(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2011-0121241
(43) 공개일자 2011년11월07일

(51) Int. Cl.
H04L 12/26 (2006.01) *H04L 29/06* (2006.01)
(21) 출원번호 10-2010-0040742
(22) 출원일자 2010년04월30일
심사청구일자 2010년04월30일

(71) 출원인
고려대학교 산학협력단
서울 성북구 안암동5가 1
시큐아이닷컴 주식회사
서울특별시 강남구 삼성동 157-16 제이엠빌딩 1~9층
(72) 발명자
김명섭
충청남도 연기군 조치원읍 신흥리 푸르지오1차 107-401
박진완
충청남도 연기군 조치원읍 서창리 고려대학교세종캠퍼스 과학기술대학 210-B
(뒷면에 계속)
(74) 대리인
특허법인다나

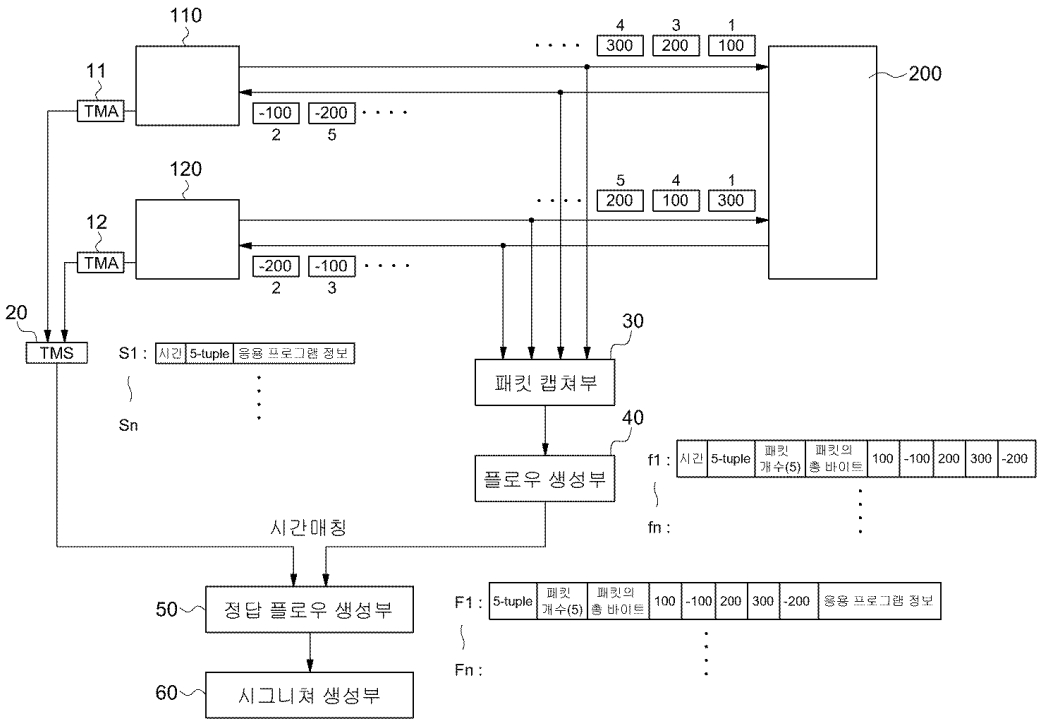
전체 청구항 수 : 총 20 항

(54) 통계 시그니쳐 생성 장치 및 방법

(57) 요 약

본 발명은 통계 시그니쳐 생성 장치 및 방법에 관한 것으로, 플로우 중 초기 N개 패킷의 페이로드 크기와 방향을 이용하여 통계 시그니쳐를 생성하는 것으로, 5-튜플, 초기 N개의 패킷의 페이로드 크기 및 방향, 응용 프로그램 정보를 포함하는 복수의 정답 플로우를 생성하고, 복수의 정답 플로우를 N차원의 복수의 벡터 플로우로 변환하고, 복수의 벡터 플로우 사이의 거리가 설정된 거리 임계값 미만인 벡터 플로우들을 하나의 그룹으로 그룹핑하여 통계 시그니쳐를 생성함으로써, 실시간 트래픽 분류에서 트래픽 분류 시간을 단축시킬 수 있는 통계 시그니쳐를 제공할 수 있다.

대 표 도



(72) 발명자

박준상

전라남도 무안군 현경면 수양리 783

윤성호

경상북도 포항시 북구 용흥1동 현대타워 104동 1205호

김경모

서울특별시 강남구 대치3동 949-11 포르쉐타워 5층

특허청구의 범위

청구항 1

복수의 클라이언트 호스트와 복수의 서버 호스트 사이의 패킷들을 캡쳐하는 패킷 캡쳐부,

상기 패킷 캡쳐부의 상기 패킷들 중 5-튜플이 동일한 초기 N개의 패킷을 이용하여 시간, 상기 5-튜플, 상기 초기 N개 패킷의 페이로드 크기 및 방향을 포함하는 플로우 데이터를 생성하는 플로우 생성부,

상기 복수의 클라이언트 호스트에 각각 설치되어 시간, 상기 5-튜플, 상기 시간에 따른 응용 프로그램 정보를 포함하는 복수의 로그 정보를 수집하는 복수의 티엠에이,

상기 복수의 티엠에이와 연결되어 상기 복수의 클라이언트 호스트에서 발생하는 상기 복수의 로그 정보를 수집하는 티엠에스,

상기 티엠에스와 상기 플로우 생성부 사이에 연결되어, 상기 복수의 로그 정보의 시간과 상기 플로우 데이터의 시간을 매칭하여 상기 5-튜플, 상기 초기 N개의 패킷의 페이로드 크기 및 방향, 상기 응용 프로그램 정보를 포함하는 복수의 정답 플로우를 생성하는 정답 플로우 생성부, 및

상기 정답 플로우 생성부와 연결되어, 상기 복수의 정답 플로우의 페이로드 크기와 방향을 이용하여 상기 복수의 정답 플로우를 N차원의 복수의 벡터 플로우로 변환하고, 상기 복수의 벡터 플로우 사이의 거리가 설정된 거리 임계값 미만인 벡터 플로우들을 하나의 그룹으로 그룹핑하여, 상기 하나의 그룹을 하나의 시그니쳐로 생성하는 시그니쳐 생성부를 포함하며,

상기 5-튜플은 소스 IP, 목적지 IP, 소스 포트, 목적지 포트, 프로토콜을 포함하는 통계 시그니쳐 생성 장치.

청구항 2

제1항에 있어서,

상기 시그니쳐 생성부는,

상기 정답 플로우 생성부와 연결되어, 상기 복수의 정답 플로우 중 하나의 응용 프로그램에 대해 설정된 제1 기준 개수 이상의 정답 플로우를 추출하는 플로우 추출부,

상기 플로우 추출부와 연결되어, 상기 플로우 추출부에서 추출된 정답 플로우를 N차원의 상기 복수의 벡터 플로우로 변환하는 벡터 변환부,

상기 벡터 변환부와 연결되어, 상기 복수의 벡터 플로우 사이의 거리가 상기 거리 임계값 미만인 벡터 플로우들을 하나의 그룹으로 그룹핑하여 복수의 그룹을 형성하는 그룹핑부, 및

상기 복수의 그룹에 포함되는 상기 복수의 벡터 플로우를 제외한 벡터 플로우를 제거하는 최적화부를 포함하는 통계 시그니쳐 생성 장치.

청구항 3

제2항에 있어서,

상기 플로우 추출부는,

상기 복수의 정답 플로우 중 하나의 응용 프로그램에 대해 상기 복수의 클라이언트 호스트 각각에서 설정된 제2 기준 개수 이상의 정답 플로우를 추출하는 통계 시그니쳐 생성 장치.

청구항 4

제2항에 있어서,

상기 그룹핑부에서 그룹핑된 상기 하나의 그룹은 프로세스 코드, 전송 프로토콜, 대표값, 대표값의 차원 및 거리 임계값의 5가지 속성을 가지며,

상기 대표값은 상기 하나의 그룹의 모든 벡터 플로우의 벡터를 산술평균을 통해 계산한 벡터로 상기 하나의 그

룹의 중심점이고,

상기 거리 임계값은 상기 대표값으로부터 가장 먼 거리에 있는 벡터 플로우와 상기 대표값 사이의 시티 블록 거리인 통계 시그니쳐 생성 장치.

청구항 5

제4항에 있어서,

상기 시티 블록 거리는 N차원에서 두 벡터 플로우 사이의 거리인 통계 시그니쳐 생성 장치.

청구항 6

제5항에 있어서,

상기 그룹핑부는

상기 벡터 변환부로부터 새로운 벡터 플로우가 입력되면, 상기 새로운 벡터 플로우와의 거리가 상기 거리 임계값 미만인 상기 복수의 그룹 중에서 상기 시티 블록 거리가 가장 가까운 그룹에 상기 새로운 벡터 플로우를 그룹핑하는 통계 시그니쳐 생성 장치.

청구항 7

제5항에 있어서,

상기 그룹핑부는,

상기 벡터 변환부로부터 새로운 벡터 플로우가 입력되고, 상기 새로운 벡터 플로우와의 거리가 상기 거리 임계값 미만인 그룹이 존재하지 않을 경우, 새로운 그룹에 상기 새로운 벡터 플로우를 그룹핑하는 통계 시그니쳐 생성 장치.

청구항 8

제4항에 있어서,

상기 최적화부는,

상기 대표값으로부터 가장 멀리 떨어진 벡터 플로우와 상기 대표값 사이의 거리가 새로운 거리 임계값이 되게 상기 거리 임계값을 상기 새로운 거리 임계값으로 보정하는 통계 시그니쳐 생성 장치.

청구항 9

제4항에 있어서,

상기 최적화부는,

상기 하나의 그룹에 포함되는 모든 벡터 플로우가 동일한 상기 목적지 포트를 가질 경우, 상기 하나의 그룹에 상기 목적지 포트를 할당하는 통계 시그니쳐 생성 장치.

청구항 10

제2항에 있어서,

상기 시그니쳐 생성부는,

상기 복수의 그룹 사이에 충돌 영역을 추출하고, 상기 복수의 그룹에서 상기 충돌 영역이 제거된 최종 그룹을 형성하는 충돌 처리부를 더 포함하는 통계 시그니쳐 생성 장치.

청구항 11

(a) 복수의 클라이언트 호스트와 복수의 서버 호스트 사이의 패킷들을 캡쳐하는 단계,

(b) 상기 패킷들 중 5-튜플이 동일한 초기 N개의 패킷을 이용하여 시간, 상기 5-튜플, 상기 초기 N개의 패킷의 페이로드 크기 및 방향을 포함하는 플로우 데이터를 생성하는 단계,

(c) 상기 복수의 클라이언트 호스트에서 시간, 상기 5-튜플, 응용 프로그램 정보를 포함하는 복수의 로그 정보를 수집하는 단계,

(d) 상기 복수의 로그 정보의 시간과 상기 플로우 데이터의 시간을 매칭시켜, 상기 5-튜플, 상기 초기 N개의 패킷의 페이로드 크기 및 방향, 상기 응용 프로그램 정보를 포함하는 복수의 정답 플로우를 생성하는 단계,

(e) 상기 정답 플로우의 페이로드 크기와 방향을 이용하여 상기 복수의 정답 플로우를 N차원의 복수의 벡터 플로우로 변환하는 단계, 및

(f) 상기 복수의 벡터 플로우 사이의 거리가 설정된 거리 임계값 미만인 벡터 플로우들을 하나의 그룹으로 그룹핑하여 복수의 그룹을 형성하는 단계를 포함하며,

상기 5-튜플은 소스 IP, 목적지 IP, 소스 포트, 목적지 포트, 프로토콜을 포함하며, 상기 하나의 그룹은 하나의 시그니처인 통계 시그니처 생성 방법.

청구항 12

제11항에 있어서,

상기 (e) 단계는,

상기 정답 플로우 중 하나의 응용 프로그램에 대해 설정된 제1 기준 개수 이상의 정답 플로우를 추출하는 단계, 및

추출된 상기 제1 기준 개수 이상의 정답 플로우를 N차원의 상기 복수의 벡터 플로우로 변환하는 단계를 포함하는 통계 시그니처 생성 방법.

청구항 13

제11항에 있어서,

상기 (e) 단계는,

상기 정답 플로우 중 하나의 응용 프로그램에 대해 상기 복수의 클라이언트호스트 각각에서 설정된 제2 기준 개수 이상의 정답 플로우를 추출하는 단계, 및

추출된 상기 제2 기준 개수 이상의 정답 플로우를 N차원의 상기 복수의 벡터 플로우로 변환하는 단계를 포함하는 통계 시그니처 생성 방법.

청구항 14

제11항에 있어서,

상기 복수의 그룹 각각은 서로 다른 5가지 속성을 가지며,

상기 5가지 속성은 프로세스 코드, 전송 프로토콜, 대표값, 대표값의 차원 및 거리 임계값이며,

상기 대표값은 상기 하나의 그룹의 모든 벡터 플로우의 벡터를 산술평균을 통해 계산한 벡터로 상기 하나의 그룹의 중심점이고,

상기 거리 임계값은 상기 대표값으로부터 가장 먼 거리에 있는 벡터 플로우와 상기 대표값 사이의 시티 블록 거리인 통계 시그니처 생성 방법.

청구항 15

제14항에 있어서,

상기 시티 블록 거리는 N차원에서 두 벡터 플로우 사이의 거리인 통계 시그니처 생성 방법.

청구항 16

제15항에 있어서,

상기 (f) 단계는,

새로운 벡터 플로우가 입력되면, 상기 새로운 벡터 플로우와의 거리가 상기 거리 임계값 미만인 상기 복수의 그룹을 검색하고, 상기 거리 임계값 미만인 상기 복수의 그룹 중에서 상기 시티 블록 거리가 가장 가까운 그룹에 상기 새로운 벡터 플로우를 그룹핑하는 단계를 더 포함하는 통계 시그니쳐 생성 방법.

청구항 17

제15항에 있어서,

상기 (f) 단계는,

새로운 벡터 플로우가 입력되고, 상기 새로운 벡터 플로우와의 거리가 상기 임계값 미만인 그룹이 존재하지 않을 경우, 새로운 그룹에 상기 새로운 벡터 플로우를 그룹핑하는 단계를 더 포함하는 통계 시그니쳐 생성 방법.

청구항 18

제14항에 있어서,

상기 (f) 단계에 이어,

상기 복수의 그룹에 포함되는 상기 복수의 벡터 플로우를 제외한 벡터 플로우를 제거하는 단계, 및

상기 대표값으로부터 가장 멀리 떨어진 벡터 플로우와 상기 대표값 사이의 거리가 새로운 거리 임계값이 되게 상기 거리 임계값을 상기 새로운 거리 임계값으로 보정하는 단계를 더 포함하는 통계 시그니쳐 생성 방법.

청구항 19

제14항에 있어서,

상기 (f) 단계에 이어,

상기 하나의 그룹에 포함되는 모든 벡터 플로우가 동일한 상기 목적지 포트를 가질 경우, 상기 하나의 그룹에 상기 목적지 포트를 할당하는 통계 시그니쳐 생성 방법.

청구항 20

제11항에 있어서,

상기 (f) 단계에 이어,

상기 복수의 그룹 사이에 충돌 영역을 추출하고, 상기 복수의 그룹에서 상기 충돌 영역이 제거된 최종 그룹을 형성하는 단계를 포함하는 통계 시그니쳐 생성 방법.

명 세 서

기 술 분 야

[0001] 본 발명은 통계 시그니쳐 생성 장치 및 방법에 관한 것으로, 플로우 중 초기N개 패킷의 페이로드 크기 및 방향을 이용하여 통계 시그니쳐를 생성하는 통계 시그니쳐 생성 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 오늘날의 네트워크에서는 다양한 응용의 등장으로 인해 트래픽이 복잡 다양해지고 있다. 이러한 상황 속에서 네트워크의 효과적인 운용과 관리를 위해 네트워크 트래픽 분석은 필수적인 요소가 되었다. 네트워크의 의존도가 앞으로 계속 증대됨에 따라 트래픽 분석의 중요성은 계속 커질 전망이다.

[0003] 엔터프라이즈 네트워크나 캠퍼스 네트워크에서는 네트워크 자원의 효율적인 운용 및 관리를 위해서 다양한 정책들을 마련하고 있다. 예를 들어, 학교나 공공 기관에서 P2P 트래픽, 게임 트래픽 등 네트워크 자원을 많이 차지하거나 업무와 관련이 없는 트래픽들을 제한하는 정책들이 있다. 이러한 정책들을 펼치기 위해서는 빠르고 정확한 응용계층 트래픽 분류가 필수적인 요소이다.

[0004] 응용계층 트래픽 분류란 네트워크 패킷을 수집하여 해당 패킷을 발생시킨 응용계층의 정체를 알아내고 이를 기

준으로 트래픽을 분류하는 것을 말한다.

[0005] 실시간의 정확한 응용계층 트래픽 분류의 중요성에 의해 현재까지 많은 연구들이 진행되어 왔다. 기존의 분류 방법들로는 포트 기반 분석, 페이로드 시그니쳐 기반 분석, 머신러닝 기반 분석, 플로우 상관관계 기반 분석 등 다양한 방법론들이 존재한다.

[0006] 이러한 다양한 방법론들이 있지만, 새로운 응용 프로그램의 등장, 응용 프로그램의 업데이트 등 응용의 변화에 따라 트래픽이 변하고 있기 때문에 오늘날의 응용 별 트래픽 분류는 더욱 더 힘들어지고 있다. 따라서, 보다 정확한 응용 별 트래픽 분류를 위하여 다양한 분류 방법론에 대한 지속적인 연구가 필요하다.

[0007] 최근 연구에서는 기존의 방법론들의 단점을 극복하고 빠르고 정확한 응용계층 트래픽 분류를 위해 패킷들의 통계 정보를 이용한 분류 방법을 연구 중이다. 통계 정보를 이용한 분류 방법은 패킷 크기, 패킷 간의 시간, 윈도우 크기 등에 대한통계적 분포를 바탕으로 통계 시그니쳐를 생성해 내고, 이를 바탕으로 트래픽을 분석하는 방법론이다. 이 방법론은 최근 증가하고 있는 암호화된 트래픽의 분석에 용이하며, 패킷의 페이로드 정보를 분석하지 않기 때문에 트래픽을 빠른 속도로 분류할 수 있다는 장점을 가진다. 하지만, 통계 정보를 생성하기 위해 플로우가 끝날 때까지 기다려야 한다는 점이 실시간 트래픽 분류를 어렵게 하고 있다.

발명의 내용

해결하려는 과제

[0008] 본 발명은 플로우가 끝나기 전에 실시간 트래픽 분류가 가능하도록 플로우 중 초기 N개 패킷의 페이로드 크기와 방향을 이용하여 통계 시그니쳐를 생성하는 통계 시그니쳐 생성 장치 및 방법을 제공한다.

과제의 해결 수단

[0009] 본 발명의 한 특징에 따르면, 복수의 클라이언트 호스트와 복수의 서버 호스트 사이의 패킷들을 캡쳐하는 패킷 캡쳐부, 상기 패킷 캡쳐부의 상기 패킷들 중 5-튜플이 동일한 초기 N개의 패킷을 이용하여 시간, 상기 5-튜플, 상기 초기 N개 패킷의 페이로드 크기 및 방향을 포함하는 플로우 데이터를 생성하는 플로우 생성부, 상기 복수의 클라이언트 호스트에 각각 설치되어 시간, 상기 5-튜플, 상기 시간에 따른 응용 프로그램 정보를 포함하는 복수의 로그 정보를 수집하는 복수의 티엠에이, 상기 복수의 티엠에이와 연결되어 상기 복수의 클라이언트 호스트에서 발생하는 복수의 로그 정보를 수집하는 로그 티엠에스, 상기 티엠에스와 상기 플로우 생성부 사이에 연결되어, 상기 복수의 로그 정보의 시간과 상기 플로우 데이터의 시간을 매칭하여 상기 5-튜플, 상기 초기 N개의 패킷의 페이로드 크기 및 방향, 상기 응용 프로그램 정보를 포함하는 복수의 정답 플로우를 생성하는 정답 플로우 생성부, 상기 정답 플로우 생성부와 연결되어, 상기 복수의 정답 플로우의 페이로드 크기와 방향을 이용하여 상기 복수의 정답 플로우를 N차원의 복수의 벡터 플로우로 변환하고, 상기 복수의 벡터 플로우 사이의 거리가 설정된 거리 임계값 미만인 벡터 플로우들을 하나의 그룹으로 그룹핑하여, 상기 하나의 그룹을 하나의 시그니쳐로 생성하는 시그니쳐 생성부를 포함하는 통계 시그니쳐 생성 장치가 제공된다.

[0010] 이때, 상기 5-튜플은 소스 IP, 목적지 IP, 소스 포트, 목적지 포트, 프로토콜을 포함한다.

[0011] 본 발명의 다른 특징에 따르면, 복수의 클라이언트 호스트와 복수의 서버 호스트 사이의 패킷들을 캡쳐하는 단계, 상기 패킷들 중 5-튜플이 동일한 초기 N개의 패킷을 이용하여 시간, 상기 5-튜플, 상기 초기 N개의 패킷의 페이로드 크기 및 방향을 포함하는 플로우 데이터를 생성하는 단계, 상기 복수의 클라이언트 호스트에서 시간, 상기 5-튜플, 응용 프로그램 정보를 포함하는 복수의 로그 정보를 수집하는 단계, 상기 복수의 로그 정보의 시간과 상기 플로우 데이터의 시간을 매칭시켜, 상기 5-튜플, 상기 초기 N개의 패킷의 페이로드 크기 및 방향, 상기 응용 프로그램 정보를 포함하는 복수의 정답 플로우를 생성하는 단계, 상기 정답 플로우의 페이로드 크기와 방향을 이용하여 상기 복수의 정답 플로우를 N차원의 복수의 벡터 플로우로 변환하는 단계, 상기 복수의 벡터 플로우 사이의 거리가 설정된 거리 임계값 미만인 벡터 플로우들을 하나의 그룹으로 그룹핑하여 복수의 그룹을 형성하는 단계를 포함하는 통계 시그니쳐 생성 방법이 제공된다.

[0012] 이때, 상기 5-튜플은 소스 IP, 목적지 IP, 소스 포트, 목적지 포트, 프로토콜을 포함하며, 상기 하나의 그룹은 하나의 시그니쳐인 통계 시그니쳐 생성 방법.

발명의 효과

[0013] 본 발명의 실시 예에서는 패킷의 페이로드 크기 및 방향을 이용하여 시그니쳐를 생성하여, 실시간 트래픽 분류

에서 트래픽 분류 시간이 빨라지는 효과가 있다.

[0014] 또한, 본 발명에서는 시그니쳐 사이에 충돌이 발생하는 충돌 영역을 제거하여 트래픽이 잘못 분류되는 경우를 방지할 수 있다.

도면의 간단한 설명

[0015] 도 1은 본 발명의 실시 예에 따른 통계 시그니쳐 생성 장치의 개략적인 구성도이다.

도 2는 본 발명의 실시 예에 따른 통계 시그니쳐 생성부의 개략적인 구성도이다.

도 3은 본 발명의 실시 예에 따른 N차원의 벡터 플로우를 2차원으로 나타내는 도면이다.

도 4는 본 발명의 실시 예에 따른 최적화부의 동작을 나타내는 도면이다.

도 5는 본 발명의 실시 예에 따른 충돌 처리부의 동작을 나타내는 도면이다.

도 6은 본 발명의 실시 예에 따른 통계 시그니쳐 생성 방법을 나타내는 플로우 차트이다.

발명을 실시하기 위한 구체적인 내용

[0016] 아래에서는 첨부한 도면을 참고로 하여 본 발명의 실시 예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시 예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0017] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 "…부", "…기", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0018] 통계 시그니쳐란 패킷의 헤더 정보(패킷 크기, 윈도우 크기 등)나 캡쳐 정보(패킷 캡쳐 시간 등)를 기반으로 하여 다른 응용 프로그램과 구별할 수 있는 응용 프로그램의 고유한 통계적 특징으로, 이를 통해 트래픽을 분류할 수 있다.

[0019] 본 발명에서 트래픽은 호스트와 호스트 사이의 모든 트래픽을 대상으로 하고, 플로우는 동일한 소스 IP(source IP), 목적지 IP(destination IP), 소스 포트(source port), 목적지 포트(destination port) 및 전송 계층 프로토콜(protocol)을 갖는 패킷의 집합으로, 양방향의 플로우로 정의한다. 플로우에서 소스 IP(source IP), 목적지 IP(destination IP), 소스 포트(source port), 목적지 포트(destination port) 및 전송 계층 프로토콜(protocol)은 5-튜플(tuple)이라 한다.

[0020] 특정 네트워크에서 트래픽을 분류하기 위해서는, 우선 네트워크에 연결된 모든 호스트 사이의 트래픽을 수집해야 한다. 예를 들어, 어떤 네트워크가 라우터를 통해 인터넷과 연결되어 있다면 하나의 라우터에서 미러링을 통해 인터넷과 통신하는 모든 트래픽을 수집할 수 있다.

[0021] 2개의 호스트 사이에서 통신이 이루어질 경우, SYN 연결 요청을 하는 호스트가 클라이언트 호스트이고, 도착지 IP가 서버 호스트로서, SYN 연결 요청이 어느 호스트에서 어느 호스트로 송신되느냐에 따라 클라이언트 호스트와 서버 호스트는 수시로 변경될 수 있다. 그러나, 본 발명에서는 하나의 서버 호스트에 2개의 클라이언트 호스트가 연결된 네트워크를 대표하여 통계 시그니쳐를 생성하는 방법을 설명한다.

[0022] 본 발명에서 통계 시그니쳐는 패킷의 페이로드 크기와 방향이다. 패킷의 페이로드의 크기는 페이로드가 존재하는 패킷의 페이로드 크기만을 의미하고, 방향은 양수와 음수로 표현되는데, 양수(+)는 클라이언트에서 서버로 향하는 패킷, 음수(-)는 서버에서 클라이언트로 향하는 패킷을 의미한다. TCP의 경우는 서버/클라이언트의 구분이 명확하지만 UDP는 서버/클라이언트의 구분이 명확하지 않기 때문에, 양수/음수의 의미는 단지 방향이 서로 반대라는 것만 표현할 수 있다. 따라서, UDP의 경우에는 첫 패킷을 양수로 표현하고 뒤에 이어지는 패킷은 첫 패킷을 기준으로 방향이 같으면 양수, 다르면 음수로 표현한다.

[0023] 이제 본 발명의 실시 예에 따른 통계 시그니쳐 생성 장치 및 방법에 대하여 도면을 참고하여 상세하게 설명한다.

[0024] 도 1은 본 발명의 실시 예에 따른 통계 시그니쳐 생성 장치의 개략적인 구성도이다.

[0025] 도 1에 도시된 바와 같이, 통계 시그니쳐 생성 시스템은 제1 및 제2 클라이언트 호스트(110, 120), 서버 호스트(200), 제1 및 제2 티엠에이(TMA, Traffic Measurement Agent)(11, 12), 티엠에스(TMS, Traffic Measurement Server)(20), 패킷 캡쳐부(30), 플로우 생성부(40), 정답지 생성부(50), 시그니쳐 생성부(60)를 포함한다.

[0026] 제1 클라이언트 호스트(110)와 서버 호스트(200) 사이에는 복수의 패킷이 송, 수신되고, 제2 클라이언트 호스트(120)와 서버 호스트(200) 사이에는 복수의 패킷이 송, 수신된다. 플로우는 클라이언트 호스트(110, 120)와 서버 호스트(200) 사이에 송수신되는 패킷들의 집합이다. 도 1에서는 하나의 응용 프로그램을 기준으로 첫 5개의 패킷을 예시적으로 도시하였다.

[0027] 일반적으로 플로우의 초기 몇 개의 패킷은 응용 프로그램에 의해 미리 정해진 정보를 전달하는 용도로 사용된다. 그러므로 플로우의 첫 N개의 패킷에 해당하는 페이로드 크기는 응용 프로그램마다 다르다. 이때, 패킷은 페이로드를 가지고 있는 패킷만을 대상으로 한다. 본 발명에서는 첫 N개의 패킷의 수를 5개로 한정하여 첫 5개 패킷의 페이로드 크기 및 방향을 이용하여 통계 시그니쳐를 생성한다. 그러나 본 발명은 이에 한정되지 않고 첫 N개의 패킷의 개수는 5개보다 많거나 적을 수 있다.

[0028] 도 1에 도시된 바와 같이, 하나의 응용 프로그램을 기준으로 호스트 사이의 전송 패킷 중 첫 5개의 패킷은 1, 2, 3, 4, 5와 같이 전송번호를 가지며, 전송번호 1번 패킷은 페이로드의 크기가 100이며 방향은(+)로서 100으로 표시할 수 있다. 전송번호 2번 패킷은 페이로드의 크기가 100이며 방향은 (-)로서 -100으로 표시이다. 전송번호 3번 패킷은 페이로드의 크기가 200이며 방향은 (+)이다. 전송번호 4번 패킷은 페이로드의 크기가 300이며 방향은(+)이다. 전송번호 5번 패킷은 페이로드의 크기가 200이며 방향은 (-)이다. 상기와 같이, 제2 클라이언트 호스트(120)와 서버 호스트(200) 사이의 패킷들도 페이로드의 크기와 방향을 포함하는 숫자로 표시할 수 있다.

[0029] 패킷 캡쳐부(30)는 클라이언트 호스트(110,120)와 서버 호스트(200) 사이에 연결되어 클라이언트 호스트(110, 120)와 서버 호스트(200) 사이의 복수의 패킷을 캡쳐한다.

[0030] 플로우 생성부(40)는 패킷 캡쳐부(30)에서 캡쳐된 복수의 패킷 중 5-튜플이 동일한 첫 5개의 패킷을 이용하여 복수의 플로우 데이터를 생성한다. 제1 클라이언트 호스트(110)와 제2 클라이언트 호스트(120)는 소스 IP가 다르므로, 5-튜플이 서로 다르다. 따라서, 제1 및 제2 클라이언트 호스트(110, 120)와 서버 호스트(200) 사이에 하나의 응용 프로그램이 존재할 경우, 플로우 생성부(40)에서는 2개의 플로우 데이터가 생성될 수 있다.

[0031] 도 1에 도시된 바와 같이, 플로우 데이터(f1~fn)는 시간, 5-튜플, 패킷 개수, 패킷의 총 바이트, 전송번호 1번 패킷부터 전송번호 5번 패킷까지 각 패킷의 페이로드 크기 및 방향 정보를 포함한다.

[0032] 제1 및 제2 티엠에이(11, 12)는 제1 및 제2 클라이언트 호스트(110, 120)에 각각 설치되어 시간, 5-튜플, 시간에 따른 응용 프로그램 정보를 포함하는 TMA 로그 정보를 수집한다.

[0033] 티엠에스(20)는 제1 및 제2 티엠에이(11, 12)에 연결되어 티엠에이(11, 12)의 TMA 로그 정보(S1~Sn)를 수집한다.

[0034] 도 1에 도시된 바와 같이, 로그 정보(S1~Sn)는 시간, 5-튜플, 응용 프로그램 정보를 포함한다.

[0035] 정답 플로우 생성부(50)는 티엠에스(20) 및 플로우 생성부(40)와 연결되어, 티엠에스(20)의 로그 정보(S1~Sn)와 플로우 생성부(40)의 플로우 데이터(f1~fn) 사이의 시간을 매칭시켜, 정답 플로우(F1~Fn)를 생성한다.

[0036] 도 1에 도시된 바와 같이, 정답 플로우(F1~Fn)는 5-튜플, 패킷 개수, 패킷의 총 바이트, 전송번호 1번 패킷부터 전송번호 5번 패킷까지 각 패킷의 페이로드 크기 및 방향, 응용 프로그램 정보를 포함한다.

[0037] 시그니쳐 생성부(60)는 정답 플로우 생성부(50)에서 생성된 정답 플로우(F1~Fn)를 페이로드 크기와 방향을 이용하여 N차원의 벡터 플로우(V1~Vn)로 변환하고, N차원의 벡터 플로우(V1~Vn) 사이의 시티 블록 거리(city-block distance)가 임계값 미만인 벡터들을 하나의 그룹으로 그룹핑하고, 그룹핑된 하나의 그룹은 하나의 통계 시그니쳐로 생성한다.

[0038] 도 2는 본 발명의 실시 예에 따른 시그니쳐 생성부(60)의 개략적인 구성도이다.

[0039] 도 2에 도시된 바와 같이, 시그니쳐 생성부(60)는 플로우 추출부(61), 벡터 변환부(62), 그룹핑부(63), 최적화부(64), 충돌 처리부(65)를 포함한다.

[0040] 플로우 추출부(61)는 정답 플로우 생성부(50)와 연결되어, 정답 플로우 생성부(50)의 정답 플로우(F1~Fn) 중 하

나의 응용 프로그램에 대해 설정된 제1 기준 개수(a) 이상의 정답 플로우를 추출한다. 이로 인해, 본 발명은 너무 적은 개수의 정답 플로우로 생성된 시그니쳐의 신뢰성 하락을 방지할 수 있으며, 응용 프로그램 내의 분류 가능한 트래픽의 양을 늘릴 수 있다.

[0041] 그리고, 플로우 추출부(61)는 정답 플로우(F1~Fn) 중 하나의 응용 프로그램에 대해 복수의 호스트 각각에서 설정된 제2 기준 개수(b) 이상의 정답 플로우를 추출한다. 이로 인해, 본 발명은 하나의 응용 프로그램에 대한 정답 플로우가 하나의 호스트에서 수집되어 트래픽의 특징이 해당 호스트에 의존적인 특징이 되는 것을 방지할 수 있다.

[0042] 벡터 변환부(62)는 플로우 추출부(61)와 연결되어, 플로우 추출부(61)에서 추출된 정답 플로우의 초기 N개 패킷의 페이로드 크기와 방향을 벡터 플로우로 변환한다. 상기 언급한 바와 같이 본 발명에서는 N을 5로 설정하여, 초기 5개 패킷의 페이로드 크기와 방향을 이용하여 정답 플로우를 벡터 플로우로 변환한다.

[0043] 벡터 플로우를 f라 하고 f의 i번째 페이로드 패킷의 페이로드 크기를 si(f)라 표현할 때, si(f)는 패킷의 방향에 따라 양수와 음수 중 하나의 값을 갖는다. τ가 벡터 플로우를 N차원의 벡터로 표현하는 함수라고 정의하면, 벡터 플로우는 하기 수학식 1과 같이 정의할 수 있다.

수학식 1

[0044] $$\tau(f)=(s_1(f),\ldots,s_N(f))$$

[0045] 여기서, N은 벡터 플로우의 초기 N개 패킷의 개수를 의미하며, 도 1의 정답 플로우(F1~Fn) 중 첫번째 정답 플로우(F1)를 수학식 1의 벡터 플로우로 표현하면, τ(f1)=(100, -100, 200, 300, -200)으로 표현될 수 있다.

[0046] 벡터 변환부(62)는 상기 수학식 1에 의해 정답 플로우(F1~Fn)를 5차원 벡터인 벡터 플로우(τ(f))로 변환한다.

[0047] 그러나, 도 3에서는 벡터 플로우(τ(f))를 설명의 편의상 2차원 벡터로 도시하였다.

[0048] 도 3은 본 발명의 실시 예에 따른 N차원의 벡터 플로우를 2차원으로 나타내는 도면이다.

[0049] 도 3에 도시된 바와 같이, 복수의 벡터 플로우(τ(f))는 복수의 점으로 표시될 수 있다.

[0050] 그룹핑부(63)는 벡터 변환부(62)와 연결되어, 벡터 변환부(62)에서 입력된 5차원의 벡터 플로우(τ(f)) 중 벡터 플로우 사이의 거리가 설정된 거리 임계값 미만인 벡터 플로우들을 하나의 그룹으로 그룹핑한다. 아래에서는 거리 임계값이 50이라 가정하고 설명한다.

[0051] 본 발명에서 그룹핑부(63)는 2개의 벡터 플로우 사이의 거리를 시티 블록 거리(city-block distance)를 이용하여 계산한다. d는 N차원에서 두 벡터 플로우간의 거리라 하고 하기 수학식 2에 의해 계산될 수 있다.

수학식 2

[0052] $$d(\tau(f),\tau(f'))=\sum_{i=1}^{N}|s_i(f)-s_i(f')|$$

[0053] 본 발명에서는 시티 블록 거리를 이용하여 두 벡터 플로우 사이의 거리를 계산함으로써, 많은 양의 데이터를 처리해야 하는 트래픽 분석에서 처리 속도가 빨라지는 장점이 있다.

[0054] 도 3에 도시된 바와 같이, 본 발명의 그룹핑부(63)는 복수의 벡터 플로우 중 시티 블록 거리가 거리 임계값(d1) 미만인 벡터 플로우들을 하나의 그룹으로 그룹핑한다. 그룹핑부(63)에서 그룹핑된 각 그룹은 표 1과 같이 총 5가지의 속성을 가진다.

표 1

[0055]

Atrributes of group	Process code
	Transport protocol
	Representative value
	Dimension of representative value
	Distance threshold

[0056] 프로세스 코드(process code)는 특정 응용 프로그램을 나타내는 것으로 프로그램을 용이하게 하기 위해 임의로 정한 정수이다.

[0057] 전송 프로토콜(transport protocol)은 응용 프로그램이 사용하는 전송계층의 프로토콜로써 TCP, UDP 등이 존재한다.

[0058] 대표값(representative value)는 그룹을 대표하는 값으로 해당 그룹의 모든 벡터 플로우의 벡터를 산술평균을 통해 계산한 벡터로서, 도 3의 2차원 벡터 상에서 다이아몬드 형상인 그룹의 중심점이다.

[0059] 대표값의 차원(dimension of representative value)은 대표값의 차원을 의미하는 것으로, 차원은 곧 페이로드 패킷의 개수를 의미하므로 1 에서 N까지의 값을 가질 수 있다. 본 발명에서는 앞서 언급한 바와 같이 N의 값을 5로 설정하였다.

[0060] 거리 임계값(distance threshold)은 대표값으로부터 가장 먼 시티 블록 거리를 의미하며, 도 3의 2차원 벡터 상에서 대표값을 다이아몬드 형상의 중심점이라 할 때 거리 임계값은 중심점과 다이아몬드 상의 점 사이의 거리 중 가장 먼 거리를 의미한다.

[0061] 상기와 같이 5가지 속성을 갖는 복수의 그룹이 그룹핑부(63)에 의해 생성된 상태에서, 벡터 변환부(62)로부터 새로운 벡터 플로우가 입력되면, 그룹핑부(63)는 새로운 벡터 플로우와 시티 블록 거리가 가장 가까운 제1 그룹을 검색하여 새로운 벡터 플로우를 제1 그룹으로 그룹핑한다. 그리고, 그룹핑부(63)는 새로운 벡터 플로우를 포함하는 제1 그룹의 대표값을 재설정한다. 새로운 벡터 플로우가 그룹에 추가될 때마다 그룹핑부(63)는 대표값을 재설정하므로, 도 3에 도시된 바와 같이 그룹의 대표값은 이동하게 되고 이에 따라 그룹의 다이아몬드 형태도 이동하게 된다.

[0062] 만약, 새로운 벡터 플로우와 제1 그룹의 시티 블록 거리가 설정된 거리 임계값(d1) 이상이면, 그룹핑부(63)는 새로운 제2 그룹을 생성하고 새로운 벡터 플로우를 새로운 제2 그룹으로 그룹핑한다.

[0063] 상기한 바와 같이, 그룹핑부(63)는 복수의 벡터 플로우를 이용하여 복수의 그룹을 형성하고, 새로운 벡터 플로우가 입력될 때마다 새로운 벡터 플로우를 기존의 그룹들과 비교하여 새로운 벡터 플로우를 해당 그룹에 그룹핑한다. 새로운 벡터 플로우에 대응하는 그룹이 존재하지 않을 경우 그룹핑부(63)는 새로운 플로우 벡터를 포함하는 새로운 그룹을 형성한다. 도 3에서 그룹(63-1)은 그룹핑부(63)에서 최종적으로 그룹핑된 하나의 그룹을 나타낸 것이다. 이때, 하나의 그룹(63-1)은 하나의 제1 시그니쳐로서, 그룹핑부(63)에서 형성된 제1 시그니쳐는 대표값, 거리 임계값을 포함한다.

[0064] 이와 같이, 본 발명의 시그니쳐 생성부(60)는 플로우 추출부(61), 벡터 변환부(62) 및 그룹핑부(63)를 이용하여 제1 시그니쳐를 생성할 수 있다.

[0065] 또한, 본 발명의 시그니쳐 생성부(60)는 최적화부(64) 및 충돌 처리부(65)를 통해 제1 시그니쳐보다 정확도가 높은 시그니쳐를 생성할 수 있다.

[0066] 보다 구체적으로, 최적화부(64)는 그룹핑부(63)와 연결되어, 그룹핑부(63)에서 그룹핑된 복수의 그룹을 최적화한다.

[0067] 도 4는 본 발명의 실시 예에 따른 최적화부의 동작을 나타내는 도면으로, 설명의 편의상 그룹핑부(63)에서 그룹핑된 하나의 그룹(63-1)을 대표하여 설명한다.

[0068] 최적화부(64)는 그룹(63-1)에 속하지 않는 벡터 플로우를 제거하고, 대표값(다이아몬드의 중심점)과 가장 멀리 떨어진 벡터 플로우 사이의 거리가 새로운 거리 임계값(d2)이 되도록 그룹핑부(63)의 그룹(63-1)의 최초 거리 임계값(d1)을 새로운 거리 임계값(d2)으로 축소한다. 그러면, 도 4에 도시된 바와 같이 그룹핑부(63)의 그룹(63-1)은 새로운 거리 임계값(d2)를 갖는 최적화 그룹(64-1)으로 축소한다. 최적화 그룹(64-1)은 제1 시그니쳐

보다 정확도가 높은 제2 시그니쳐로서, 제2 시그니쳐 대표값, 새로운 거리 임계값을 포함한다.

[0069] 이와 같이, 본 발명에서는 그룹핑부(63)에서 그룹핑된 복수의 그룹을 최적화하여 그룹핑부(63)의 제1 시그니쳐보다 속성의 범위가 축소된 제2 시그니쳐를 생성할 수 있다.

[0070] 또한, 최적화부(64)는 최적화 그룹(64-1)에 포함되는 모든 벡터 플로우가 동일한 목적지 포트를 가질 경우, 최적화 그룹(64-1)에 포트를 할당한다. 포트가 할당된 하나의 그룹은 제3 시그니쳐로서, 제3 시그니쳐는 대표값, 새로운 거리 임계값 및 목적지 포트 정보를 포함한다.

[0071] 이로 인해, 본 발명은 최적화 그룹(64-1)(제3 시그니쳐)은 패킷의 페이로드 크기 및 방향뿐만 아니라 포트 정보를 포함하여, 좀더 정확하고 빠르게 트래픽을 분류할 수 있는 시그니쳐가 된다.

[0072] 충돌 처리부(65)는 최적화부(64)와 연결되어, 최적화부(64)에서 형성된 최적화 그룹 사이의 충돌 영역을 제거하여 제3 시그니쳐에서 충돌 영역이 제거된 제4 시그니쳐를 생성한다.

[0073] 도 5는 본 발명의 실시 예에 따른 충돌 처리부의 동작을 나타내는 도면이다.

[0074] 도 5에 도시된 바와 같이, 최적화부(64)에서 최적화된 2개의 최적화 그룹(64-1, 64-2) 사이에 충돌 영역(65)이 존재할 경우, 충돌 처리부(65)는 최적화 그룹(64-1, 64-2)에서 각각 충돌 영역(65)을 제거하여 최종 그룹을 형성한다.

[0075] 최종 그룹은 제3 시그니쳐에서 충돌 영역(65)이 제거된 것으로, 최종 그룹은제4 시그니쳐가 된다.

[0076] 이와 같이, 본 발명에서는 패킷의 페이로드 크기 및 방향을 이용하여 시그니쳐를 생성하여, 실시간 트래픽 분류에서 트래픽 분류 시간이 빨라지는 효과가 있다. 또한, 시그니쳐 사이에 충돌이 발생하는 충돌 영역을 제거하여 트래픽이 잘못 분류되는 경우를 방지할 수 있다.

[0077] 도 6은 본 발명의 실시 예에 따른 통계 시그니쳐 생성 방법을 나타내는 플로우 차트이다.

[0078] 도 6에 도시된 바와 같이, 패킷 캡쳐부(30)는 제1 클라이언트 호스트(110)와 서버 호스트(200) 사이의 패킷들 및 제2 클라이언트 호스트(120)와 서버 호스트(200) 사이의 패킷들을 캡쳐한다(S601).

[0079] 플로우 생성부(40)는 패킷 캡쳐부(30)의 패킷들 중 5-튜플이 동일한 초기 N개의 패킷을 이용하여 시간, 5-튜플, 초기 N개 패킷의 페이로드 크기 및 방향을 포함하는 플로우 데이터를 생성한다(S602).

[0080] 패킷 캡쳐부(30)가 패킷들을 캡쳐하는 시간과 동일한 시간에 제1 및 제2 티엠에이(11, 12)는 각각 복수의 클라이언트 호스트(110, 120)에서 시간, 5-튜플, 시간에 따른 응용 프로그램 정보를 포함하는 복수의 로그 정보를 수집하고(S603), 티엠에이(20)는 제1 및 제2 티엠에이(11, 12)에서 수집된 복수의 로그 정보를 수집한다(S604).

[0081] 정답 플로우 생성부(50)는 티엠에스(20)와 플로우 생성부(40) 사이에 연결되어, 복수의 로그 정보의 시간과 플로우 데이터의 시간을 매칭하여 5-튜플, 초기 N개의 패킷의 페이로드 크기 및 방향, 응용 프로그램 정보를 포함하는 복수의 정답 플로우를 생성한다(S605).

[0082] 시그니쳐 생성부(60)는 정답 플로우 생성부(50)와 연결되어, 복수의 정답 플로우의 페이로드 크기와 방향을 이용하여 복수의 정답 플로우를 N차원의 복수의 벡터 플로우로 변환하고(S606), 복수의 벡터 플로우 사이의 거리가 설정된 거리 임계값 미만인 벡터 플로우들을 하나의 그룹으로 그룹핑한다(S607).

[0083] 여기서, 하나의 그룹은 하나의 시그니쳐가 된다.

[0084] 단계(S607)에 이어, 시그니쳐 생성부(60)의 최적화부(64)는 복수의 그룹에 포함되는 복수의 벡터 플로우를 제외한 벡터 플로우를 제거하고, 그룹의 거리 임계값(d1)이 대표값으로부터 가장 멀리 떨어진 벡터 플로우와 대표값 사이의 새로운 거리 임계값(d2)이 되게 거리 임계값(d1)을 새로운 거리 임계값(d2)으로 보정하여, 복수의 그룹을 최적화시킨다(S608).

[0085] 이때, 단계(S608)에서 최적화부(64)는 하나의 그룹에 포함되는 복수의 벡터 플로우가 동일한 목적지 포트를 가질 경우, 하나의 그룹에 목적지 포트를 할당할 수도 있다.

[0086] 충돌 처리부(65)는 상기와 같이 최적화된 복수의 그룹 사이의 충돌 영역을 추출하고, 복수의 그룹에서 충돌 영역을 제거하여 최종 그룹을 형성한다(S609)

[0087] 이와 같이, 본 발명에서는 패킷의 페이로드 크기 및 방향을 이용하여 시그니쳐를 생성하여, 실시간 트래픽 분류에서 트래픽 분류 시간이 빨라지는 효과가 있다. 또한, 시그니쳐 사이에 충돌이 발생하는 충돌 영역을 제거하여

트래픽이 잘못 분류되는 경우를 방지할 수 있다.

[0088] 이상에서 본 발명의 실시 예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

부호의 설명

[0089] 110, 120: 클라이언트 호스트

200: 서버 호스트

11, 12: 티엠에이(TMA)

20: 티엠에스(TMS)

30: 패킷 캡쳐부

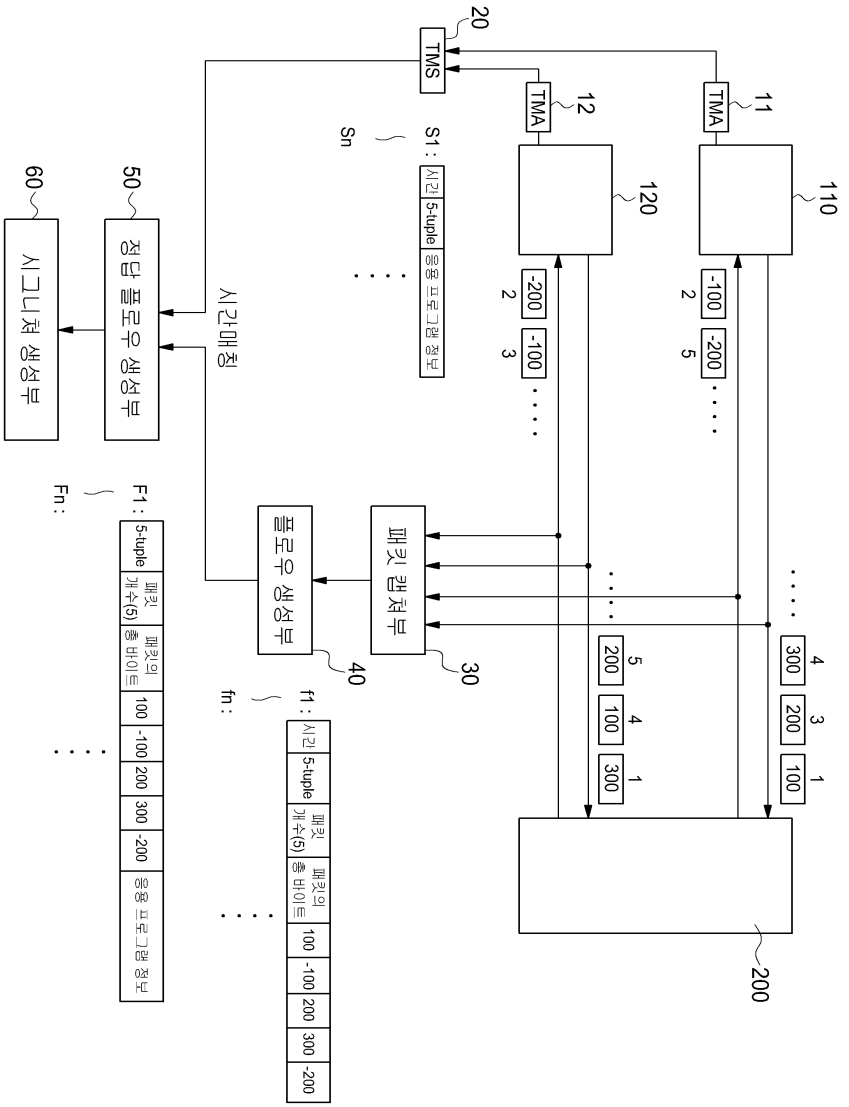
40: 플로우 생성부

50: 정답지 생성부

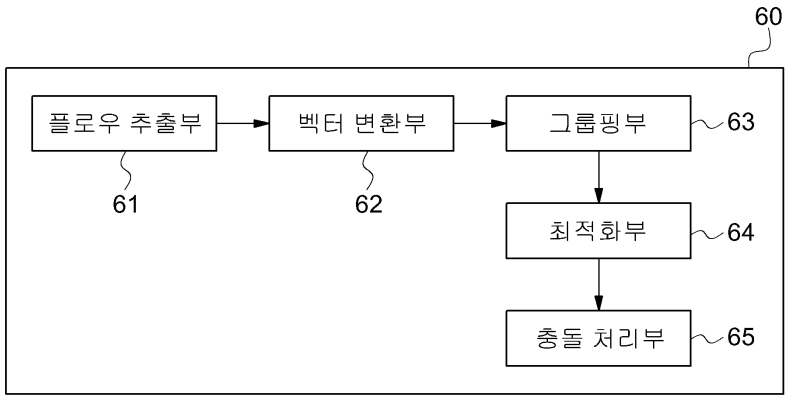
60: 시그니쳐 생성부

도면

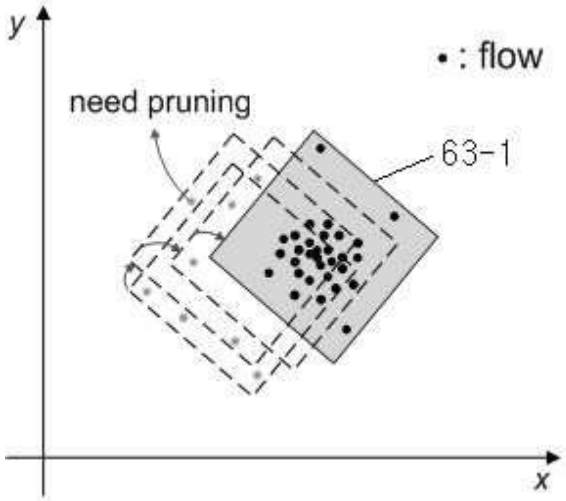
도면1


TMS
TMA
TMA
패킷 캡쳐부
플로우 생성부
정답 플로우 생성부
시그니쳐 생성부
시간매칭
S1 :
Sn
시간
5-tuple
응용 프로그램 정보
F1 :
Fn :
5-tuple
패킷 개수(5)
패킷의 총 바이트
f1 :
fn :
110
120
200

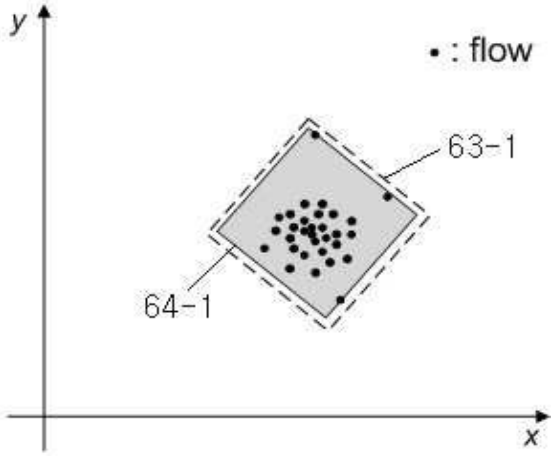
도면2


60
플로우 추출부
61
빅딘 변환부
62
그룹핑부
63
최적화부
64
충돌 처리부
65

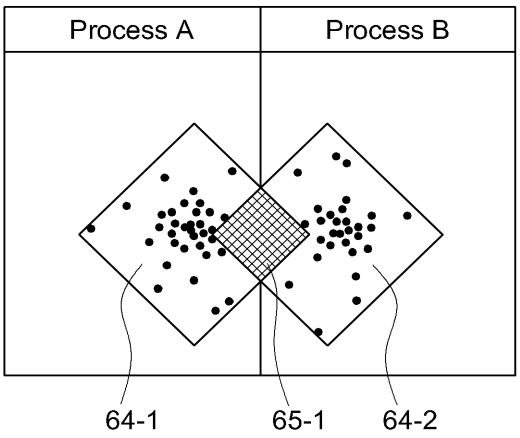
도면3


y
• : flow
need pruning
63-1
x

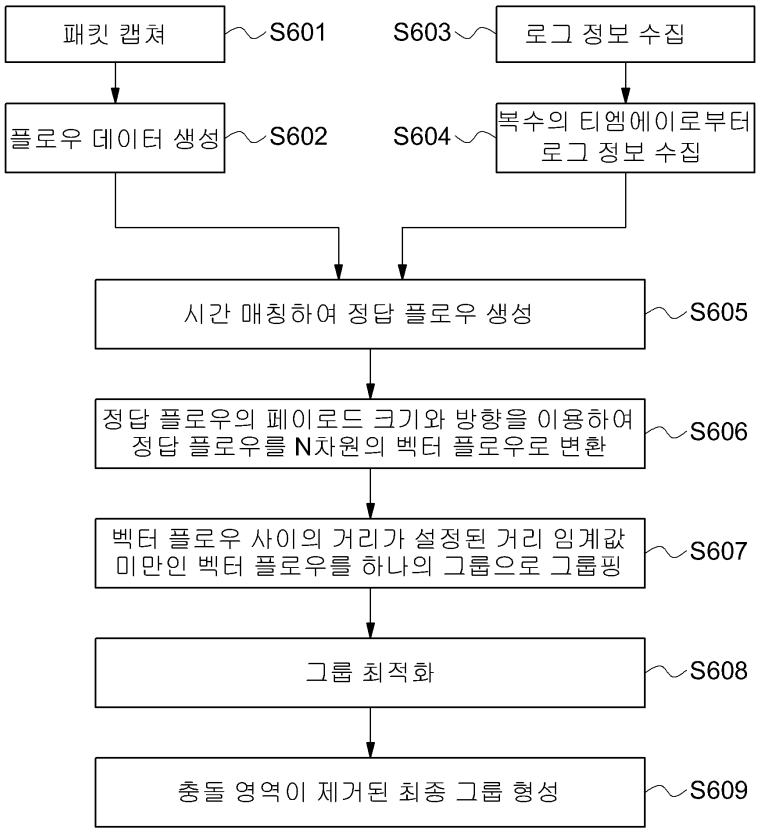
도면4


y
• : flow
63-1
64-1
x

도면5


Process A
Process B
64-1
65-1
64-2

도면6


패킷 캡쳐
S601
플로우 데이터 생성
S602
S603
로그 정보 수집
S604
복수의 티엠에이로부터 로그 정보 수집
시간 매칭하여 정답 플로우 생성
S605
정답 플로우의 페이로드 크기와 방향을 이용하여 정답 플로우를 N차원의 벡터 플로우로 변환
S606
벡터 플로우 사이의 거리가 설정된 거리 임계값 미만인 벡터 플로우를 하나의 그룹으로 그룹핑
S607
그룹 최적화
S608
충돌 영역이 제거된 최종 그룹 형성
S609