201515410

(19)中華民國智慧財產局

(12)發明說明書公開本 (11)公開編號：TW 201515410 A

(43)公開日： 中華民國 104 (2015) 年 04 月 16 日

(21)申請案號：102136709 (22)申請日： 中華民國 102 (2013) 年 10 月 11 日

(51)Int. Cl.： *H04L12/803 (2013.01)* *H04L12/24 (2006.01)*

(71)申請人：國立交通大學 (中華民國) NATIONAL CHIAO TUNG UNIVERSITY (TW)
新竹市大學路 1001 號

(72)發明人：呂俊男 LU, CHUN NAN (TW) ；黃俊穎 HUANG, CHUN YING (TW) ；林盈達 LIN, YING DAR (TW) ；賴源正 LAI, YUAN CHENG (TW)

(74)代理人：陳昭誠

申請實體審查：有 申請專利範圍項數：10 項 圖式數：6 共 29 頁

(54)名稱

基於訊息長度序列之網路流量辨識系統及其方法

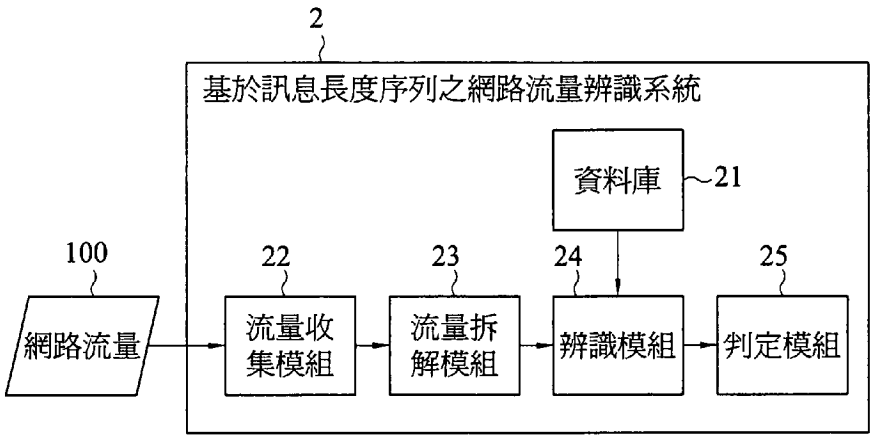
A TRAFFIC CLASSIFICATION SYSTEM BASED ON MESSAGE SIZE SEQUENCE AND METHOD THEREOF

(57)摘要

一種基於訊息長度序列之網路流量辨識系統，包括：預存對應各種網路應用程式之長度共用子序列集合的資料庫、用於收集網路流量的流量收集模組、將網路流量拆解成多條連線並產生對應每一連線之長度特徵序列的流量拆解模組、比對長度特徵序列與資料庫中各種網路應用程式之長度共用子序列集合的辨識模組、以及判定所偵測之連線為已知網路應用程式或未知網路應用程式的判定模組。透過本發明之萃取連線行為特徵作網路流量辨識，可解決現有加密通訊或刻意隱藏封包內容之應用程式的偵測問題。

Proposed is a traffic classification system based on message size sequence, comprising: a collection for common message size subsequence corresponding to a variety of network applications, a traffic collection module for collecting network traffic, a traffic decomposition module for decomposing the network traffic into multiple flows and producing a message size sequence corresponding to each flow, a flow classification module for comparing the message size sequence with the common message size subsequence collections of each network application in the database, and an arbitration module for determining the flow to be a known or an unknown network application. The present invention uses extraction of flow behavior characteristics for network traffic identification to resolve existing application packet detection issues about encrypted communication or deliberately hiding contents.

TW 201515410 A



第2圖

2・・・基於訊息長度序列之網路流量辨識系統

21・・・資料庫

22・・・流量收集模組

23・・・流量拆解模組

24・・・辨識模組

25・・・判定模組

100・・・網路流量

201515410

發明摘要

※申請案號：102136709

※申請日：102.10.11　　※IPC分類：

H04L 12/803 (2013.01)

H04L 12/24 (2006.01)

【發明名稱】（中文/英文）

基於訊息長度序列之網路流量辨識系統及其方法

A TRAFFIC CLASSIFICATION SYSTEM BASED ON MESSAGE SIZE SEQUENCE AND METHOD THEREOF

【中文】

一種基於訊息長度序列之網路流量辨識系統，包括：預存對應各種網路應用程式之長度共用子序列集合的資料庫、用於收集網路流量的流量收集模組、將網路流量拆解成多條連線並產生對應每一連線之長度特徵序列的流量拆解模組、比對長度特徵序列與資料庫中各種網路應用程式之長度共用子序列集合的辨識模組、以及判定所偵測之連線爲已知網路應用程式或未知網路應用程式的判定模組。透過本發明之萃取連線行爲特徵作網路流量辨識，可解決現有加密通訊或刻意隱藏封包內容之應用程式的偵測問題。

【英文】

Proposed is a traffic classification system based on message size sequence, comprising: a collection for common message size subsequence corresponding to a variety of network applications, a traffic collection module for collecting network traffic, a traffic decomposition module for decomposing the network traffic into multiple flows and producing a message size sequence corresponding to each flow, a flow classification module for comparing the message size sequence with the common message size subsequence collections of each network application in the database, and an arbitration module for determining the flow to be a known or an unknown network application. The present invention uses extraction of flow behavior characteristics for network traffic identification to resolve existing application packet detection issues about encrypted communication or deliberately hiding contents.

【代表圖】

【本案指定代表圖】：第（ 2 ）圖。

【本代表圖之符號簡單說明】：

2　基於訊息長度序列之網路流量辨識系統

21　資料庫

22　流量收集模組

23　流量拆解模組

24　辨識模組

25　判定模組

100　網路流量

【本案若有化學式時，請揭示最能顯示發明特徵的化學式】：

本案無化學式。

發明專利說明書

（本說明書格式、順序，請勿任意更動）

【發明名稱】（中文/英文）

基於訊息長度序列之網路流量辨識系統及其方法

A TRAFFIC CLASSIFICATION SYSTEM BASED ON MESSAGE SIZE SEQUENCE AND METHOD THEREOF

【技術領域】

本發明係關於一種網路流量分類技術，詳而言之，係關於一種基於訊息長度序列之分析封包屬性之網路流量辨識系統及其方法。

【先前技術】

隨著網際網路的蓬勃發展，需要各種網路應用程式透過網路進行訊息傳遞、資料傳送或溝通。在網路應用程式傳遞資料的過程中，資料將分成多個封包傳送，由於現行網路上存在許多有害的垃圾封包，爲確保資料傳遞的安全性，有效地辨識流量內容變得十分重要。

對於網路流量內所包含之網路應用程式而言，由於封包內含有網路位址（IP address）和連接埠（port），故早期在封包辨識時可採用公認的連接埠（well-known port）方式進行判定，也就是預先規劃連接埠號碼(port number)，規範不同網路應用程式採用不同連接埠號碼通行，例如port 80是給http protocol使用，但目前許多通訊軟體也都使用port 80，使得以連接埠號碼辨識封包所屬之網路應用程式的方式變得不可行，再者，現行許多網路上的有害封

包是採用隨機配置連接埠(port randomization)的技術，所以，使用前述方法是無法有效辨認出有害封包。此外，爲了解決封包辨識問題，另一種“封包內容特徵值”的比對方法現今被廣泛使用，即透過比對封包內容是否存在某些特定式樣(pattern)以辨識封包，舉例來說，防毒軟體可比對封包內容是否等同於資料庫內之關鍵字來判定是否排除，然而越來越多網路應用程式使用封包加密技術，使得無法利用前述方法來剖析封包內容，同時，還有些惡意程式會利用僞裝封包內容的方式意圖躲避內容特徵值的比對偵測，因此，可能產生誤擋或漏擋的問題，且利用剖析封包內容的偵測方式有侵害個人隱私的問題。其他需考量者，還包括這類的傳輸層行爲特徵比對方式皆需要收集足夠傳輸層資訊方能獲得正確判斷能力，也導致判斷時間過長。

因此，如何克服現有網路流量分類技術，特別是在不侵犯個人隱私且不受封包加密技術影響下，提供有效地網路流量辨識，實已成目前亟欲解決的課題。

【發明內容】

鑒於上述習知技術之缺點，本發明之目的係提出一種基於訊息長度序列之網路流量辨識系統及其方法，透過辨識封包大小與順序以判斷流量所屬網路應用程式爲何。

爲達成前述目的及其他目的，本發明提出一種基於訊息長度序列之網路流量辨識系統，係包括：資料庫、流量收集模組、流量拆解模組、辨識模組以及判定模組。該資料庫係預存對應各種網路應用程式之長度共用子序列集

合，該流量收集模組係用於收集網路流量，該流量拆解模組用於依據流量資訊將該網路流量拆解成多條連線，且擷取各該連線中複數封包的傳遞方向及長度大小，以由該複數封包產生對應各該連線之長度特徵序列，該辨識模組用於比對該長度特徵序列與該資料庫中之各種網路應用程式之長度共用子序列集合，以由該長度共用子序列集合中得到與該長度特徵序列之相似度最高者，最後判定模組依據該辨識模組所得到之該相似度最高者的數量判定該連線爲已知網路應用程式或未知網路應用程式。

於一實施例中，該流量拆解模組係自各該連線之複數封包中移除封包長度爲最大傳輸單位之封包以及封包內容(payload)長度爲零之封包。

於另一實施例中，該基於訊息長度序列之網路流量辨識系統更包括應用程式代表集合產生模組。該應用程式代表集合產生模組係利用已知的網路流量進行訓練，透過該流量拆解模組拆解已知的網路流量以產生各連線之長度特徵序列，並以兩兩一組的方式計算任兩條連線之最長長度共有子序列，收集各種組合所計算出之最長長度共有子序列以產生對應該應用程式之長度共用子序列集合。

本發明還提出一種基於訊息長度序列之網路流量辨識方法，係包括：提供對應各種網路應用程式之長度共用子序列集合；收集網路流量並拆解該網路流量成多條連線，擷取各該連線中複數封包的傳遞方向及長度大小，以產生對應各該連線之長度特徵序列；比對該長度特徵序列

與該各種網路應用程式之長度共用子序列集合，以由該長度共用子序列集合中取得與該長度特徵序列之相似度最高者；以及依據該長度共用子序列集合之相似度最高者的數量判斷該連線爲已知網路應用程式或未知網路應用程式。

於一實施例中，於擷取各該連線中複數封包的傳遞方向及長度大小之前，更包括自各該連線之複數封包中移除封包長度爲最大傳輸單位之封包以及封包內容長度爲零之封包。

於又一實施例中，該長度共用子序列集合係於辨識之前利用已知的網路流量進行訓練，包括：將已知的網路流量拆解成多條連線，並擷取各該連線中複數封包的傳遞方向及長度大小，以產生對應各該連線之長度特徵序列，各該連線之長度特徵序列以兩兩一組的方式計算該兩條連線之最長長度共有子序列，並收集各種組合所計算出之最長長度共有子序列，以產生該長度共用子序列集合。

相較於先前技術，本發明所提出之基於訊息長度序列之網路流量辨識系統及其方法，可用於偵測被加密或是刻意隱藏通訊協定及通訊內容之網路應用程式，藉此提供網路管理者取得充份資訊以執行流量控管。相較於傳統的流量偵測方式，不論是一般網路應用程式所用的公認的連接埠方式之判別或是利用封包內容特徵值作比對，不僅容易誤判且判斷時間較長，然而本發明透過萃取傳輸層行爲特徵的方式，可以網路應用程式連線行爲特徵作爲分析之依據，而非如往剖析封包內容之特徵資料，因而可對加密通

訊協定之網路應用程式作偵測，同時在無需解析封包內容下而不侵犯個人隱私，因此，透過本發明可解決傳統對於加密通訊或刻意隱藏封包內容之網路應用程式無法偵測等問題。

【圖式簡單說明】

第 1 圖係說明本發明之基於訊息長度序列之網路流量辨識系統的訓練與分類兩階段之流程圖；

第 2 圖係說明本發明之基於訊息長度序列之網路流量辨識系統於分類階段之系統架構圖；

第 3 圖係說明本發明之基於訊息長度序列之網路流量辨識系統於訓練階段之系統架構圖；

第 4 圖係說明本發明之基於訊息長度序列之網路流量辨識系統一訓練過程實施例之示意圖；

第 5 圖係說明本發明之基於訊息長度序列之網路流量辨識系統一辨識過程實施例之示意圖；以及

第 6 圖係說明本發明之基於訊息長度序列之網路流量辨識方法之步驟圖。

【實施方式】

以下係藉由特定的實施例說明本發明之實施方式，熟悉此技術之人士可由本說明書所揭示之內容輕易地瞭解本發明之其他特點與功效。本發明亦可藉由其他不同的具體實施例加以施行或應用。

參閱第 1 圖，其係說明本發明之基於訊息長度序列之網路流量辨識系統的訓練與分類兩階段之流程圖。如圖所

示，流程圖 1 顯示本發明可分爲第一階段 11 和第二階段 12，爲了得到網路應用程式連線行爲特徵資料作爲辨識依據，本發明需通過預先訓練以取得各種不同網路應用程式其連線行爲特徵資料，並藉由該些連線行爲特徵資料來判斷封包所屬網路應用程式爲何。

第一階段 11 爲代表訓練階段。首先，於流程 110 中，是追蹤網路應用程式以完成流量收集（traffic collection），爲了解析各網路應用程式的行爲特徵，本發明透過一套『網路應用程式流量收集技術』，在一台主機上執行想要收集的網路應用程式，限定該網路應用程式及其使用的埠號，使得只有該網路應用程式的網路流量才能通過主機網路介面，並且在網路流量出入端利用流量錄製技術將所需的流量錄製下來作爲分析之用，也就是說，流程 110 是收集網路應用程式的流量，且要收集足夠多網路流量（網路流量完整性會影響判斷準確性）。因此，利用一次訓練一種網路應用程式，錄製通過主機的封包，以用於網路應用程式之連線行爲的分析。

接著於流程 111 中，是擷取各連線的連線特徵（flow characterizing），以將所錄製到的網路流量拆解成多條連線（flow），雖然訓練階段一次僅訓練一種網路應用程式，但是網路應用程式在不同對象下會有不同的連線行爲，如此導致封包序列也有所不同，因而對網路流量依據傳遞對象進行分類，例如以 IP 位址爲分辨依據，將不同傳遞對象的流量分開，才能避免將多位傳遞對象的流量混在一起而造

成誤判。

於流程 112 中，是開發各流量中的最長共有子序列。簡單來說，自各流量可得到其所表示的子序列，接著，對各不同流量找出共同子序列中的最長者，即可用此共同子序列來代表訓練中的網路應用程式，如此一來，當某一網路流量具有類似於該最長共有子序列時，則有機會判斷該網路流量可能屬於該最長共有子序列所對應之網路應用程式。

於流程 113 中，是找出網路應用程式之代表長度特徵序列集合，亦即找出該網路應用程式於不同流量中各子序列的長度共用子序列集合。於前一流程 112 中，可能會得到多個最長共有子序列，因此，於本流程 113 中，將該些最長共有子序列組成集合，此集合即可代表所訓練的網路應用程式。

接著，進入第二階段 12，第二階段 12 爲實際分類階段，也就是利用第一階段 11 所得到的代表各種網路應用程式之長度共用子序列集合，作爲與真實網路流量比對的基準，藉由比對與代表各網路應用程式之長度共用子序列集合之間的相似度（similarity）差距，來推論所擷取到的網路連線是屬於哪種網路應用程式。

首先，於流程 120 和流程 121 中，同樣也是收集實際網路流量並進行流量拆解（traffic decomposition），也是依據如 IP 位址作爲分辨依據，將所擷取到網路流量拆解成多條連線，使每一個對象的網路流量能夠分開，以利於之後

對於封包序列的識別。

接著，於流程 122 中，是對網路流量進行分類以達到連線辨識（flow classification）的目的。簡單來說，即將流程 121 所取得之網路流量的封包序列與第一階段 11 所取得的各種代表不同網路應用程式之長度共用子序列集合進行比對，藉此找出最接近該網路流量之封包序列者。

最後，於流程 123 中，即利用前一流程 122 所得到之與欲判定網路流量之封包序列最接近者，找出其所對應之網路應用程式，即可得到所擷取到網路流量其所屬之網路應用程式。

由上可知，通過訓練找出網路應用程式其連線行爲，之後可以此與之後擷取到的新網路連線作比對，藉此判斷新網路連線所屬網路應用程式爲何者。需說明者，上述僅是針對單一網路應用程式比對的操作流程敘述，因此，如有多種網路應用程式需要比對，則僅需針對各不同的網路應用程式多次操作本流程即可。

參閱第 2 圖，其係說明本發明之基於訊息長度序列之網路流量辨識系統於分類階段之系統架構圖。如圖所示，基於訊息長度序列之網路流量辨識系統 2 主要說明於第 1 圖的第二階段中所執行的網路流量辨識，其中，基於訊息長度序列之網路流量辨識系統 2 係包括：資料庫 21、流量收集模組 22、流量拆解模組 23、辨識模組 24 以及判定模組 25。

資料庫 21 是用於預存對應各種網路應用程式之長度

共用子序列集合。如前面所述，長度共用子序列集合是用於代表某一網路應用程式，因而在辨識之前，可透過訓練方式得到可代表每一個網路應用程式之長度共用子序列集合，關於訓練方式，之後會有更詳盡說明。

流量收集模組 22 是用於收集網路流量 100，也就是將通過網路設備的網路流量 100 進行收集，該網路流量 100 可能混雜多種網路應用程式的封包，或屬於同一網路應用程式但與不同對象溝通的封包，因而收集後之後，需經拆解才能進行比對判斷。

流量拆解模組 23 是用於依據流量資訊將流量收集模組 22 所收集之網路流量 100 拆解成多條連線，亦即依據不同傳遞對象進行拆解，前述的流量資訊可爲來源 IP 位址、來源埠號、目的地 IP、目的地埠號及傳輸協定等，藉此拆解成多條連線，之後，再擷取各連線中複數封包的傳遞方向及長度大小，以由該連線之該些複數封包產生對應該連線之長度特徵序列（message size sequence）。

更具體來說，當拆解成多條連線後，每一連線即是與單一對象之間的封包傳遞，以其中一條連線爲例，將同一連線中之多個封包依序排列，並且找出該些封包的傳遞方向及長度大小，其中，傳遞方向可以連線發起者的傳遞方向爲正，反之爲負，最後，利用該連線中之複數封包產生對應此連線之長度特徵序列，即可作爲代表該連線之封包序列組合。

此外，於具體實施時，流量拆解模組 23 係自各連線之

複數封包中移除封包長度爲最大傳輸單位（maximum transmission unit）之封包以及封包內容長度爲零之封包。由於擷取網路流量時，所取得的封包對於找出網路應用程式的封包序列並不一定是有用的，例如：封包可分爲帶著控制訊息（control message）的封包以及帶著資料訊息（data message）的封包，由於伺服器端與使用者端連線後，將會以最大封包方式進行資料傳遞（以縮短傳遞時間），也就是說，用於攜帶資料訊息的封包其大小會達到最大傳輸長度，如此一來無法由該些封包看出差異，反觀，帶著控制訊息的封包可能是具有帳號密碼的訊息，因而封包大小會因連線過程而會有所差異。因此，流量拆解模組 23 會先將每一連線的複數封包中，其封包長度爲最大傳輸單位者及封包內容大小爲零者移除，以提高之後封包序列建立的可用性。

辨識模組 24 是用於將流量拆解模組 23 所取得某一連線之長度特徵序列與預存於資料庫 21 中的各種網路應用程式之長度共用子序列集合進行比對，藉此由長度共用子序列集合中找到與該連線之長度特徵序列間相似度最高者，也就是說，相似度越高者表示兩者網路連線行爲特徵越相近。

最後，判定模組 25 是依據辨識模組 24 所得到之相似度最高者的數量以判定該連線屬於已知網路應用程式或者是屬於未知網路應用程式。具體而言，於辨識模組 24 進行辨識時，可能找不到或找到一個以上的長度特徵序列與欲

辨識之連線的長度特徵序列相似，此時，若數量爲零或超過一個，則判定該連線爲未知網路應用程式，反之，若最相似的數量僅有一個，則可判定該連線爲已知網路應用程式。

於具體實施時，流量中複數封包所形成之長度特徵序列，係依據複數封包中的每一封包的傳遞方向及長度大小予以數值定義，並且依序排列該些數值而產生。如前所述，傳遞方向可以正負表示，封包大小可給予數值定義，例如，一封包是由連線發起端至接收端且封包大小爲 20KB，則可給予+20 的數值定義。因此，長度特徵序列可爲多個封包依其傳遞方向和封包長度大小所給予數值加以定義，所定義出的序列可供之後相似度判斷之用。

參閱第 3 圖，其係說明本發明之基於訊息長度序列之網路流量辨識系統於訓練階段之系統架構圖。如圖所示，基於訊息長度序列之網路流量辨識系統 3 內除包括資料庫 31、流量收集模組 32、流量拆解模組 33、辨識模組 34、判定模組 35 外，還包括在訓練的階段中，用於產生各網路應用程式之代表長度特徵序列集合的應用程式代表集合產生模組 36。

需先說明者，若在訓練的階段中，辨識模組 34 和判定模組 35 則無需運作，此外，基於訊息長度序列之網路流量辨識系統 3 所接收到的資料，是知道封包所屬網路應用程式爲何的已知的網路流量 200，而不會是不知道封包所屬網路應用程式爲何的網路流量 100，因此，辨識模組 34 和

判定模組 35 在訓練階段中是無需運作的，而網路流量 100 則無需提供。

應用程式代表集合產生模組 36 係利用已知的網路流量 200 進行訓練，透過該流量拆解模組 33 拆解該已知的網路流量以產生各連線之長度特徵序列，之後，將該些連線之長度特徵序列以兩兩一組的方式計算該兩條連線之最長長度共有子序列（longest size subsequence），並收集各種組合所計算出之最長長度共有子序列以產生該長度共用子序列集合。由上可知，爲了知悉各種網路應用程式的可能封包序列，因而利用應用程式代表集合產生模組 36 以找出對應不同網路應用程式之各種長度共用子序列集合。

具體來說，已知的網路流量 200 同樣由流量收集模組 32 進行收集，這裡的已知的網路流量 200 是指僅對單一網路應用程式的封包收集，接著傳送至流量拆解模組 33，流量拆解模組 33 也是將已知的網路流量 200 拆解成多條連線，亦即可依據前述的流量資訊，例如來源 IP 位址、來源埠號、目的地 IP、目的地埠號及傳輸協定等，之後，再擷取各連線中複數封包的傳遞方向及長度大小，以由該連線之該些複數封包產生對應該連線之長度特徵序列，流量收集模組 32 和流量拆解模組 33 與第 2 圖中的流量收集模組 22、流量拆解模組 23 的作用是相同。

接著，不同連線所產生之長度特徵序列會傳送至應用程式代表集合產生模組 36，應用程式代表集合產生模組 36 將以兩兩一組的方式，亦即任兩條連線一組，以計算兩條

連線之最長長度共有子序列，也就是取兩條連線中共有子序列的最長者，之後，在收集各種組合（任兩條一組）所計算出之最長長度共有子序列後，以產生該長度共用子序列集合。由此可知，長度共用子序列集合是包含網路應用程式所傳送之封包的可能集合。

此外，於前述計算過程中，若發現某長度較短之序列存在於某長度較長之共有子序列中時，即較短的序列已被長度較長的其他共有子序列所包含，此時可將長度較短者刪除，已被其他共有子序列所包含者是無需成爲長度共用子序列集合之一員。

參閱第 4 圖，其係說明本發明之基於訊息長度序列之網路流量辨識系統一訓練過程實施例之示意圖，該圖主要是說明如何找出最長長度共有子序列。如圖所示，某一網路應用程式其連線特徵的可能封包序列爲 1-2-3-4-5，在訓練過程中，得到 A 連線的長度特徵序列爲 1-2-3-4，而 B 連線的長度特徵序列爲 2-3-4-5，之後，透過第 3 圖之應用程式代表集合產生模組 36 以兩兩一組的方式進行計算，以找出兩條連線之間的最長長度共有子序列，於本範例中，即爲 2-3-4。

由上可知，將收集的已知的網路流量拆解成多條連線，再以兩兩一組進行最長長度共有子序列的計算，最後會得到多個最長長度共有子序列，即爲第 3 圖之應用程式代表集合產生模組 36 所產生該對應網路應用程式之長度共用子序列集合。此外，若有一組找到的最長長度共有子

序列例如 3-4 時，則該最長長度共有子序列（3-4）可由圖中之 A 連線和 B 連線所得到之最長長度共有子序列(2-3-4)所包含，故包含 3-4 的最長長度共有子序列是無需加入長度共用子序列集合中。

參閱第 5 圖，其係說明本發明之基於訊息長度序列之網路流量辨識系統一辨識過程實施例之示意圖，即第 5 圖主要是說明如何辨識長度特徵序列之間的相似度。在訓練某一網路應用程式後，其得到的長度共用子序列集合中的一個可能封包序列爲 1-2-3-4-5，爲方便進行相似度比對，則可依據封包方向和封包長度大小給予數值定義，如圖所示，該可能封包序列依序可被定義爲+10-10+15-10+20 的數值，其中 10、15、20 等數字僅爲封包大小的舉例。

之後，在新網路流量辨識過程中，網路流量也會被拆解成多條連線，其中，有一條連線的封包序列如情況 1，其包含有 2-3-4-5 的封包序列，且依封包方向和大小得到 -10+15-10+20 的數值，與可能封包序列（1-2-3-4-5）相比較，序列中有四個封包符合且數值相等，也就是說，情況 1 的連線與可能封包序列相似度高。

再考量情況 2，另一條連線包含有 1-2-3-4-5 的封包序列，且依封包方向和大小得到+10-8+16-10+19 的數值，與可能封包序列（1-2-3-4-5）相比較，序列中有五個封包順序符合，但其數值與封包序列（+10-10+15-10+20）相近，因而兩者相似度也很高。前述數值差異主要表示封包大小的不同，然而僅要正負號相同即表示傳遞方向相同，但封

包大小可與給予誤差值的容忍（視情況設立），以避免因封包過小差異導致數值不同的可能誤判。

若考量情況 1 和情況 2 的相似度高低，情況 2 本身已比情況 1 多一個封包相符，再者，若情況 2 的數值也都在可容忍範圍內下，則可判定情況 2 的相似度高於情況 1。

本發明對於所擷取之封包是不考慮先後時間，是僅考量封包出現的先後順序，當發送端發出數個封包後，接收端會有暫存區暫存該些封包，之後，再判斷是否有預存之封包序列與所收集之該些封包的順序類似，簡單來說，封包順序不對（不同於某一網路應用程式的可能封包序列）就判定不屬該網路應用程式的封包，只有當順序需符合資料庫內定義的封包順序時，才會進一步考慮封包大小是否相近。

上述是避免不同順序的封包被誤判，因爲，同一個網路應用程式中，不同的封包順序代表不同應用含意，例如訓練得到的封包序列爲 1-2-3，若新網路流量所拆解出連線的封包序列爲 3-1-2，則不能因兩者都有相同大小的封包就認爲兩者相似度高。再者，以往在判定相似度高低時，僅是考量封包是否相似（可能是封包數量或封包內容），因而對於具有相同大小的封包但有不同排列順序的兩連線可能會判定相似，但多個封包的排列不同下，仍有可能是不同網路應用程式所產生的，因此，本發明首重考量封包順序是否正確，以提高判斷時的準確性。

參閱第 6 圖，其係說明本發明之基於訊息長度序列之

網路流量辨識方法之步驟圖。如圖所示，於步驟 S601 中，係提供對應各種網路應用程式之長度共用子序列集合，於此所述之長度共用子序列集合可透過預先訓練得到，長度共用子序列集合是用來代表網路應用程式之封包序列的許多可能集合。接著至步驟 S602。

於步驟 S602 中，係收集網路流量並拆解該網路流量成多條連線，擷取各該連線中複數封包的傳遞方向及長度大小，以產生對應各該連線之長度特徵序列。於此步驟中，係將所收集到之網路流量依據流量資訊進行拆解，以將網路流量拆解成多條連線，其中，流量資訊可包括來源 IP 位址、來源埠號、目的地 IP、目的地埠號及傳輸協定等，亦即可分辨出傳遞對象的資訊。

此外，還包括於擷取各連線中複數封包的傳遞方向及長度大小之前，將各連線之複數封包中移除封包長度爲最大傳輸單位之封包以及封包內容長度爲零之封包，該些封包長度大小爲最大或內容爲零是無法成爲封包序列之一員，因爲該些封包是無法提供辨識效果，反而容易混淆判斷，因而在建立長度特徵序列前，需先去除該些無用封包。接著至步驟 S603。

於步驟 S603 中，係比對該長度特徵序列與該各種網路應用程式之長度共用子序列集合，以由該長度共用子序列集合中取得與該長度特徵序列之相似度最高者。於此步驟中，將步驟 S602 所取得之各連線之長度特徵序列與預先定義之各種網路應用程式之長度共用子序列集合進行比對，

以由各種網路應用程式之長度共用子序列集合找出最符合者，於此所述的比對中，要考量的包括封包序列的順序、個數和大小等相似程度，封包序列的順序即是指連線中有用封包的先後順序，若有明顯排列差異則完全不需比對，接著，再以封包個數和大小來判斷相似度。

舉例來說，可利用複數封包中每一封包之傳遞方向及長度大小給予數值定義，並且依序排列該些數值以產生代表連線之長度特徵序列。如前面第 5 圖所示的利用正負號及數值來表示方向和大小，必要時給予適當容忍值，以避免微小差距的封包大小所造成的誤判。接著至步驟 S604。

於步驟 S604 中，係依據該長度共用子序列集合的數量判斷該連線爲已知網路應用程式或未知網路應用程式。於此步驟中，在任一連線之長度特徵序列與長度共用子序列集合比對後，可能找出一個或一個以上相似度最高的長度特徵序列，因此，若僅有一個相似度最高，那可判定該連線屬於該長度共用子序列集合之長度特徵序列所對應之網路應用程式，反之，若超過一個以上時，則無法判定該連線所屬之網路應用程式爲何者，故判定其爲未知的網路應用程式。

於另一實施例中，本發明之步驟 S601 係提供有對應各種網路應用程式之長度共用子序列集合，然該長度共用子序列集合是透過預先訓練得到的。具體而言，長度共用子序列集合是預先利用已知的網路流量進行訓練，其包括：將已知的網路流量拆解成多條連線，並擷取各該連線中複

數封包的傳遞方向及長度大小，藉此得到對應各該連線之長度特徵序列，之後，將各該連線之長度特徵序列以兩兩一組的方式計算兩條連線之間的最長長度共有子序列，最後，收集各種組合所計算出之最長長度共有子序列，即可得到長度共用子序列集合。

上述長度共用子序列集合的訓練過程是利用一次訓練一種網路應用程式來進行，透過計算多個連線之間所共有的封包序列，以作爲該網路應用程式之代表。如此，當新網路流量所拆解出之連線，與長度共用子序列集合中之任何最長長度共有子序列比對後具有高相似度時，則可判定該連線是歸屬某類網路應用程式。

綜上所述，本發明所提出之基於訊息長度序列之網路流量辨識系統及其方法，即利用傳輸層連線行爲特徵，以網路應用程式之特定訊息長度序列作爲辨認網路流量中網路應用程式之依據，辨識時，將網路應用程式連線在傳輸層行爲所萃取出的連線行爲長度序列作爲該連線之代表特徵，與已知的各種網路應用程式代表長度特徵序列做相似度比對，並以相似度最大之網路應用程式作爲最後歸屬。與習知技術相比較，傳統的網路應用程式流量偵測和辨識技術多是採用網路應用程式所使用的已知連接埠或者是採用封包內容特徵值的比對方式，因而本發明克服習知技術的兩個缺點：（1）無法偵測使用動態連接埠之網路應用程式，以及（2）封包內容如果被網路應用程式加密傳送就無法透過封包內容特徵值比對辨認。因此，本發明解決了現

有無法利用封包內容辨認的問題以及使用動態連接埠無法辨認的問題，並且提供一種可以用來作爲線上閘道器使用之辨認機制。

上述實施例僅例示性說明本發明之原理及其功效，而非用於限制本發明。任何熟習此項技藝之人士均可在不違背本發明之精神及範疇下，對上述實施例進行修飾與改變。因此，本發明之權利保護範圍，應如後述之申請專利範圍所列。

【符號說明】

1	流程圖
11	第一階段
110~113	流程
12	第二階段
120~123	流程
2、3	基於訊息長度序列之網路流量辨識系統
21、31	資料庫
22、32	流量收集模組
23、33	流量拆解模組
24、34	辨識模組
25、35	判定模組
36	應用程式代表集合產生模組
100	網路流量
200	已知的網路流量
S601~S604	步驟

申請專利範圍

1. 一種基於訊息長度序列之網路流量辨識系統，係包括：

資料庫，係用於預存對應各種網路應用程式之長度共用子序列集合；

流量收集模組，係用於收集網路流量；

流量拆解模組，係用於依據流量資訊將該網路流量拆解成多條連線，且擷取各該連線中複數封包的傳遞方向及長度大小，以由該複數封包產生對應各該連線之長度特徵序列；

辨識模組，係用於比對該長度特徵序列與該資料庫中之各種網路應用程式之長度共用子序列集合，以由該長度共用子序列集合中得到與該長度特徵序列之相似度最高者；以及

判定模組，係依據該辨識模組所得到之該相似度最高者的數量判定該連線爲已知網路應用程式或未知網路應用程式。

2. 如申請專利範圍第 1 項所述之基於訊息長度序列之網路流量辨識系統，其中，該流量拆解模組係自各該連線之複數封包中移除封包長度爲最大傳輸單位之封包以及封包內容長度爲零之封包。

3. 如申請專利範圍第 1 項所述之基於訊息長度序列之網路流量辨識系統，其中，該長度特徵序列係指依據該複數封包中每一封包的傳遞方向及長度大小予以數值定義，且依序排列該些數值所產生者。

4. 如申請專利範圍第 1 項所述之基於訊息長度序列之網路流量辨識系統，其中，該流量資訊係包括來源 IP 位址、來源埠號、目的地 IP、目的地埠號及傳輸協定。

5. 如申請專利範圍第 1 項所述之基於訊息長度序列之網路流量辨識系統，更包括應用程式代表集合產生模組，其係利用已知的網路流量進行訓練，透過該流量拆解模組拆解該已知的網路流量以產生各連線之長度特徵序列，將該些連線之長度特徵序列以兩兩一組的方式計算該兩條連線之最長長度共有子序列，並收集各種組合所計算出之最長長度共有子序列以產生該長度共用子序列集合。

6. 一種基於訊息長度序列之網路流量辨識方法，係包括：

提供對應各種網路應用程式之長度共用子序列集合；

收集網路流量並拆解該網路流量成多條連線，擷取各該連線中複數封包的傳遞方向及長度大小，以產生對應各該連線之長度特徵序列；

比對該長度特徵序列與該各種網路應用程式之長度共用子序列集合，以由該長度共用子序列集合中取得與該長度特徵序列之相似度最高者；以及

依據該長度共用子序列集合的數量判斷該連線爲已知網路應用程式或未知網路應用程式。

7. 如申請專利範圍第 6 項所述之基於訊息長度序列之網路流量辨識方法，其中，於擷取各該連線中複數封包

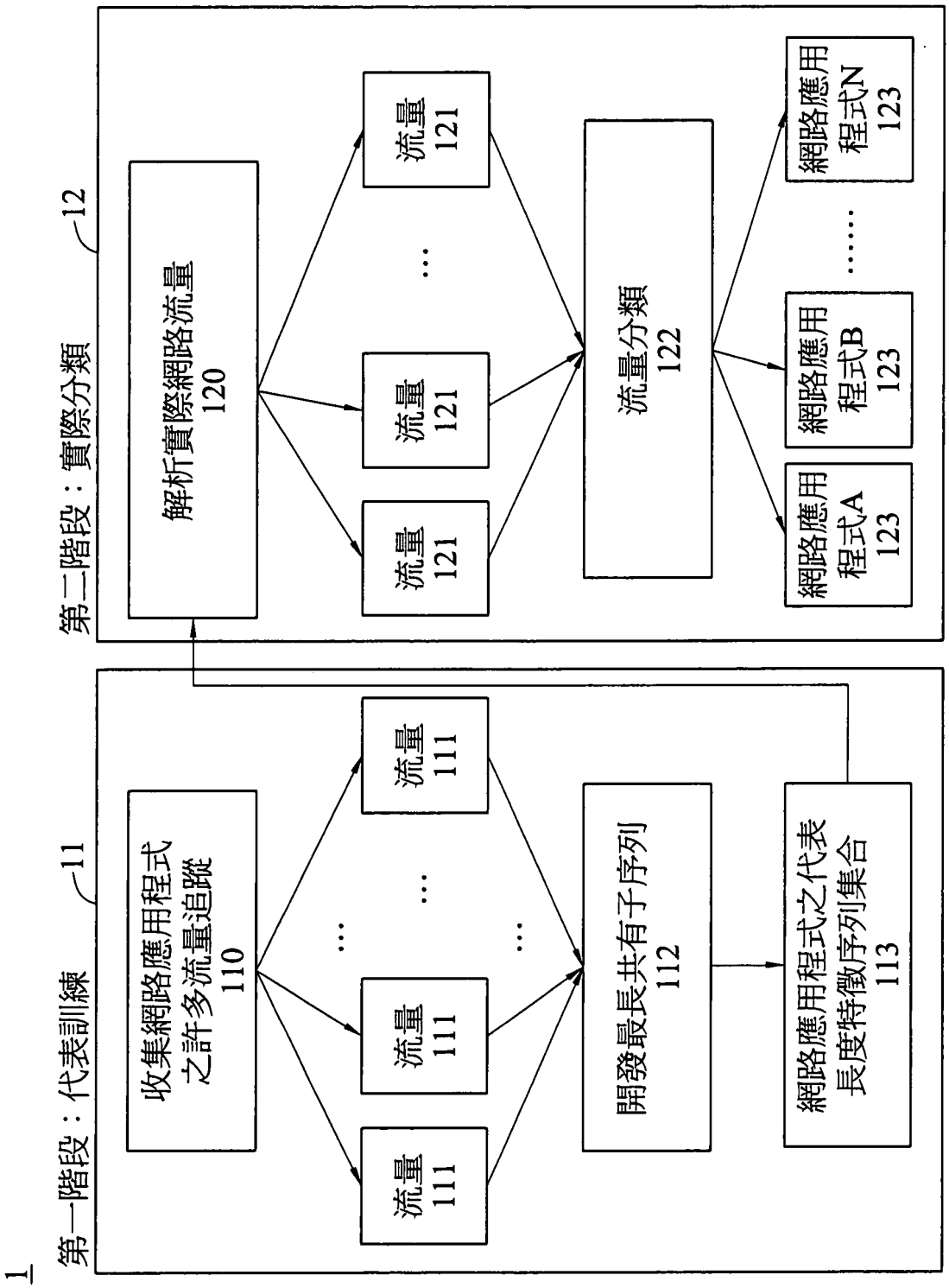
的傳遞方向及長度大小之前，更包括自各該連線之複數封包中移除封包長度爲最大傳輸單位之封包以及封包內容長度爲零之封包。

8. 如申請專利範圍第 6 項所述之基於訊息長度序列之網路流量辨識方法，其中，該長度特徵序列係指依據該複數封包中每一封包的傳遞方向及長度大小予以數值定義，且依序排列該些數值所產生者。

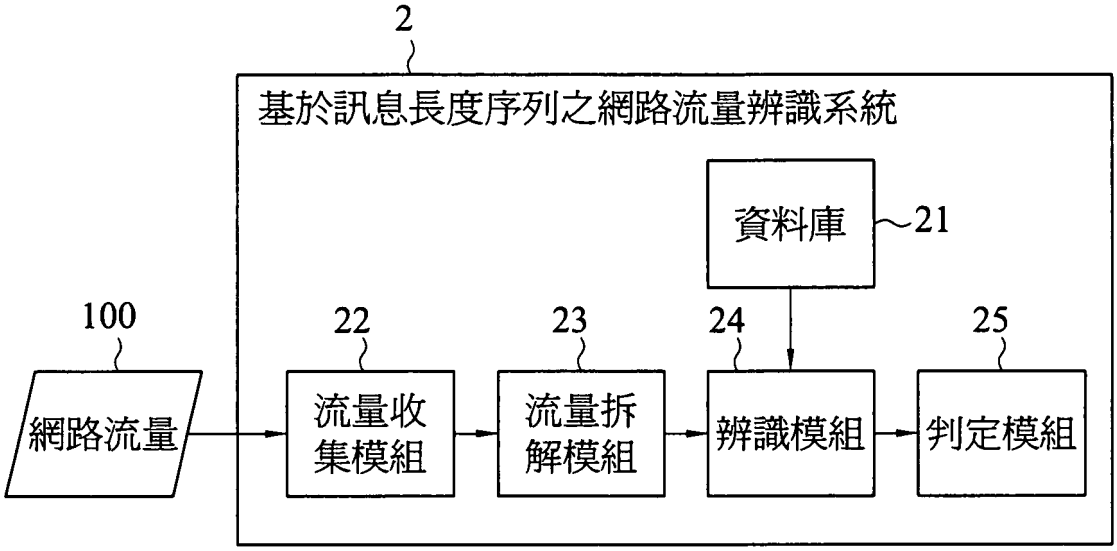
9. 如申請專利範圍第 6 項所述之基於訊息長度序列之網路流量辨識方法，其中，拆解該網路流量成多條連線係包括以來源 IP 位址、來源埠號、目的地 IP、目的地埠號及傳輸協定來拆解該網路流量。

10. 如申請專利範圍第 6 項所述之基於訊息長度序列之網路流量辨識方法，其中，該長度共用子序列集合係於辨識之前利用已知的網路流量進行訓練，包括：將已知的網路流量拆解成多條連線，並擷取各該連線中複數封包的傳遞方向及長度大小，以產生對應各該連線之長度特徵序列，將各該連線之長度特徵序列以兩兩一組的方式計算該兩條連線之最長長度共有子序列，並收集各種組合所計算出之最長長度共有子序列，以產生該長度共用子序列集合。

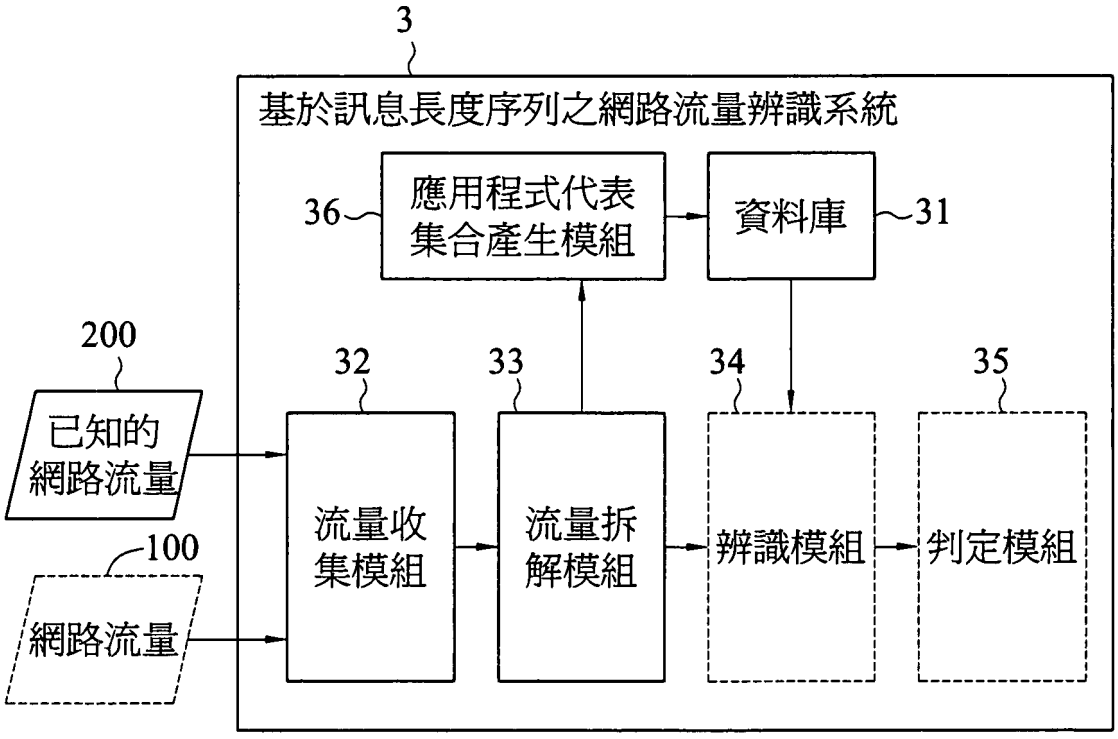
圖式


1
第一階段：代表訓練
11
收集網路應用程式之許多流量追蹤 110
流量 111
流量 111
流量 111
開發最長共有子序列 112
網路應用程式之代表長度特徵序列集合 113
第二階段：實際分類
12
解析實際網路流量 120
流量 121
流量 121
流量 121
流量分類 122
網路應用程式A 123
網路應用程式B 123
網路應用程式N 123

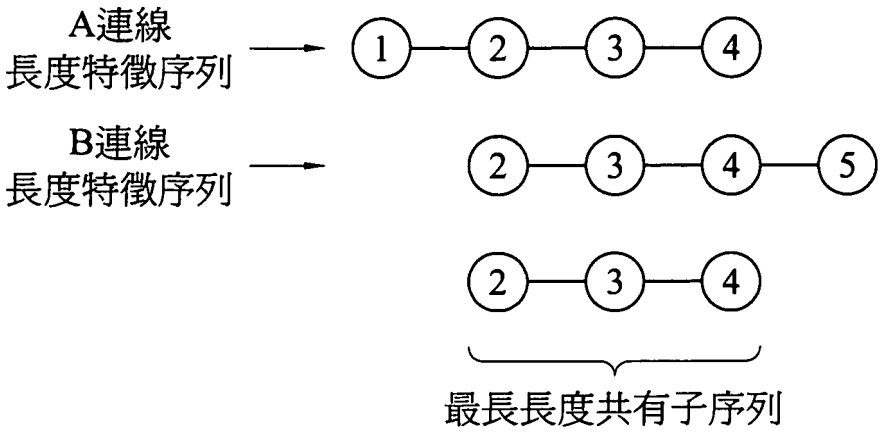
第1圖


2
基於訊息長度序列之網路流量辨識系統
資料庫
21
100
22
23
24
25
網路流量
流量收集模組
流量拆解模組
辨識模組
判定模組

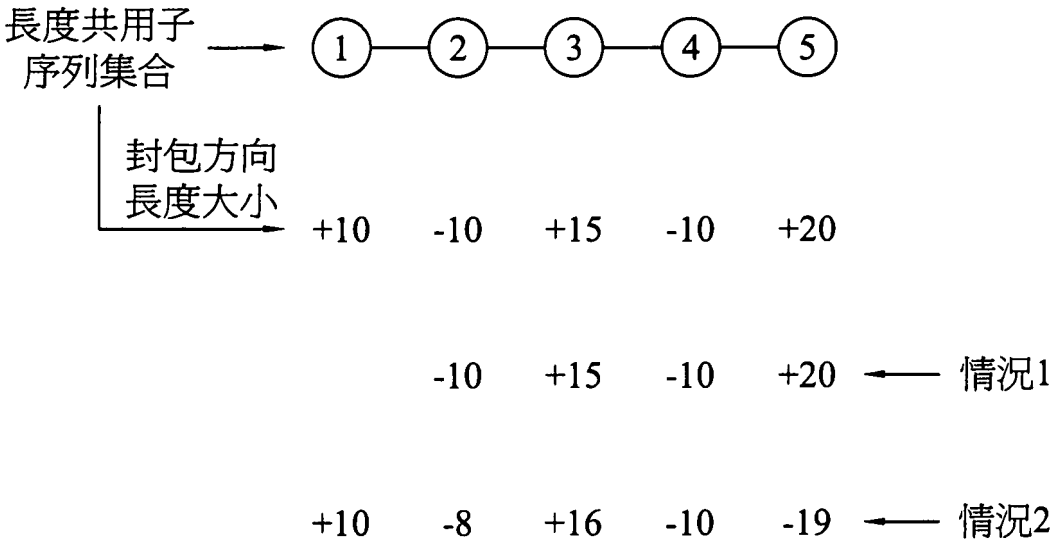
第2圖


3
基於訊息長度序列之網路流量辨識系統
36
應用程式代表集合產生模組
資料庫
31
200
32
33
34
35
已知的網路流量
100
網路流量
流量收集模組
流量拆解模組
辨識模組
判定模組

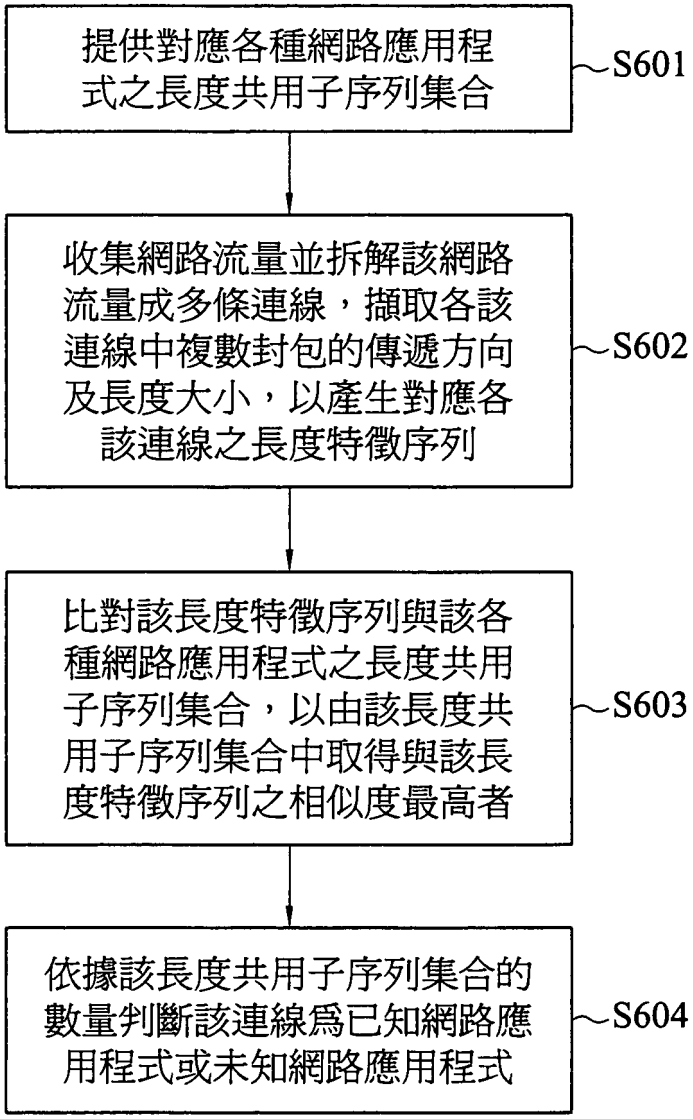
第3圖


A連線
長度特徵序列
1
2
3
4
B連線
長度特徵序列
2
3
4
5
2
3
4
最長長度共有子序列

第4圖


長度共用子
序列集合
1
2
3
4
5
封包方向
長度大小
+10
-10
+15
-10
+20
-10
+15
-10
+20
情況1
+10
-8
+16
-10
-19
情況2

第5圖


提供對應各種網路應用程式之長度共用子序列集合
S601
收集網路流量並拆解該網路流量成多條連線，擷取各該連線中複數封包的傳遞方向及長度大小，以產生對應各該連線之長度特徵序列
S602
比對該長度特徵序列與該各種網路應用程式之長度共用子序列集合，以由該長度共用子序列集合中取得與該長度特徵序列之相似度最高者
S603
依據該長度共用子序列集合的數量判斷該連線爲已知網路應用程式或未知網路應用程式
S604

第6圖