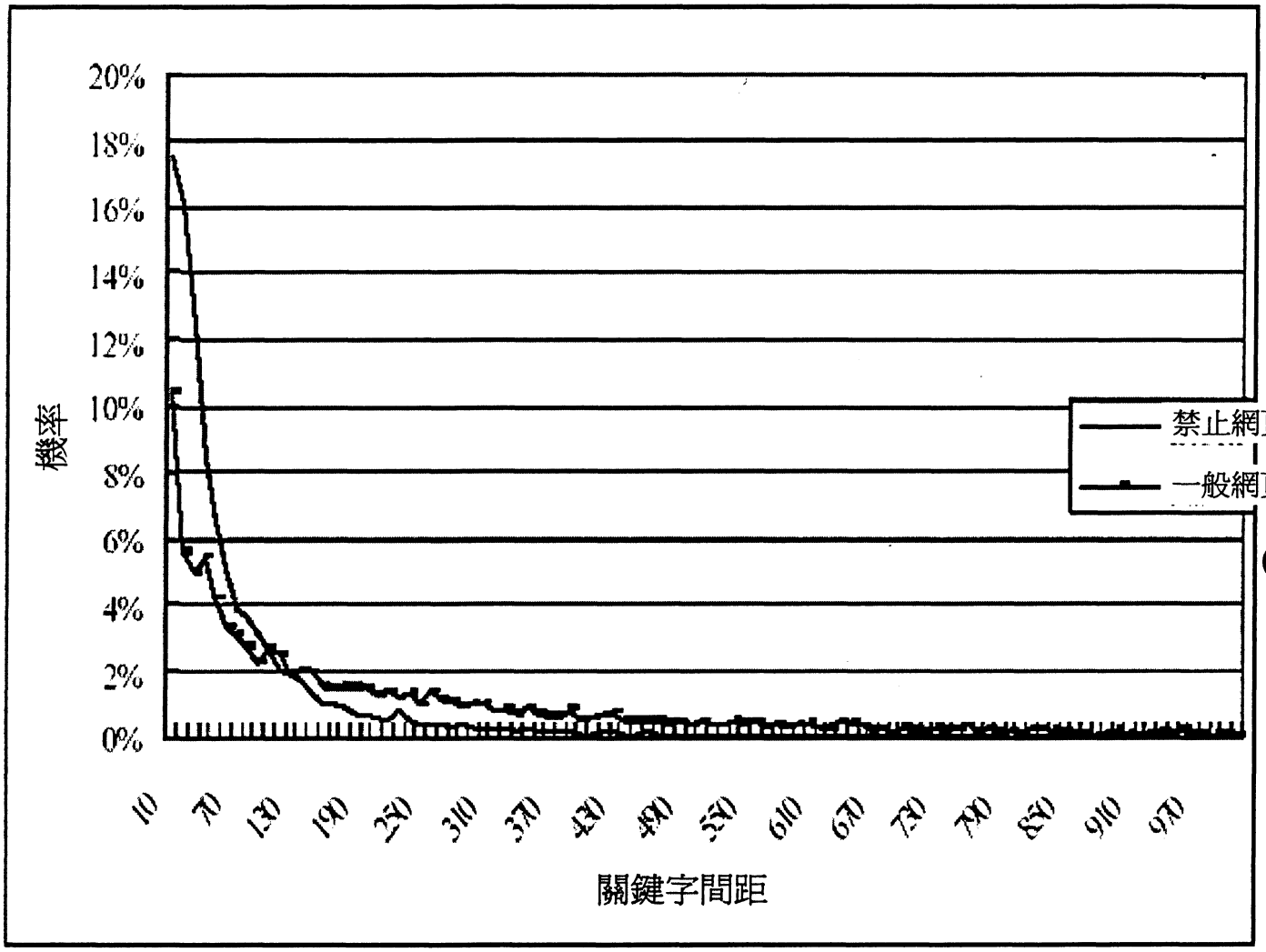
第二圖


0
-200
-400
-600
-800
-1000
-1200
-1400
-1600
-1800
2
6
7
8
9
10
11
12
13
14
15
16
17
18
19
文件類別之評比積分
累計關鍵字數
遊戲
色情
購物
金融

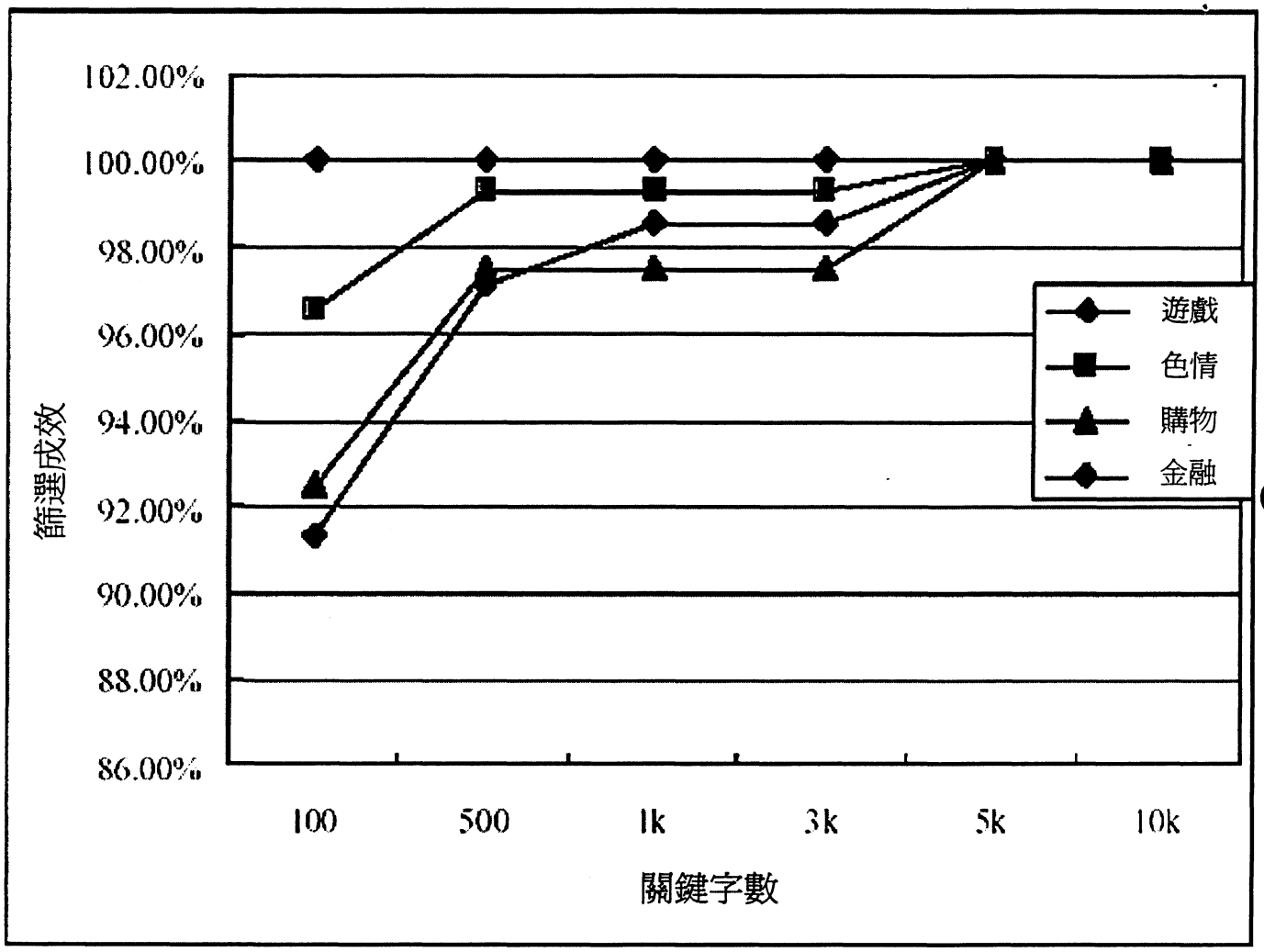
第三圖


70
60
50
40
30
20
10
0
差異級數
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93
累計關鍵字數
一般網頁
通行閥值

第四圖


20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
機率
禁止網頁
一般網頁
10
70
130
190
250
310
370
430
490
550
610
670
730
790
850
910
970
關鍵字間距

第五圖


102.00%
100.00%
98.00%
96.00%
94.00%
92.00%
90.00%
88.00%
86.00%
篩選成效
100
500
1k
3k
5k
10k
關鍵字數
遊戲
色情
購物
金融

第六圖