I355168

97年5月2日修正替換頁

公告本

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：96146824

※ 申請日期：96.12.7　　※IPC 分類：H04L 12/24 (2006.01)
H04L 12/26 (2006.01)
H04L 12/56 (2006.01)

一、發明名稱：（中文/英文）

在網路流量中分類所屬應用程式之方法

APPLICATION CLASSIFICATION METHOD IN NETWORK TRAFFIC

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

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, Taiwan, R.O.C.

國　籍：（中文/英文）中華民國/TW

三、發明人：（共 5 人）

姓　名：（中文/英文）

林盈達/ LIN, YING-DAR

彭偉豪/ PENG, WEI-HAO

賴源正/ LAI, YUAN-CHENG

呂俊男/ LU, CHUN-NAN

陳一瑋/ CHEN, I-WEI

國　籍：（中文/英文）（皆同）中華民國/TW

97年1月24日修正補充頁

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：2007 年 6 月。

☐ 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：

本發明提出了一種網路流量分類方法，其利用網路傳輸層行為特徵以及應用程式連線封包大小分佈與結合埠關聯特性，做為辨認流量中所屬應用程式之依據。利用應用程式在傳輸層中行為計算出的向量值與已知的代表特徵點比對辨認，並且利用埠關聯的特性一併將與應用程式相關的連線辨識出來。

六、英文發明摘要：

An application classification method in network traffic is disclosed. The invention uses Packet Size Distribution and Ports Association to achieve the classification of application. Every succeeded connection would be transformed into one vector in the multi-dimensional coordinate spaces and classified into some specified applications or other unknown ones. Besides the Euclidean distances of every connection between all individual centers and the representatives of the applications will also be computed. Once a connection is identified and classified into some certain session, ports association algorithm is used to associate and accelerate other connection in the same session.

七、指定代表圖：

(一)、本案代表圖為：第 1 圖

(二)、本案代表圖之元件代表符號簡單說明:

100	第一階段
200	第二階段
S110-S140	訓練過程之步驟
S205-S240	分類過程之步驟

八、本案若有化學式時，請揭示最能顯示發明特徵的化學式：

無

九、發明說明：

【發明所屬之技術領域】

本發明係有關一種網路流量分類方法，特別是在網路流量中分類所屬應用程式之方法。

【先前技術】

在網路流量的分析中，一種使用非封包內容判斷流量種類的方法，其利用封包的內部到來時間與大小的變化加上埠號的判定來決定該流量屬於何種類型，但是此方法僅能判定流量是屬於即時訊息（Instant Messaging, IM）的傳輸、某種應用程式的命令資料傳輸或是某種應用程式的資料傳輸，並無法辨認出流量屬於何種應用程式。

傳統的網路偵測技術皆依靠應用程式已知埠號與封包內容特徵值比對方式，這樣的方法已知有兩個缺點：(1)無法偵測動態決定埠號使用之應用程式、(2)封包內容如果被應用程式加密就無法透過內容特徵值比對辨認。

另外一種方式是在點對點流量模式（P2P flow pattern）中，先檢查兩點之間是否同時存在 TCP 及 UDP 連線，接下來去除掉一些已知（well-known）應用的連線(ex. HTTP, SMTP, FTP)，若兩點之間的連線數目等同於埠號對（port pair）的數目的話，即將這些連線當作是點對點 P2P 流量。此方法的限制在於現行的 P2P 流量很多都是跑在已知埠號（well-known port）。利用已知埠消去法並無法保證對於 P2P 應用程式的偵測正確率。

再者，稱作 BlinC 的方法，透過三個層面(Social, Functional, Application)來分析流量。Social 層面將每一來源（source）與哪些目的地（destination）有溝通標示出來；若有某一群的來源（source）同時與很多且同樣的目地（destination）溝通的情況，很有可能是病毒

攻擊流量（(ex. Blaster)；若目的地（destination）的數目是正常的，很有可能是同時有一群人在瀏覽同一個網站或是串流（streaming）的應用。Functional 層面則是決定主機(host)所扮演的角色偏向 Server、Client 或是 P2P；至於 Application 層面透過來源 IP（source IP）、目的地 IP(destination IP)、來源埠號(source Port)、目的地埠號(destination Port)這些變數值組(4-tuple)再進一步來分辨流量是屬於哪一種應用程式。此方法雖然有極高的準確度，但是因為需要大量的傳輸層資訊做為判讀之用，所以相當耗時。

在美國專利 US Patent. 6,157,955 中提出一個可以針對網路介面加上分類引擎機制的架構來對網路上的封包做分析與分類。在分類引擎的部分包含了兩個主體：封包頭端資訊解析(Packet header parsing)與雜湊表查詢機制（hash table lookups），而引擎的過濾機制，則是由主機端定義來決定什麼樣的應用程式封包可以通過。此專利提供了一個彈性的機制可以任意增加新的過濾方針，並且可以動態決定，儘量節省所需偵測的封包內容資訊。此篇專利類似於本申請案的架構（利用一套分類機制分類網路中的流量），但卻沒有進一步地對使用加密協定的應用程式做偵測機制。

在美國專利 US Patent. 6,597,660 中提出一套可以分析、預測及分辨網路即時流量的架構，包含了一個可以儲存及處理封包時間資訊的裝置，利用在不同的時間點、不同的時間範圍內，同時累計接收到的封包時間資訊，再利用統計出的封包時間資訊分類封包。此篇專利與本申請案均採用統計計算後所得到的資訊作為封包判斷的依據。不過，此篇專利使用的是封包到達時間資訊，與本申請案不同；另外，此篇專利也不能進一步偵測經過加密後的封包。

在美國專利 US Patent. 6,754,662 中採用了一套發送引擎及一組雜湊表結構來分類封包，雜湊表內儲存的是一組判斷識別名，發送引擎則會在收到封包後，根據收到的封包資訊，計算出封包的雜湊值，

再嘗試以雜湊值當作索引，到儲存的雜湊表中去尋找。雜湊表內的項目則會根據網路流量的統計，包含存取頻率、最近存取時間、應用程式種類，以及流量長度等作判斷，決定存在雜湊表內的時間長短。

在美國專利US. Patent. 6,839,751 中採用了一組封包擷取裝置及一組資料庫；資料庫主要是用來儲存已經處理過的對話流量資訊。當封包擷取裝置接受封包後，會到資料庫中查詢是否已經處理過；如果處理過，則會根據包含的統計資訊，包括該連線擁有的封包總個數、封包到達時間、及此次封包與上次收到的封包到達時間差等更新資料庫。如果沒有處理過，則在資料庫中新增項目。

傳統的偵測分析技術針對應用程式使用已知埠（well-known port）方式做判定，但是現行許多存在於網路上的有害物質，由於採用了動態埠（dynamic port）的技術，皆無法使用該方法來辨認。

現今廣泛使用的“封包內容特徵值”比對方法，因為越來越多P2P/IM 軟體使用封包加密技術而無法利用此方法對其封包內容判別偵測，造成管理上的漏洞。

又，某些惡意軟體利用偽裝封包內容的方式意圖躲避內容特徵值比對的偵測，傳統方法有可能會因此產生誤擋或是漏擋之問題。但是現今的封包內容偵測方式，有侵害個人隱私的問題。

目前的傳輸層特徵比對方式，皆有需要收集足夠傳輸層資訊方能獲得正確判斷能力之缺陷，並且判斷時間過長，無法適用在需要快速決定網路流量管理政策的閘道器或是防火牆之上。

【發明內容】

為了解決上述問題，本發明目的之一係提出可以用來偵測被加密或是刻意隱藏通訊協定之應用程式，以提供網路管理者流量處理上足夠的資訊。

本發明另一目的係提出一種在網路流量中分類所屬應用程式之方法，其提出了一個利用傳輸層行為特徵，計算應用程式連線封包大小分佈與結合埠關聯特性之方法，來做為辨認流量中應用程式之依據。利用應用程式在傳輸層中行為計算出的特徵值（向量值）與已知的代表特徵值比對辨認，並且利用埠關聯的特性一併將與應用程式相關的連線辨識出來。

為了達到上述目的，本發明一實施例之在網路流量中分類所屬應用程式之方法，包括：計算一指定應用程式之複數個代表特徵值；將複數個實際網路封包流量拆解成一第二組連線；搜尋第二組連線是否存在於一埠關連表格中；以及，若是沒有存在於埠關連表格中，則計算第二組連線之特徵值，並與些代表特徵值作比較，選擇最接近之代表特徵值以歸屬為指定應用程式。

【實施方式】

第 1 圖所示為本發明一實施例之網路流量分類方法之執行步驟，包括第一階段 100 之訓練過程與第二階段 200 之分類過程。

在第一階段 100 之訓練過程中，分析已錄製的流量並根據應用程式的不同作分類，以求得各分類的代表特徵，其包括：步驟 110 流量收集（Traffic Collection），流程一開始是經由流量收集，先收集想要比對的應用程式流量，得到足夠的封包個數後（至少需要超過 400 個封包個數）；步驟 120 計算各連線特徵（Connection Characterizing），將流量拆解成多個連線（connection）；步驟 130 計算應用程式代表特徵值，以各連線為處理單位再分別計算其代表特徵值，包含有支配值（Dominating Size, DS）、支配值比例（Dominating Size Proportion, DSP），及變動週期（Change Cycle, CC）；以及最後的步驟 140 應用程式代表特徵值之集合，得到一個應用程式代表特徵值之集合，並儲

存經由上述步驟所計算出的應用程式代表特徵值，以作為第二階段200分類過程之線上模式比對流量的基準。

根據上述各步驟的動作，在步驟110流量收集（Traffic Collection）中，採用應用程式流量收集技術，利用網路流量過濾器的概念，執行想要比對的應用程式，限定應用程式及其使用的埠號，使得只有所需要的應用程式封包才能通過網路介面，並且在網路流量出入口端利用流量錄製技術將所需的流量錄製下來做為分析之用。

在步驟120計算各連線特徵（Connection Characterizing）中，依據來源IP、來源埠號、目的IP及目的埠號，將錄製到的流量分類，拆解成多條連線。以各連線為處理單位，分別計算各連線的特徵值，亦為向量值（vector），包含有支配值（DS）、支配值比例（DSP），及變動週期（CC）。其中支配值與支配值比例各是指連線中佔有較大比例之各個封包大小及相對應的佔有比例數，變動週期則是當某一連線中所含的封包大小有劇烈變化時，用來作為輔助辨識的依據。

在步驟130計算應用程式代表特徵值中，有了各連線的特徵值後，再從處在相同交談（session）的各連線推出可代表此類的代表特徵值。本實施例中是對各連線的特徵值平均計算，將計算所得的平均值作為某一類應用程式的代表特徵值。

接著在第二階段200之分類過程中，利用第一階段100之訓練過程得到的各應用程式代表特徵值，作為與網路中真實流量比對的基準，藉著與各代表特徵值之間的差距來推論擷取到的封包屬於哪種應用程式。包括：步驟205接入網路中真實流量；步驟210流量拆解，將流量拆解成多個連線（connection），並依照第一階段之步驟120計算各連線特徵；步驟220建立埠關連表格（Port Association Table, PAT），以各連線為處理單位，內以<SrcIP，SrcPort>、<DstIP，DstPort>作為索引，至埠關連表格（PAT）去搜尋是否已有相關資訊存在；如果沒有，則進入步驟230封包辨識，先分別計算各連線之特徵值，再

與第一階段步驟130中獲得之各應用程式代表特徵值做歐幾里得距離（Euclidean Distance）比較，選擇差距最小者之應用程式特徵值做為該連線之歸屬應用程式；如果該連線已有資訊存在埠關連表格(PAT)中，則可依據埠關連表格（PAT）中之記錄，直接判定該連線之歸屬應用程式。最後，如果發現該連線所計算出之特徵值並不存在於PAT中，且與第一階段100所得應用程式特徵值集合之差距也過大而無法判定歸屬之應用程式種類時，可將該連線判定為『未知應用程式』；最後的步驟240判定應用程式，等待被辨識的封包就可以被判定成『已知的某類應用程式』，或是『未知的應用程式』。

請參閱第2圖所示為本發明一實施例之網路流量分類方法在第二階段200之分類過程示意圖，其步驟與第1圖之步驟相同。分類過程包括：步驟205接入網路中真實的封包流量；步驟210將封包流量拆解成多個連線；步驟220比對連線若已有資訊存在埠關連表格，則進入最後步驟240的應用程式判定為程式A或程式B，若比對連線沒有資訊存在埠關連表格，則進入步驟230的封包辨識以判定為程式A、程式B或未知程式。

根據上述各步驟的動作，在步驟210流量拆解中，接入網路上真實流量後，依據來源IP（SrcIP）、來源埠號（SrcPort）、目的IP（DstIP），及目的埠號（DstPort），將想要分析的流量分類，拆解成多條連線。

在步驟220埠關連表格中，先尋找該連線的<SrcIP，SrcPort>、<DstIP，DstPort>是否出現在埠關連表格(PAT)中，埠關連表格(PAT)中儲存的是已經辨識出的連線及所屬之交談（session）資訊；以<SrcIP，DstIP，SrcPort，DstPort>來代表一條被辨認出的連線，依照下列步驟操作：

1.記錄使用該SrcIP與DstIP的主機(host)有使用辨認出的應用程式。

2.將其SrcPort、DstPort記錄於埠關連表格（PAT）中。

3.若有某條連線符合<SrcPort，SrcPort+1>或是<DstPort，DstPort+1>的情況，則認定該連線亦屬於該交談（session）。

在步驟 230 封包辨識中，依照 210 處所拆解的各連線，分別計算其特徵值，再與第一階段之應用程式代表特徵值之集合得到的應用程式代表特徵值作歐幾里得距離（Euclidean Distance）運算；若是連線封包大小分佈與某個應用程式代表特徵值相似，則之間的歐幾里得距離一定會比較接近，故可用來判斷一個連線與哪種應用程式最為類似。同時，我們也會對辨識出的各連線作交談（session）關聯性分析，將屬於相同交談（session）的各連線組合在一起，以期得到較全面性的資訊。

以上所述是針對某一應用程式比對的操作流程敘述，如果有多種應用程式需要比對，本發明僅需要針對不同的應用程式多次操作本運作流程即可。

綜合上述，本發明為一在網路流量中分類所屬應用程式之方法，應用程式連線封包大小分佈與結合埠關聯特性之方法來做為辨認流量中應用程式之依據。利用應用程式在傳輸層中行為計算出的特徵值（向量值）與已知的代表特徵點比對辨認，並且利用埠關聯的特性一併將與應用程式相關的連線辨識出來。本發明解決了無法利用封包內容辨認的問題與動態埠號之使用而無法辨認之問題，並且提供一個可以用來做為線上閘道器使用之辨認機制。

以上所述之實施例僅係為說明本發明之技術思想及特點，其目的在使熟習此項技藝之人士能夠瞭解本發明之內容並據以實施，當不能以之限定本發明之專利範圍，即大凡依本發明所揭示之精神所作之均等變化或修飾，仍應涵蓋在本發明之專利範圍內。

【圖式簡單說明】

第 1 圖所示為本發明一實施例之網路流量分類方法之執行步驟。

第 2 圖所示為本發明一實施例之網路流量分類方法之分類過程示意圖。

【主要元件符號說明】

100	第一階段
200	第二階段
S110-S140	訓練過程之步驟
S205-S240	分類過程之步驟

十、申請專利範圍：

1. 一種在網路流量中分類所屬應用程式之方法，包括：

計算一指定應用程式之複數個代表特徵值；

將複數個實際網路封包流量拆解成一第二組連線；

搜尋該第二組連線是否存在於一埠關連表格中；以及

若是沒有存在於該埠關連表格中，則計算該第二組連線之特徵值，並與該些代表特徵值作比較，選擇最接近之該些代表特徵值以歸屬為該指定應用程式。

2. 如請求項 1 所述之在網路流量中分類所屬應用程式之方法，其中計算該些代表特徵值更包括：

收集該指定應用程式之複數個封包流量；

將該些封包流量拆解成一第一組連線；以及

以該第一組連線為處理單位再分別計算數個特徵值，以得到該指定應用程式的代表特徵值集合。

3. 如請求項 2 所述之在網路流量中分類所屬應用程式之方法，其中該些代表特徵值包含一支配值、一支配值比例及一變動週期。

4. 如請求項 3 所述之在網路流量中分類所屬應用程式之方法，其中計算該些代表特徵值係對該第一組連線的特徵值作平均計算作為該指定應用程式的代表特徵值。

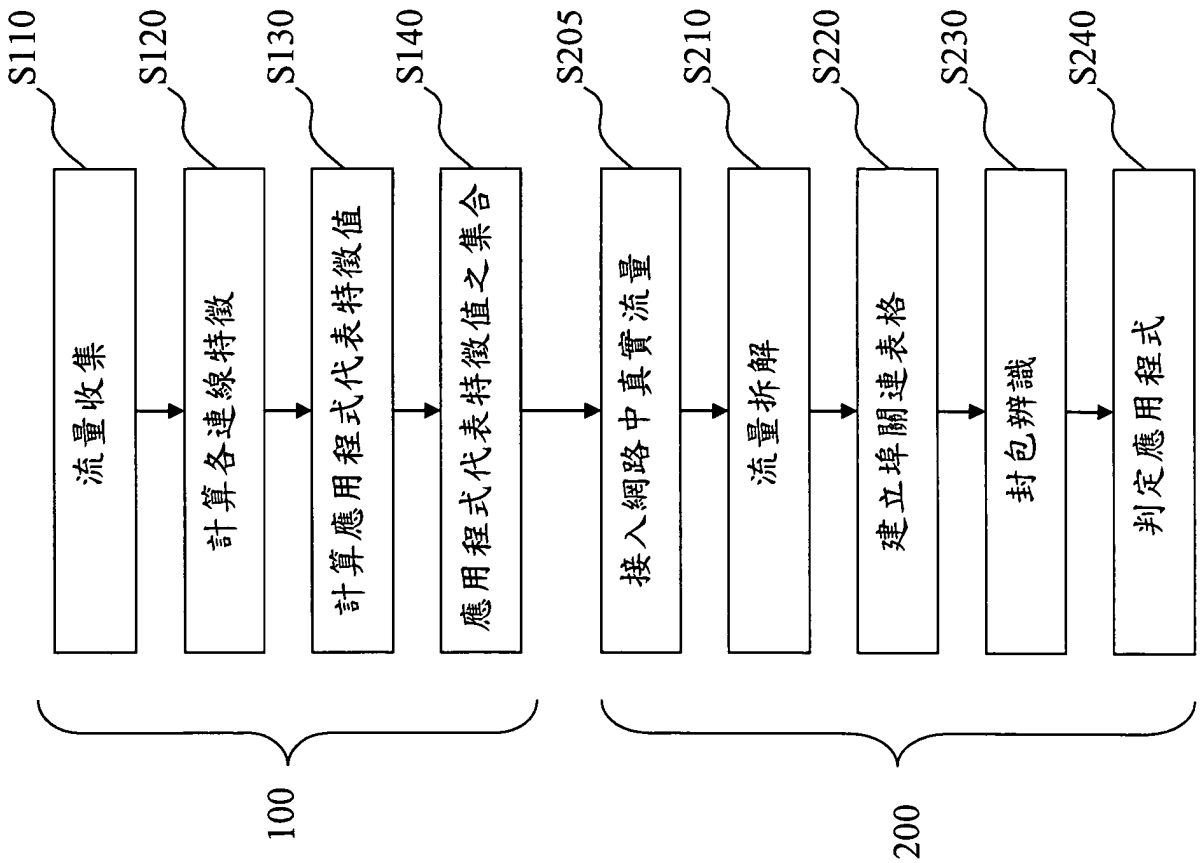
5. 如請求項 1 所述之在網路流量中分類所屬應用程式之方法，其中該埠關連表格位於網路傳輸層之處理程式中。

6. 如請求項 1 所述之在網路流量中分類所屬應用程式之方法，其中計算該第二組連線之特徵值與該些代表特徵值作比較是使用歐幾里得距離比較法。

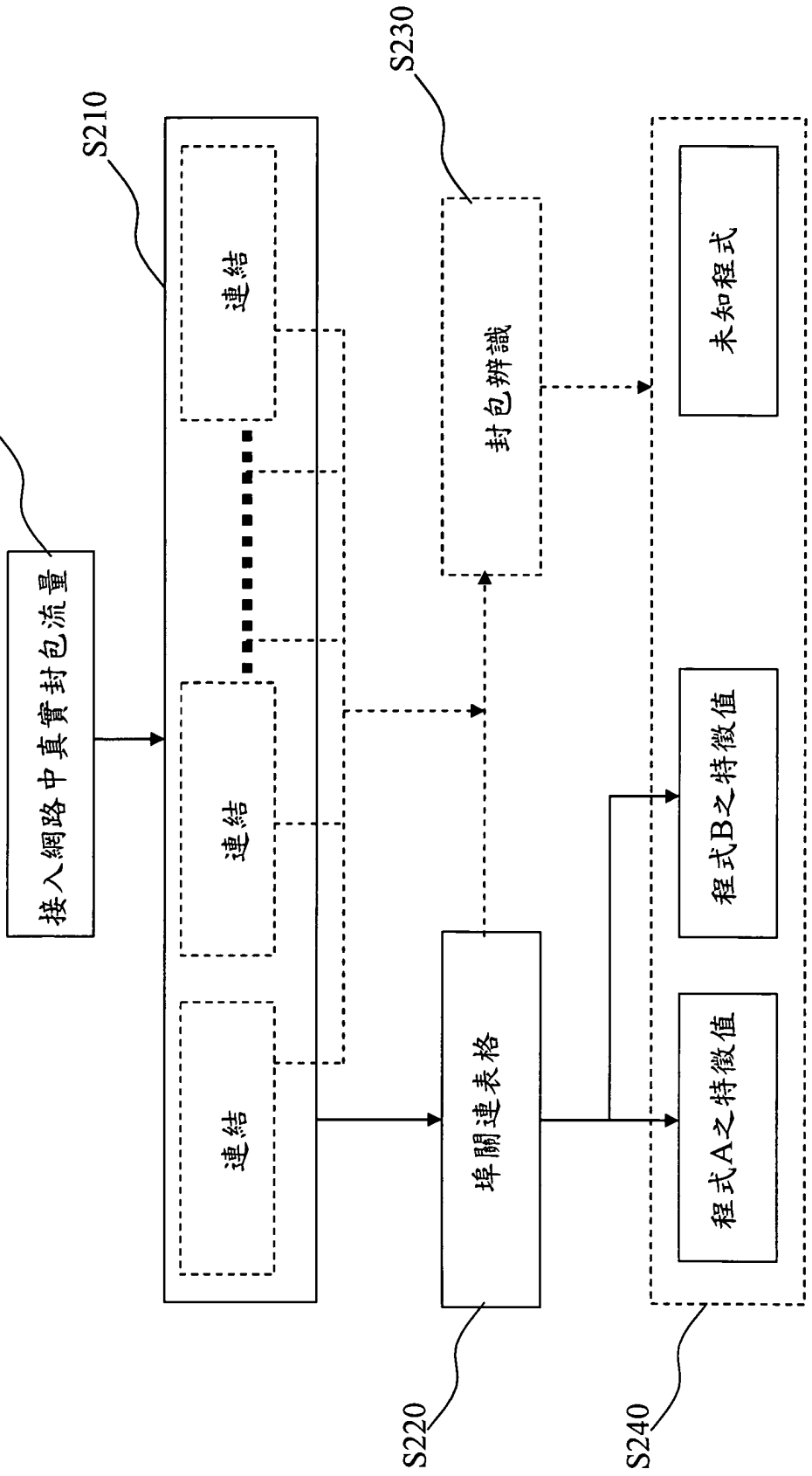
7. 如請求項 1 所述之在網路流量中分類所屬應用程式之方法，更包括若該第二組連線存在於該埠關連表格中，則判定為一已知的應用程式。

8. 如請求項 1 所述之在網路流量中分類所屬應用程式之方法，更包括若該第二組連線之特徵值無最接近之該些代表特徵值，則判定為一未知的應用程式。

9. 如請求項 1 所述之在網路流量中分類所屬應用程式之方法，其中該埠關連表格包含有來源 IP 欄位、來源埠號欄位、目的 IP 欄位、以及目的埠號欄位。


S110
流量收集
S120
計算各連線特徵
S130
計算應用程式代表特徵值
S140
應用程式代表特徵值之集合
100
S205
接入網路中真實流量
S210
流量拆解
S220
建立埠關連表格
S230
封包辨識
S240
判定應用程式
200

第1圖


S205
接入網路中真實封包流量
S210
連結
連結
連結
S220
埠關連表格
S230
封包辨識
S240
程式A之特徵值
程式B之特徵值
未知程式

第2圖