(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2011-0041149
(43) 공개일자 2011년04월21일

(51) Int. Cl.
H04L 12/26 (2006.01)
(21) 출원번호 10-2009-0098199
(22) 출원일자 2009년10월15일
심사청구일자 2009년10월15일

(71) 출원인
고려대학교 산학협력단
서울 성북구 안암동5가 1
(72) 발명자
강철희
서울특별시 중랑구 묵1동 신내대림아파트 513-503
박현희
서울특별시 강북구 번1동 465-4 밀레니엄 오피스텔 904호
김미정
서울특별시 은평구 불광동 245-68
(74) 대리인
나승택, 조영현

전체 청구항 수 : 총 5 항

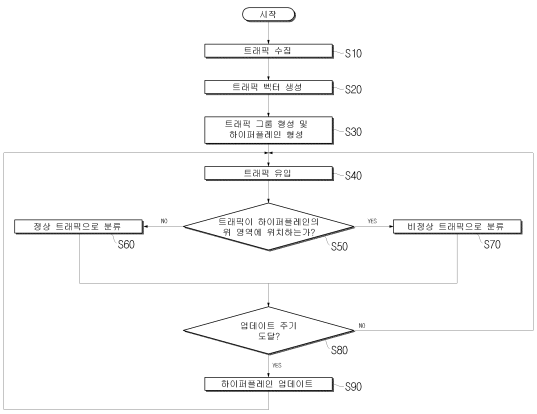
(54) 피셔 선형 분류법을 이용한 비정상 트래픽 탐지 방법

(57) 요 약

본 발명은 서브 네트워크에서의 비정상 트래픽을 탐지하는 방법에 관한 것으로, 특히 수집 기간 동안의 트래픽들을 데이터로 하여, 카이스퀘어 분포(chi-square distribution)와 피셔(Fisher) 선형 분류법을 이용하여 정상 그룹 및 비정상 그룹으로 분류할 수 있는 하이퍼플레인을 형성한 후, 이후에 유입되는 트래픽이 정상 그룹에 속하는지 또는 비정상 그룹에 속하는지를 용이하게 판별할 수 있도록 하는 피셔 선형 분류법을 이용한 비정상 트래픽 탐지 방법에 관한 것이다.

본 발명인 비정상 트래픽 탐지 방법을 이루는 구성수단은, 라우터로부터 특정 수집 기간 동안 특정 주기로 트래픽을 수집한 후, 수집된 각 트래픽의 특성을 나타내는 복수개의 성분으로 구성된 트래픽 벡터들을 생성하고, 상기 트래픽 벡터들을 데이터로 하여 피셔(Fisher) 선형 분류법을 이용하여 정상 그룹과 비정상 그룹으로 분류할 수 있는 하이퍼플레인(hyperplane)을 형성하는 제1 단계, 상기 하이퍼 플레인을 형성한 후에, 유입되는 트래픽이 하이퍼플레인의 면을 포함한 아래 영역에 위치하는 경우에는 정상 그룹으로 판단하고, 하이퍼플레인의 위 영역에 위치하는 경우에는 비정상 그룹으로 판단하는 제2 단계, 상기 판단 후, 사전에 세팅된 업데이트 주기에 도달했는지 판단하는 제3 단계, 상기 판단 결과, 업데이트 주기에 도달한 경우에는, 업데이트 주기에 해당하는 기간 동안 수집된 가장 오래된 트래픽 벡터들을 버리고, 업데이트 주기 동안 새롭게 생성된 트래픽 벡터들을 포함시켜 새로운 하이퍼플레인을 형성한 후, 상기 제2 단계로 귀환하는 제4 단계를 포함하여 이루어진 것을 특징으로 한다.

대 표 도 - 도2



특허청구의 범위

청구항 1

라우터로부터 특정 수집 기간 동안 특정 주기로 트래픽을 수집한 후, 수집된 각 트래픽의 특성을 나타내는 복수개의 성분으로 구성된 트래픽 벡터들을 생성하고, 상기 트래픽 벡터들을 데이터로 하여 피셔(Fisher) 선형 분류법을 이용하여 정상 그룹과 비정상 그룹으로 분류할 수 있는 하이퍼플레인(hyperplane)을 형성하는 제1 단계;

상기 하이퍼 플레인을 형성한 후에, 유입되는 트래픽이 하이퍼플레인의 면을 포함한 아래 영역에 위치하는 경우에는 정상 그룹으로 판단하고, 하이퍼플레인의 위 영역에 위치하는 경우에는 비정상 그룹으로 판단하는 제2 단계;

상기 판단 후, 사전에 세팅된 업데이트 주기에 도달했는지 판단하는 제3 단계;

상기 판단 결과, 업데이트 주기에 도달한 경우에는, 업데이트 주기에 해당하는 기간 동안 수집된 가장 오래된 트래픽 벡터들을 버리고, 업데이트 주기 동안 새롭게 생성된 트래픽 벡터들을 포함시켜 새로운 하이퍼플레인을 형성한 후, 상기 제2 단계로 귀환하는 제4 단계를 포함하여 이루어진 것을 특징으로 하는 피셔 선형 분류법을 이용한 비정상 트래픽 탐지 방법.

청구항 2

청구항 1에 있어서,

상기 각 트래픽 벡터의 성분은 대역폭(bandwidth)의 용량, 플로우(flow)의 용량 및 패킷(packet)의 용량인 것을 특징으로 하는 피셔 선형 분류법을 이용한 비정상 트래픽 탐지 방법.

청구항 3

청구항 2에 있어서,

상기 트래픽 벡터의 각 성분에 대한 상기 수집 기간 동안의 분포는 카이스퀘어 분포(chi-square distribution)를 따르고, 각 성분의 용량이 하나라도 상위 4% ~ 6%에 해당하는 경우에는 비정상 그룹으로 분류하고, 모든 성분의 용량이 상위 4% ~ 6%에 해당하지 않는 경우에는 정상 그룹으로 분류하는 것을 특징으로 하는 피셔 선형 분류법을 이용한 비정상 트래픽 탐지 방법.

청구항 4

청구항 3에 있어서,

상기 카이스퀘어 분포는 자유도가 상기 수집 기간인 것을 특징으로 하는 피셔 선형 분류법을 이용한 비정상 트래픽 탐지 방법.

청구항 5

청구항 1에 있어서,

상기 하이퍼플레인은 법선벡터에 수직이고, 상기 정상 그룹 및 비정상 그룹을 포함한 전체 평균점을 지나는 평면인 것을 특징으로 하는 피셔 선형 분류법을 이용한 비정상 트래픽 탐지 방법.

명 세 서

발명의 상세한 설명

기 술 분 야

[0001] 본 발명은 서브 네트워크에서의 비정상 트래픽을 탐지하는 방법에 관한 것으로, 특히 수집 기간 동안의 트래픽들을 데이터로 하여, 카이스퀘어 분포(chi-square distribution)와 피셔(Fisher) 선형 분류법을 이용하여 정상 그룹 및 비정상 그룹으로 분류할 수 있는 하이퍼플레인을 형성한 후, 이후에 유입되는 트래픽이 정상 그룹에 속하는지 또는 비정상 그룹에 속하는지를 용이하게 판별할 수 있도록 하는 피셔 선형 분류법을 이용한 비정상 트래픽 탐지 방법에 관한 것이다.

배 경 기 술

[0002] 오늘날의 통신 네트워크는 전통적인 인터넷 응용 프로그램뿐만 아니라 멀티미디어 스트리밍, P2P, 실시간 음성 및 화상 통신, 네트워크 게임 등 인터넷 기반 응용 프로그램들이 증가하면서 네트워크의 트래픽들이 복잡화되고 다양화되어 이에 대한 트래픽 모니터링과 분석의 중요성이 대두되고 있다.

[0003] 이렇게 복잡화되고 다양화된 네트워크를 통하여 비정상 트래픽의 발생과 각종 침해사고는 점점 증가하고 있으며 이로 인한 피해가 지속적으로 발생하고 있다. 비정상 트래픽은 네트워크 트래픽 폭주, DoS 공격, 웜과 같은 다양한 공격들의 원인이 되며 지능적이고 복합적인 형태로 발생되고 있기 때문에 이러한 이상 징후를 조기에 탐지할 수 있는 보다 진보된 기술이 요구되고 있다.

[0004] 따라서 기존의 많은 연구들이 비정상 트래픽을 탐지하기 위한 방법을 다루어 왔다. 대표적인 두 가지 방법으로는 규칙(rule) 기반의 방식과 측정(measurement) 기반의 방식이 있다.

[0005] 먼저 규칙(rule) 기반의 방식은 비정상 트래픽의 패턴을 미리 파악해서 새롭게 발생하는 트래픽의 패턴은 미리 파악한 패턴을 비교해서 검출하는 방식이다.

[0006] 패턴 매칭과 같은 이러한 접근 방식은 알려진 공격에 대해서 높은 탐지율을 보인다. 그러나 알려지지 않은 새로운 공격에 대해서는 비교적 낮은 탐지율을 보이며, 패킷을 일일이 비교해야 하는 문제 때문에 백본과 같은 고속의 대용량 네트워크에서는 접합하지 않다는 단점이 있다.

[0007] 이러한 단점을 극복하기 위하여 측정(measurement) 기반의 방식에서는 인터넷 백본 트래픽이 주기성을 띈다는 점을 전제로 하여, 네트워크 상에서 얻을 수 있는 간단한 트래픽 정보를 바탕으로 트래픽을 모델링함으로써 미래의 트래픽을 예측하는 방법과 많은 트래픽 정보 중 유용한 특정 정보들만을 샘플링하여 모델링하는 방법들이 제안되었다.

[0008] 그러나 모델링을 통해 탐지하는 방법은 정상의 여부만을 구분할 뿐 비정상의 종류를 알기 위해 여러 단계를 거쳐야 한다는 단점이 있다.

[0009] 한편, 비정상 트래픽의 탐지는 트래픽 모니터링이 중요한 부분을 차지하고 있기 때문에 트래픽의 특징을 잘 표현할 수 있는 파라미터들을 정의하고 트래픽의 분석 시스템을 개발하는 것이 중요하다.

발명의 내용

해결 하고자하는 과제

[0010] 본 발명은 상기와 같은 종래 기술의 문제점들을 해결하기 위하여 창안된 것으로, 수집 기간 동안의 트래픽들을 데이터로 하여, 카이스퀘어 분포(chi-square distribution)와 피셔(Fisher) 선형 분류법을 이용하여 정상 그룹 및 비정상 그룹으로 분류할 수 있는 하이퍼플레인을 형성한 후, 이후에 유입되는 트래픽이 정상 그룹에 속하는지 또는 비정상 그룹에 속하는지를 용이하게 판별할 수 있도록 하는 피셔 선형 분류법을 이용한 비정상 트래픽 탐지 방법을 제공하는 것을 그 목적으로 한다.

과제 해결수단

[0011] 상기와 같은 과제를 해결하기 위하여 제안된 본 발명인 비정상 트래픽 탐지 방법을 이루는 구성수단은, 라우터

로부터 특정 수집 기간 동안 특정 주기로 트래픽을 수집한 후, 수집된 각 트래픽의 특성을 나타내는 복수개의 성분으로 구성된 트래픽 벡터들을 생성하고, 상기 트래픽 벡터들을 데이터로 하여 피셔(Fisher) 선형 분류법을 이용하여 정상 그룹과 비정상 그룹으로 분류할 수 있는 하이퍼플레인(hyperplane)을 형성하는 제1 단계, 상기 하이퍼 플레인을 형성한 후에, 유입되는 트래픽이 하이퍼플레인의 면을 포함한 아래 영역에 위치하는 경우에는 정상 그룹으로 판단하고, 하이퍼플레인의 위 영역에 위치하는 경우에는 비정상 그룹으로 판단하는 제2 단계, 상기 판단 후, 사전에 세팅된 업데이트 주기에 도달했는지 판단하는 제3 단계, 상기 판단 결과, 업데이트 주기에 도달한 경우에는, 업데이트 주기에 해당하는 기간 동안 수집된 가장 오래된 트래픽 벡터들을 버리고, 업데이트 주기 동안 새롭게 생성된 트래픽 벡터들을 포함시켜 새로운 하이퍼플레인을 형성한 후, 상기 제2 단계로 귀환하는 제4 단계를 포함하여 이루어진 것을 특징으로 한다.

[0012] 여기서, 상기 각 트래픽 벡터의 성분은 대역폭(bandwidth)의 용량, 플로우(flow)의 용량 및 패킷(packet)의 용량인 것을 특징으로 한다.

[0013] 여기서, 상기 트래픽 벡터의 각 성분에 대한 상기 수집 기간 동안의 분포는 카이스퀘어 분포(chi-square distribution)를 따르고, 각 성분의 용량이 하나라도 상위 4% ~ 6%에 해당하는 경우에는 비정상 그룹으로 분류하고, 모든 성분의 용량이 상위 4% ~ 6%에 해당하지 않는 경우에는 정상 그룹으로 분류하는 것을 특징으로 한다.

효 과

[0014] 상기와 같은 과제 및 해결 수단을 가지는 본 발명인 서브 네트워크에서의 피셔 선형 분류법을 이용한 비정상 트래픽을 탐지하는 방법에 의하면, 수집 기간 동안의 트래픽들을 데이터로 하여, 카이스퀘어 분포(chi-square distribution)와 피셔(Fisher) 선형 분류법을 이용하여 정상 그룹 및 비정상 그룹으로 분류할 수 있는 하이퍼플레인을 형성한 후, 이후에 유입되는 트래픽이 상기 하이퍼플레인의 아래 영역에 해당되면 정상 그룹으로 분류하고, 하이퍼플레인의 위 영역에 해당되면 비정상 그룹으로 분류할 수 있기 때문에, 유입되는 트래픽 중, 비정상 트래픽을 용이하게 판별할 수 있는 장점이 있다.

[0015] 또한, 하이퍼플레인을 주기적으로 업데트하기 때문에, 변화되는 트래픽의 추세에 부합할 수 있는 장점이 있다.

발명의 실시를 위한 구체적인 내용

[0016] 이하, 첨부된 도면을 참조하여 상기와 같은 과제, 해결수단 및 효과를 가지는 본 발명인 피셔 선형 분류법을 이용한 비정상 트래픽 탐지 방법에 관한 바람직한 실시예를 상세하게 설명한다.

[0017] 네트워크 관리자의 입장에서 사용자에게 효율적이고, 정상적인 서비스를 제공하기 위해서는 유해 트래픽뿐 아니라, 악의성은 없지만 네트워크 사용자가 정상적인 서비스를 제공받는데 문제가 되는 비정상 트래픽들에 대한 탐지가 이루어져야 한다.

[0018] 또한 각각의 비정상 트래픽들은 네트워크의 특성에 따라서 정상 트래픽으로 분류되는 경우도 있을 수 있다. 예를 들어, 어떤 네트워크 안에서는 ftp 트래픽으로 인해 다른 정보 전송에 문제를 야기한다면 해당하는 ftp 트래픽은 비정상 트래픽으로 분류되어야 하고, 네트워크의 효율 측면에서 일정 수준의 ftp 트래픽만을 허용할 수 있다면 그보다 더 많은 양의 ftp 트래픽에 대해서는 비정상 트래픽으로 탐지를 해야만 제어할 수 있을 것이다.

[0019] 이러한 관점에서 네트워크 트래픽들이 가진 특성에 따라 분류하고 유입되는 트래픽이 어떠한 트래픽의 그룹에 속하는지 분류를 하는 작업이 필요하다.

[0020] 본 발명에서 제안하는 방법은 네트워크 상에서 관측되는 트래픽들이 지니고 있는 특성들을 성분으로 하는 데이터의 집합인 트래픽 벡터를 정의한 후 트래픽 벡터를 사용하여 트래픽을 그 특성에 따라 그룹을 짓고 그 그룹이 잘 구분이 되는지 판별한 후에 유입되는 트래픽이 어떠한 그룹에 속하게 되는지 탐지하는 방법이다.

[0021] 본 발명인 피셔(Fisher) 선형 분류법을 이용한 비정상 트래픽 탐지 방법에 대하여 구체적으로 설명하기 전에, 본 발명에 적용되는 전체 시스템에 대하여 도 1을 참조하여 간략하게 설명한다.

[0022] 본 발명인 피셔 선형 분류법을 이용한 비정상 트래픽 탐지 방법을 구현하기 위하여 도 1에 도시된 시스템이 구비된다. 즉, 트래픽 모니터링 수단(10), 분석 수단(20), 탐지 수단(30) 및 제어 수단(40)으로 구성된 시스템을

통하여 본 발명이 구현된다.

[0023] 상기 트래픽 모니터링 수단(10)은 크게 트래픽 캡쳐부(15), 트래픽 샘플링부(17) 및 트래픽 저장 및 모니터링부(19)를 포함하여 구성되고, 상기 분석 수단(20)은 저장부(21), 트래픽 분류부(23) 및 수학적 분석부(25)를 포함하여 구성된다.

[0024] 트래픽 로(raw) 데이터는 네트워크(11) 상에 연결되는 라우터(13)를 통하여 상기 트래픽 캡쳐부(15)에 의하여 수집될 수 있다. 상기 트래픽 로(raw) 데이터는 Netflow v5의 형식으로 구성한다. Netflow는 기본적으로 단일 방향의 특성을 가지며 출발지 및 도착지 주소와 포트번호, 프로토콜 등 5개의 튜플(tuple)로 구성된다. 그 외에도 TOS 및 인터페이스 정보, AS 정보를 포함한다.

[0025] 상기 Netflow는 이러한 정보를 규칙으로 적용하여 다양한 프로토콜 플로우를 생성하고 구분할 수 있다. 이는 처리 속도가 빠르고 트래픽의 흐름과 동시에 라우터에서 생성되어 네트워크 관리자가 그 정보를 즉시 받아볼 수 있다는 장점이 있다.

[0026] 상기 트래픽 로(raw) 데이터는 트래픽 샘플링부(17)에 의하여 소정의 샘플링 과정을 거쳐 트래픽 저장 및 모니터링부(19)에 의하여 저장되고 모니터링된다.

[0027] 상기와 같이 트래픽 저장 및 모니터링부(19)에 저장된 트래픽들은 분석 수단(20)의 저장부(21)로 소정의 가공을 거쳐 저장된다. 즉, 상기 저장부(21)는 상기 트래픽 저장 및 모니터링부(19)에 저장된 트래픽을 특정 주기마다 가공하여 저장한다.

[0028] 물론 이와 같은 가공 과정은 제어수단(40)의 제어에 따라 진행되고, 가공된 트래픽은 특정 주기마다 엑셀 파일 형식으로 상기 저장부(21)에 저장된다. 그러면, 상기 저장부(21)에 수집되어 저장된 트래픽의 특성에 따라 트래픽 벡터의 성분을 결정하고 각 트래픽 벡터의 공통된 특성에 따라 트래픽 그룹을 결정하게 된다. 본 발명에서는 트래픽 그룹의 기준을 결정하기 위하여 MS-SQL 쿼리를 이용하여 상기 분류부(23)가 두개의 트래픽 그룹(정상 그룹 및 비정상 그룹)을 만든다.

[0029] 그리고, 상기 수학적 분석부(25)는 상기 그룹의 정확한 기준을 만들기 위하여 카이스퀘어 분포를 적용하고, 피셔(Fisher) 선형 분류법을 통해 하이퍼플레인(hyperplane)을 형성한다.

[0030] 상기 하이퍼플레인이 형성된 후에는, 정상 또는 비정상으로 분류될 수 있는 트래픽들을 유입한다. 그러면, 탐지수단(30)이 상기 하이퍼플레인을 이용하여 유입되는 트래픽이 정상 그룹에 포함되는지 또는 비정상 그룹에 포함되는지 판단하고, 비정상 트래픽들을 탐지한다.

[0031] 한편, 트래픽의 추이는 지속적으로 변화기 때문에, 새롭게 발생하는 트래픽의 추이를 반영하기 위해서 하이퍼플레인을 주기적으로 상기 제어수단(40)의 제어에 따라 상기 수학적 분석부(25)가 업데이트한다. 즉, 업데이트 주기를 P라고 했을 때, 가장 오래된 P 기간의 트래픽은 P기간 동안 새롭게 들어온 트래픽으로 교체되어, 새로운 하이퍼플레인을 형성시킨다.

[0032] 상기와 같은 시스템을 통하여 본 발명인 피셔(Fisher) 선형 분류법을 이용한 비정상 트래픽 감지 방법이 구현된다.

[0033] 다음은 본 발명에서 중요하게 적용되는 피셔 선형 분류법과 카이스퀘어(chi-square) 분포에 관하여 설명한다.

[0034] 피셔 선형 분류법은 그룹 분류가 가능한 선형 함수로서 다차원 데이터를 하이퍼플레인(hyperplane)의 법선 벡터에 사영시켜 그룹을 분류하는 기법이다([A. Shashua, "On the Relationship Between the Support Vector Machine for Classification and Sparsified Fisher's Linear Discriminant," Neural Processing Letters, vol. 9, pp. 129-139, April 1999.], [R. Johnson, and D.Wichern, Applied Multivariate Statistical Analysis, 6th ed., Prentice-Hall, 2007, pp. 576-593, 623-633.]).

[0035] $X=\{\mathbb{X}_i=(x_{i1},....x_{in}):\mathbb{X}_i\in\mathbb{R}^n,i=1,...,k\}$ 를 n개의 성분으로 이루어진 벡터로 구성된 데이터 집합이라 하자. 이 데이터 집합은 m 종류의 데이터로 그룹으로 분류된다고 가정하고, j 그룹의 데이터 집합을 Y_j로 표기한다.

[0036] 즉, $Y = \{Y_j\}^m_{j=1}$는 $\sum_{j=1}^{m}|Y_j|=k$ 를 만족하는 X의 디스조인트(disjoint) 서브셋(subset)이다. 여기서 $|Y_j|$는 Y_j

집합의 원소의 개수를 나타낸다. u와 uj는 각각 X와 Yj의 평균이라 하자. 여기서 $u=(u_1,....u_n)^t$와 $u_j=(u_{j1},....u_{jn})^t$의 성분은 각각 $\mu_l = \frac{1}{n}\sum_{i=1}^{n} x_{il}$ 와 $\mu_{jl} = \frac{1}{|Y_j|}\sum_{X_i \in Y_j} x_{il}$ 로 주어진다. 여기서 t는 행렬의 전치(transpose)를 나타낸 것이다.

[0037] 이제 법선 벡터(direction vector)를 소개하고 데이터 그룹 내의 산란행렬(scatter matrix)과 그룹 간의 산란행렬을 정의한다. 본 발명에서는 각 그룹의 평균 u_j를 어떤 법선 벡터 w에 사영시켰을 때 각 그룹이 분리가 잘 될 수 있도록 최대의 분산을 갖게 하는 하이퍼플레인(hyperplane) $W^t\mathbb{X}+d=0$ (여기서 $\mathbb{X}=(x_1,...,x_n)$ 이고, d는 원점으로부터 하이퍼플레인까지의 거리)과 수직인 법선 벡터 w를 찾는 것이 목적이다.

[0038] 그룹 내의 산란 행렬과 그룹 간의 산란 행렬을 각각 $\mathbb{S}_{intra}$ 와 $\mathbb{S}_{inter}$로 나타낸다. 이 행렬들을 정의하기 위하여 $\mathbb{S}_j$ 를 집합 j의 산란 행렬로 다음과 같이 정의한다.

[0039] $$\mathbb{S}_j = \sum_{\mathbb{X}_i \in Y_j} (\mathbb{X}_i - \mu_j)(\mathbb{X}_i - \mu_j)^t \qquad \text{수식 (1)}$$

[0040] 이를 이용하여 $\mathbb{S}_{intra}$ 와 $\mathbb{S}_{inter}$는 다음과 같이 정의한다.

[0041] $$\mathbb{S}_{intra} = \sum_{j=1}^{m} \mathbb{S}_j \quad , \quad \mathbb{S}_{inter} = \sum_{j=1}^{m} (\mu_j - \mu)(\mu_j - \mu)^t$$

[0042] 그러므로 각 그룹의 사영된 중심으로부터 그 그룹의 사영된 원소들 간의 정규화된 분산의 합은 다음과 같이 나타난다.

[0043] $$\frac{1}{\| W\|} W^t \mathbb{S}_{intra} W \quad , \quad \frac{1}{\| W\|} W^t \mathbb{S}_{inter} W \qquad \text{수식 (2)}$$

[0044] 여기서 ‖w‖는 유클리드 놈(Euclidean norm)을 나타내고, $W^t\mathbb{S}_{inter}W$ 을 계산하면 $\sum_{j=1}^{m} (W^t(\mu_j - \mu))^2$ 이 된다.

[0045] 그룹을 잘 분리하기 위하여 그룹 내의 데이터는 최대한 뭉쳐있어야 하고, 그룹간의 데이터는 최대한 멀리 떨어져 있어야 한다. 즉, 수식 (2)의 값이 크다는 것은 그룹간의 구별이 명확하다는 것을 의미한다. 그러므로 수식 (1)과 수식 (2)를 이용하여 최적화 정식을 통한 전형적인 피셔 표준(classical Fisher criterion) 함수를 다음과 같이 정의할 수 있다.

[0046] $$(OPT)_{W \in D}^{\max} OPT(W) = \max_{W \in D} \frac{W^t \mathbb{S}_{inter} W}{W^t \mathbb{S}_{intra} W}$$

[0047] 여기서 $D=\{W: W^t\mathbb{X}+d=0, \mathbb{X}\in X\}$ 이다.

[0048] 최적화 정식 (OPT)의 해는 분모를 최소화하고 분자를 최대화하여 얻을 수 있으므로, 각 그룹 내의 트래픽들은 동일한 특성들로 구성되어 있고 그룹간의 트래픽들은 서로 상이한 특성을 나타낼 수 있어야 한다. 최적화 정식은 다음과 같은 고유치 문제(eigenvalue problem)와 동치이다.

[0049] $$\mathbb{S}_{inter}W=\lambda\mathbb{S}_{intra}W$$ 수식 (3)

[0050] 또한 수식 (3)의 최대 고유값에 대응하는 고유벡터가 구하는 법선 벡터 w이다([A. Shashua, "On the Relationship Between the Support Vector Machine for Classification and Sparsified Fisher's Linear Discriminant," Neural Processing Letters, vol. 9, pp. 129-139, April 1999.]).

[0051] 한편, 후술하겠지만, 본 발명에서 수집된 매일의 트래픽의 양은 크게 정규분포(nomal distribution)를 따름을 관찰할 수 있었으므로, 이에 따라 30일 이상 수집한 트래픽 데이터는 다음 정리에 의하여 카이스퀘어 분포를 따른다고 가정해도 크게 문제가 없다. 다음 정리는 표준정규분포와 카이스퀘어 분포와의 관계를 나타낸다.

[0052] 정리 1 : 확률변수 Z1,.....Zk가 표준정규분포 N(0,1)을 따르고 서로 독립이면 $Z^2_1+Z^2_2+.....+Z^2_k$는 자유도 (degrees of freedom) k인 카이스퀘어 분포를 따른다. 이것은 $Z^2_1+Z^2_2+....+Z^2_k \sim \chi^2(k)$로 나타낸다([Z. Zhang and H. Shen "Online Training of SVMs for Real-time Intrusion Detection," in Proc. AINA 2004, pp. 568-573, March 2004.], [T. Hamada, K. Chujo, T. Chujo, and X. Yang, "Peer-to-peer traffic in metro networks: analysis, modeling, and policies," in Proc. IEEE/IETF NOMS 2004, pp. 425-438, April 2004.])

[0053] 이상에서 설명한 피셔 선형 분류법 및 카이스퀘어 분포와 도 1을 참조하여 설명한 시스템에 기반하여 비정상 트래픽을 탐지하는 방법에 대하여 첨부된 도 2를 참조하여 설명한다.

[0054] 비정상 트래픽을 탐지하기 위해서, 본 발명은 유입되는 트래픽이 정상 트래픽인지, 비정상 트래픽인지를 용이하게 판단하기 위한 기준면이 되는 하이퍼플레인(hyperplane)을 카이스퀘어 분포와 피셔 선형 분류법을 이용하여 형성한다.

[0055] 상기와 같이 하이퍼플레인을 형성하기 위하여, 로(raw) 데이터로 사용되는 트래픽들을 라우터를 통하여 수집한다(S10). 라우터로부터 로(raw) 데이터로 사용되는 트래픽들은 특정 수집 기간 동안 수집된다. 후술하겠지만, 본 발명의 실시예에서는 30일 동안 트래픽들을 수집한다.

[0056] 상기 수집되는 트래픽들은 소정의 특정 주기로 수집된다. 예를 들어서, 5분마다 한번씩 트래픽을 수집하여 저장한다. 따라서, 상기 수집되는 트래픽들은 특정 수집 기간 동안 특정 주기로 수집된다. 예를 들어, 30일 동안 5분마다 트래픽들을 수집하여 저장한다.

[0057] 상기 수집된 트래픽들을 이용하여 특정 정보를 갖는 트래픽 벡터를 생성한다(S20). 즉, 트래픽을 수집한 후, 수집된 각각의 트래픽의 특성을 나타내는 복수개의 성분으로 구성된 트래픽 벡터들을 생성한다.

[0058] 상기 각 트래픽 벡터의 성분은 대역폭(bandwidth)의 용량, 플로우(flow)의 용량 및 패킷(packet)의 용량을 포함한다. 즉, 각각의 트래픽 벡터는 해당 트래픽에 대한 정보인 대역폭(bandwidth)의 용량, 플로우(flow)의 용량 및 패킷(packet)의 용량을 나타낸다.

[0059] 상기와 같이 트래픽 벡터들이 생성되면, 상기 트래픽 벡터들을 데이터로 하여, 트래픽 그룹을 형성하고 하이퍼플레인을 형성한다(S30). 즉, 상기 트래픽 벡터들을 데이터로 하여, 피셔 선형 분류법을 이용하여 정상 그룹과 비정상 그룹으로 분류할 수 있는 하이퍼플레인을 형성한다.

[0060] 본 발명에 따른, 상기 각 트래픽 벡터들의 성분인 대역폭(bandwidth)의 용량, 플로우(flow)의 용량 및 패킷(packet)의 용량의 분포는 카이스퀘어 분포를 따른다.

[0061] 즉, 상기 트래픽 벡터의 각 성분에 대한 상기 수집 기간 동안(예를 들어, 30일 동안)의 분포는 카이스퀘어 분포를 따른다. 여기서, 상기 트래픽 벡터가 정상 트래픽에 해당하는가 또는 비정상 트래픽에 해당하는가를 결정짓

는 기준을 정한다.

[0062] 본 발명에서는 각 성분 별로 용량이 상대적으로 매우 큰 범위를 비정상 범위로 간주한다. 구체적으로, 수집 기간 동안의 카이스퀘어 분포에서 대역폭의 용량이 상위 특정 범위에 속하는 경우에는 비정상 트래픽에 해당하고, 수집 기간 동안의 카이스퀘어 분포에서 플로우의 용량이 상위 특정 범위에 속하는 경우에는 비정상 트래픽에 해당하고, 수집 기간 동안의 카이스퀘어 분포에서 패킷의 용량이 상위 특정 범위에 속하는 경우에는 비정상 트래픽에 해당하는 것으로 결정한다.

[0063] 즉, 트래픽 벡터의 각 성분의 용량이 하나라도 상위 분포 범위에 해당하는 경우에는 비정상 트래픽으로 분류하고, 모든 성분의 용량이 상위 분포 범위에 해당하지 않는 경우에는 정상 트래픽으로 분류한다.

[0064] 본 발명에서는 상기 상위 분포 범위를 4% ~ 6% 사이로 한다. 실험적으로, 상기 범위 내를 기준으로 하이퍼플레인을 형성한 후, 트래픽의 정상/비정상을 탐지한 경우에 탐지 정확도가 높았기 때문이다.

[0065] 상기 정상 트래픽으로 분류되는 트래픽 군들은 정상 그룹으로 분류되고, 비정상 트래픽으로 분류되는 트래픽 군들은 비정상 그룹으로 분류된다. 구체적으로 정리하면, 트래픽 벡터의 각 성분의 용량이 하나라도 상위 4% ~ 6%에 해당하는 경우에는 비정상 그룹으로 분류하고, 모든 성분의 용량이 상위 4% ~ 6%에 해당하지 않는 경우에는 정상 그룹으로 분류한다.

[0066] 상기와 같은 정상 그룹 또는 비정상 그룹의 기준면은 피셔 선형 분류법에 의해 형성되는 하이퍼플레인이다. 상기 하이퍼플레인은 상술한 피셔 선형 분류법을 이용하여 법선 벡터 w구하고, 이 법선 벡터에 수직이면서 각 성분에 대한 전체 평균값에 의하여 공간에 표시되는 평균점을 지나는 평면을 구함으로써 형성할 수 있다.

[0067] 상기와 같이, 하이퍼플레인이 형성되면, 정상 그룹 영역과 비정상 그룹 영역이 구분되어진다. 즉, 하이퍼플레인의 면을 포함한 아래 영역은 정상 그룹 영역이고, 하이퍼플레인의 위 영역은 비정상 그룹의 영역이다.

[0068] 지금까지 과정에 의하여, 유입되는 트래픽에 대하여 정상 또는 비정상 트래픽으로 분류할 수 있는 기준면(하이퍼플레인)을 형성하였다. 그러면, 이후부터는 실제 라우터를 통하여 유입되는 트래픽이 정상 트래픽인지, 비정상 트래픽인지 판단할 수 있다.

[0069] 따라서, 트래픽을 라우터를 통하여 유입받는다(S40). 그런 다음, 유입되는 트래픽에 대하여 세개의 성분(대역폭(bandwidth)의 용량, 플로우(flow)의 용량 및 패킷(packet)의 용량)을 나타내는 트래픽 벡터를 생성하고, 이 세개의 성분에 의하여 표시되는 위치가 하이퍼플레인의 위 영역에 위치하는가 판단한다(S50). 즉, 유입되는 트래픽이 하이퍼플레인의 위 영역에 위치하는가 판단한다. 물론, 하이퍼플레인의 면을 포함한 아래 영역에 위치하는가로 판단할 수도 있다.

[0070] 상기 판단 결과, 유입되는 트래픽이 하이퍼플레인의 면을 포함한 아래 영역에 위치하는 경우에는 정상 그룹(트래픽)으로 판단하고(S60), 하이퍼플레인의 위 영역에 위치하는 경우에는 비정상 그룹으로 판단한다(S70).

[0071] 이와 같은 방법을 통하여, 유입되는 트래픽이 정상 트래픽인지, 비정상 트래픽인지를 용이하게 분류할 수 있다.

[0072] 한편, 트래픽들의 추이는 지속적으로 변화기 때문에, 이를 반영할 필요가 있다. 따라서, 본 발명에서는 유입되는 트래픽의 정상/비정상 여부를 결정하기 위한 기준면인 하이퍼플레인을 업데이트하는 과정을 거친다.

[0073] 따라서, 업데이트 주기를 사전에 설정해 놓고, 시간이 지남에 따라 증가되는 업데이트 값이 상기 업데이트 주기에 도달하면, 상기 하이퍼플레인을 업데이트하는 과정을 수행한다. 예를 들어, 업데이트 주기가 1시간이고, 시스템 내부에 구비되는 타이머가 1시간에 도달하면, 하이퍼플레인의 업데이트를 수행하고, 상기 타이머는 리셋된다.

[0074] 상기와 같이, 업데이트 주기에 도달했는지, 판단한 결과, 업데이트 주기에 도달한 경우에는, 기존의 트래픽 벡터들은 버리고, 새롭게 유입된 트래픽 벡터들을 포함시켜 새로운 하이퍼플레인을 형성한다. 구체적으로, 업데이트 주기에 도달한 경우에는 업데이트 주기에 해당하는 기간 동안 수집된 가장 오래된 트래픽 벡터들을 버리고, 업데이트 주기 동안 새롭게 생성된 트래픽 벡터들을 포함시켜 새로운 하이퍼플레인을 형성한다(S80~S90).

[0075] 예를 들어, 업데이트 주기가 10분이라고 가정할 때, 정상/비정상 트래픽인지를 판단받기 위하여 10분 동안 유입된 트래픽 벡터들은 새로운 하이퍼플레인을 형성하기 위하여 사용되어지고, 사전에 수집되어 사용되어진 트래픽 벡터들 중에 10분 동안 수집된 가장 오래된 트래픽 벡터들은 버려진다.

[0076] 이와 같은 수행 과정을 통하여, 하이퍼플레인은 주기적으로 업데이트되기 때문에, 가변적인 트래픽 추이에 용이

하게 대응할 수 있다. 새로운 하이퍼플레인이 형성되면, 다시 새로운 트래픽을 유입하고, 정상/비정상 여부를 판단하는 절차(S40~S70))를 반복해서 수행한다.

[0077] 이상에서 설명한 방법에 따라, 서브 네트워크 상의 트래픽에 대하여 비정상 트래픽을 용이하게 탐지할 수 있으며, 가변적인 트래픽 추이에 효과적으로 대응할 수 있다.

[0078] 이하에서는, 본 발명을 실제 적용한 구체적인 실시예 대하여 상세하게 설명한다.

[0079] 먼저, 본 발명에 적용되는 용어인 트래픽 벡터와 트래픽 그룹을 구체적으로 정의한다.

[0080] $\mathbb{X}=(x_1,\ldots,x_n)$ 는 트래픽 정보를 성분으로 갖는 벡터인데, 이 벡터를 트래픽 벡터라 정의하고, $X=\{\mathbb{X}_i\}_{i=1}^{k}$ 중에서 동일한 특성을 갖는 트래픽 벡터들의 집합을 트래픽 그룹이라 정의한다.

[0081] 트래픽 벡터는 각 트래픽 그룹의 특성을 나타낼 수 있는 측정 가능한 항목들을 그 성분으로 가지며 그 성분은 라우터로부터 관찰되는 정보인 출발지 IP, 목적지 IP, 출발지 포트, 목적지 포트, 프로토콜, 플로우 등이 될 수 있다. 또한 트래픽 그룹은 이러한 동일한 특성을 갖는 트래픽 벡터들의 모음이라 할 수 있다.

[0082] 트래픽 그룹은 정상 그룹, 비정상 그룹, 웜(worm) 그룹, 트래픽 혼잡 그룹 등 여러 그룹으로 분류할 수 있고, 특히 비정상 그룹의 경우 네트워크 오동작 그룹, flash crowd 비정상 그룹, 네트워크 남용 그룹 등 구체적으로 세분화될 수 있다([P. Barford and D. Plonkz, "Characteristics of Network Traffic Flow Anomalies," in Proc. ACM SIGCOMM 2001, pp. 69-73, August 2001.]).

[0083] 본 실시예에서는 고려대학교의 백본 라우터를 통해 수집된 트래픽을 분석하였다. 트래픽은 2008년 5월 14일 자정을 시작으로 매 5분 단위로 2008년 6월 12일 23:55분까지 30일 동안 수집하였다. 하루 288개의 관찰 시점으로 얻은 트래픽 집합은 총 8640개로 구성된다.

[0084] 트래픽 벡터는 세 개의 성분을 가지며 그 성분은 대역폭의 용량(b), 플로우의 용량(f), 패킷의 용량(p)으로 정의하였다. 또한 그룹을 분리하기 위하여 정상 그룹과 비정상 그룹을 고려한다. 도 3은 30일 동안 수집한 트래픽과 그의 일부인 하루 트래픽의 대역폭(bandwidth)의 용량에 대한 경향을 보여준다.

[0085] 상기 도 3을 통하여 주목해야할 점은 밤 11시부터 오전 6시까지는 매우 적은 양의 트래픽이 발생한다는 것인데, 이는 이 시간 동안 학교 도서관이 닫혀진 결과이다. 상기 도 3은 대역폭의 용량에 대해서만 보여주고 있으나, 플로우의 용량 및 패킷의 용량에 대한 경향도 상기 대역폭의 용량에 대한 경향과 유사한 패턴을 가진다.

[0086] 수집된 트래픽 집합으로부터 두 개의 그룹을 분리하기 위한 기준을 만들기 위하여 수집한 트래픽의 분포를 관찰한다. 도 4는 일일 데이터의 대역폭 용량의 분포이고, 도 5는 30일간 수집한 데이터의 대역폭 양의 분포를 나타내고 있다. 도 4를 통하여 비록 연속적이지는 않지만 0 ~ 20Mbps를 제외한 일일 트래픽의 분포가 표준정규분포를 따름을 알 수 있다.

[0087] 0 ~ 20Mbps를 제외한 이유는 밤 11시부터 오전 6시까지 매우 적은 양의 데이터가 포함되어 있어서, 이는 비정상 기준에 포함되지 않기 때문이다. 도 5에 나타난 바와 같이, 30일 동안 수집한 트래픽은 왼쪽으로 치우쳐 있고 오른쪽으로 긴 꼬리 모양을 갖는 카이스퀘어 분포를 따르고 있음을 알 수 있다.

[0088] 일일 데이터가 표준정규분포를 따르고 이를 제곱하여 더한 그래프가 카이스퀘어 분포를 따르고 있으므로, 30일 동안의 데이터는 자유도가 30인 카이스퀘어 분포를 따른다고 가정해도 무방함을 알 수 있다.

[0089] 여기서, 약 5%의 데이터가 전체 트래픽의 비정상 범위에 속한다고 가정하고 그룹을 나누기 위한 기준을 만들었다. 카이스퀘어 분포의 표(Table C.5, [K. Trivedi, Probability and Statistics with Reliability, Queuing, and Computer Science Applications, 2nd ed., John Wiley and Sons, 2002, pp. 658-664.])를 참고하여 $\chi^2_{30:0.05,b}=437.730$ 의 결과를 얻었다.

[0090] 이는 437.730Mbps 이상의 대역폭(bandwidth) 용량을 가지는 트래픽은 비정상 범위에 속한다는 것을 의미한다. 이 기준에 의하면, 8640의 트래픽 중 8468개의 트래픽이 정상에 속하고, 남은 172개의 트래픽이 비정상 그룹에 속함을 알 수 있다. 그리고, 같은 방법으로 플로우 용량 및 패킷의 용량에 대해서도 적용한 결과, 각각 8617개와 8263개가 정상 그룹에 속함을 알 수 있었다.

[0091] 또한 MS-SQL의 쿼리에 OR 조건을 적용하여 3개의 트래픽 벡터의 성분 중 하나라도 비정상이라면 비정상 그룹에 속한다는 기준을 정하였다. 이 기준에 따라 정상 그룹과 비정상 그룹이 각각 $|Y_n|=8241$와 $|Y_{an}|=399$로 나

타났다.

[0092] 여기서 $|Y_n|$는 정상 그룹에 속하는 트래픽의 개수이고, $|Y_{an}|$는 비정상 그룹에 속하는 트래픽의 개수이다.

[0093] 다음은 상기와 같은 결과를 이용하고, 피셔 선형 분류법을 이용하여 하이퍼플레인(hyperplane)을 얻는 과정을 설명하면 다음과 같다.

[0094] 먼저 법선 벡터 w를 구하기 위해 각 그룹의 평균과 전체 평균을 모두 구하고, S_{intra}와 S_{inter}의 값을 각각 계산한다. 트래픽에 관한 각각의 평균은 표 1에 나타나 있다.

[0095]

average	bandwidth(bps)	flow(fps)	packet(pps)
average for normal traffic μ_n	μ_{nb}=87,945,088.43	μ_{nf}=5,220.758403	μ_{np}=12,951.99879
average for abnormal traffic μ_{an}	μ_{anb}=429,317,861.4	μ_{anf}=5,789.884712	μ_{anp}=129,538.0727
average for total traffic μ_{total}	$\mu_{total,b}$=103,709,872.7	$\mu_{total,f}$=5,247.040972	$\mu_{total,p}$=18,336.00845

[0096] - 표 1 -

[0097] w를 찾기 위해 수식 (1)에 주어진 고유값 문제의 특성 방정식 $|S_{intra}^{-1} S_{inter} - \lambda I| = 0$ 을 이용한다. 이 특성 방정식으로부터 3개의 다른 고유값과 그에 대응하는 고유벡터를 구하면 다음과 같다.

[0098] $\lambda_1 = 0.1882e^{002}$, $\lambda_2 = -0.1083e^{-14}$, $\lambda 3 = 0.7932^{-008}$

[0099] $w_1 = (-28.6341, -0.9995e^{-002}, -0.4927)^t$

[0100] $w_2 = (0.2597e^{-003}, -0.3358e^{-002}, 9.8352)^t$

[0101] $w_3 = (9995.6493, -0.1654e^{-002}, -0.1577e^{-001})^t$

[0102] 이렇게 구한 고유값들 중 가장 큰 값에 대응하는 고유벡터가 법선벡터 w가 된다. 따라서, $\lambda_1 = 0.1882e^{002}$에 대응하는 $w_1 = (-28.6341, -0.9995e^{-002}, -0.4927)^t$이 법선벡터가 된다.

[0103] 이제 이렇게 구한 법선벡터와 수직이고 평균점을 지나는 평면을 구함으로써 찾고자 하는 하이퍼플레인을 정의할 수 있다. 이에 따라 구한 하이퍼플레인의 방정식은 다음과 같다.

[0104] $-28.6341(x - 103,709,872.7) -$

[0105] $0.9995e^{-002}(y - 5,247,040) - 0.4927(z - 18,336,008)=0$

[0106] 도 6은 각각의 트래픽과 w, 하이퍼플레인, 분리된 두 개의 그룹을 보여준다. 상기에서 구한 법선 벡터와 하이퍼플레인, 그리고 도면을 단순화시키기 위하여 정상 트래픽과 비정상 트래픽의 1/10만을 표현하였다.

[0107] 이 하이퍼플레인을 새로 유입되는 트래픽에 적용함으로써 그 트래픽이 아래 영역인 정상 그룹에 속하는지 위 영역인 비정상 그룹에 속하는지 판단할 수 있다. 이 하이퍼플레인은 탐지하고자 하는 트래픽의 추세를 반영할 수 있도록 계속적으로 변하는 평균값과 하이퍼플레인을 통해 실시간 탐지가 가능하다.

[0108] 한편, 이렇게 구한 하이퍼플레인의 신뢰성을 입증하기 위하여 특정 하루의 데이터인 2008년 7월 2일의 데이터를 이 하이퍼플레인에 적용하여 실험해 보았다. 이 날은 대학의 1학기 성적이 공시되어 학교 포털시스템의 트래픽이 매우 혼잡한 날이었고, 학교 전산처에서 관찰된 트래픽의 추이로 보아 비정상적으로 폭주한 트래픽이 많았던 날로 입증된다.

[0109] 도 7은 모니터링 서버로부터 수집한 비정상성을 보인 테이블의 일부를 나타낸 것이다. 상기 도 7을 통하여 비

정상적인 트래픽의 종류와 발생 시점, 그리고 그 패턴 등을 알 수 있다. 선택된 이 특정 하루에는 비정상적으로 보이는 점프(jump) 트래픽과 급격히 폭주하는 모습의 프로파일(profile) 트래픽이 많이 관찰되었고, 이러한 비정상적 트래픽이 관찰되는 주기도 길고 빈번히 나타남을 알 수 있었다.

[0110] 또한 새로 들어오는 트래픽을 5분 단위로 업데이트하여 기존의 가장 오래된 5분의 트래픽을 삭제하고 새로운 5분의 트래픽을 삽입하여 업데이트된 하이퍼플레인을 만드는 동작을 수행하도록 한다.

[0111] 5분 단위로 계속해서 하이퍼플레인을 업데이트하는 이유는 기존 바이러스나 공격과 달리 매우 빠른 속도로 침해사고를 일으키는 슬램머 웜(slammer worm)과 같은 영향력 있는 공격을 차단하기 위해서이다([D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, and N. Weaver, "Inside the Slammer worm," IEEE Security & Privacy Magazine 1(4), pp. 33-39, July/Aug. 2003.]).

[0112] 슬램머 웜은 출몰과 동시에 마이크로소프트 SQL 데이터베이스 소프트웨어의 약점을 이용해 급속히 확산되었다. 이는 네트워크를 통해 자기 복제를 하며 이동하여 매 8.5초마다 두 배로 확산, 불과 10분만에 취약한 호스트의 90%를 감염시켰다. 전파 속도가 최고조에 달한 출현 후 3분 정도 되는 시점에선 인터넷 네트워크를 통해 초당 5천500만회의 위치 검색 요청 신호를 보낸 것으로 나타났다.

[0113] 도 8은 기존 2008년 5월 14일부터 5월 20일까지의 수집된 트래픽을 삭제하고 새로운 2008년 6월 13일부터 6월 19일까지의 일주일의 수집된 트래픽 집합으로 교체해서 구한 하이퍼플레인의 모습을 보여준다. 새로 형성된 하이퍼플레인의 경향도 약간의 기울기만 변했을 뿐 거의 유사한 모습을 보임을 알 수 있다.

[0114] 5분 단위의 하이퍼플레인의 변화를 살펴보았을 때, 거의 유사한 모습을 보이므로 육안으로는 변화의 모습을 볼 수가 없어서 편의상 일주일 단위의 하이퍼플레인의 변화를 보여준다. 상기 도 8을 통하여 새로 형성된 하이퍼플레인의 경향도 약간의 기울기만 변했을뿐 거의 유사한 모습을 보임을 알 수 있다.

[0115] 다음에서는 본 발명의 성능을 평가한다. 여기서 성능 평가를 위하여 용어를 정의한다.

[0116] 실제로 정상 그룹에 속하면서 정상 그룹으로 옳게 판단된 경우를 true positive라 하고, 실제로 비정상 그룹에 속하면서 비정상 그룹으로 옳게 판단된 경우를 true negative라 하고, 실제로는 정상 그룹에 속하나 비정상 그룹에 속하는 것으로 잘못 판단된 경우를 false positive라 하며, 실제로는 비정상 그룹에 속하나 정상 그룹에 속하는 것으로 잘못 판단된 경우를 false negative라 한다.

[0117] 이제 얻어진 결과에 따라 탐지가 얼마만큼 정확한지를 알아보기 위해 다음과 같이 correct detection probability와 miss detection probability를 정의한다.

[0118] $$P_{cor} = \frac{|TP|+|TN|}{|Total\,traffic|}, P_{mis} = \frac{|FP|+|FN|}{|Total\,traffic|}$$

[0119] 여기서 |FP|, |FN|, |TP|, |TN|은 각각 false positive, false negative, true positive, true negative를 나타낸다. 특정한 날을 기준으로 288개의 트래픽에 탐지과정을 적용하여 표 2와 같은 결과를 얻었다.

[0120]

Categories	normal group	abnormal group
using criterion for FLD	103	185
information from server	108	180
observed results	TP: 101, FN: 7, TN: 179, FP: 1	
probabilities	P_{cor}=0.97, P_{mis}=0.03	

[0121] - 표 2 -

[0122] 비정상 트래픽의 정보를 제공하는 학교 전산처 서버를 통해 보았을 때, 본 발명에서 고려한 특정한 날인 7월

2일에는 108개의 정상 트래픽과 180개의 비정상 트래픽이 있었음을 알 수 있었다. 한편, 본 발명에서 구한 하이퍼플레인에 이 트래픽을 적용해 보았을 때, 103개의 정상 트래픽과 185개의 비정상 트래픽을 탐지해 냈다.

[0123] 상기 표 2는 탐지율과 오탐지율의 결과를 보여주고 있다. 표 2에서 보는 바와 같이, 3%의 오탐지율이 발생하였으며, 이 오탐지율은 플로우 양의 차이가 매우 적어 트래픽 그룹을 나눔에 있어 생긴 오차로 보인다.

[0124] 본 발명에서 제안한 탐지 기법의 성능을 알아보기 위하여 유사한 환경에서 비정상 트래픽 탐지를 다룬 기술(1)([K. H. Ramah, H. Ayari, and F. Kamoun, "Traffic Anomaly Detection and Characterization in the Tunisian National University Network," in Proc. NETWORKING 2006, LNCS 3976, pp. 136-147, May 2006.])의 결과와 비교해 본다.

[0125] 실험 데이터가 완전히 일치하지는 않으나 상기 기술(1)에서는 Tunisian National University Network(TNUN)을 이용하여 비정상 트래픽을 탐지하였다. 이를 위하여 상기 기술(1)에서는 45일간 주기적으로 캠퍼스 네트워크의 데이터를 수집하고 Anomaly Detection System(ADS)을 개발하였다.

[0126] TNUN 중앙의 방화벽으로부터 MIB(Management Information Base)를 통해 네트워크 트래픽 trace들을 수집한다. 수집한 결과를 통하여 inter-anomaly 시간 분포와 anomaly duration 분포를 계산하여 5분 이내에 비정상 트래픽을 탐지하게 된다. 비록 같은 실험 데이터는 아닐지라도 TNUN의 트래픽도 본 발명에서 사용한 실험 데이터와 거의 유사한 실험 데이터를 사용하였다.

[0127] 그 실험 결과, 본 발명에서 고려한 것과 동일한 탐지 상황 하에서 약 90%의 탐지율을 보임을 알 수 있다. 즉 이 결과와 비교했을 때, 본 발명에서의 기법이 약 7% 우수함을 알 수 있다.

[0128] 이상에서 살펴본 바와 같이, 본 발명에서는 트래픽 벡터와 트래픽 그룹을 정의하고 피셔 선형분류법에 의해 얻은 최적화된 하이퍼플레인을 이용하여 새 트래픽 데이터가 생성되었을 때 그 데이터가 어떤 그룹에 속하는지를 결정하였다. 본 발명에서 제안한 방법은 기존의 방법에 비해 트래픽의 비정상 여부를 실시간으로 보다 간단하고 정확하게 탐지할 수 있었다.

[0129] 또한 트래픽 추이를 주기적으로 반영함으로서 새롭게 나타나는 이상적 현상을 탐지할 수 있었다.

도면의 간단한 설명

[0130] 도 1은 본 발명인 피셔 선형 분류법을 이용한 비정상 트래픽 탐지 방법을 구현하기 위한 시스템도이다.

[0131] 도 2는 본 발명인 선형 분류법을 이용한 비정상 트래픽 탐지 방법의 순서도이다.

[0132] 도 3은 본 발명에서 적용되는 대역폭 용량의 특성 그래프이다.

[0133] 도 4는 본 발명에 적용되는 대역폭 용량의 일일 데이터 분포도이다.

[0134] 도 5는 본 발명에 적용되는 30일간 수집한 데이터의 대역폭 용량의 분포도이다.

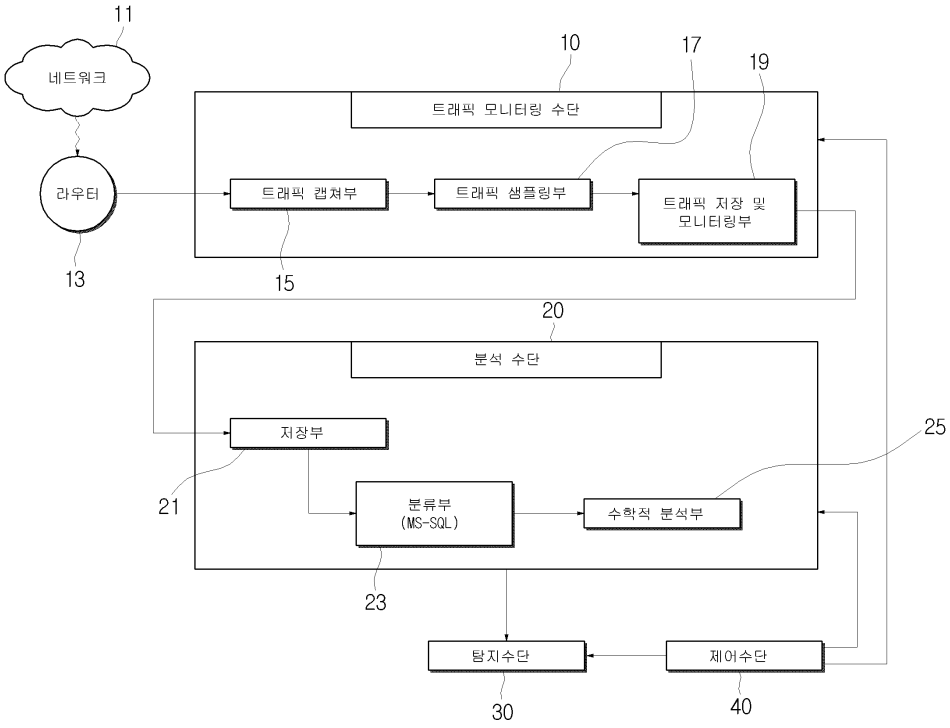
[0135] 도 6은 본 발명에 적용되는 하이퍼플레인을 보여주는 그래프이다.

[0136] 도 7은 본 발명의 실시예에 따른 모니터링 서버에 의해 관찰된 비정상 트래픽의 예시도이다.

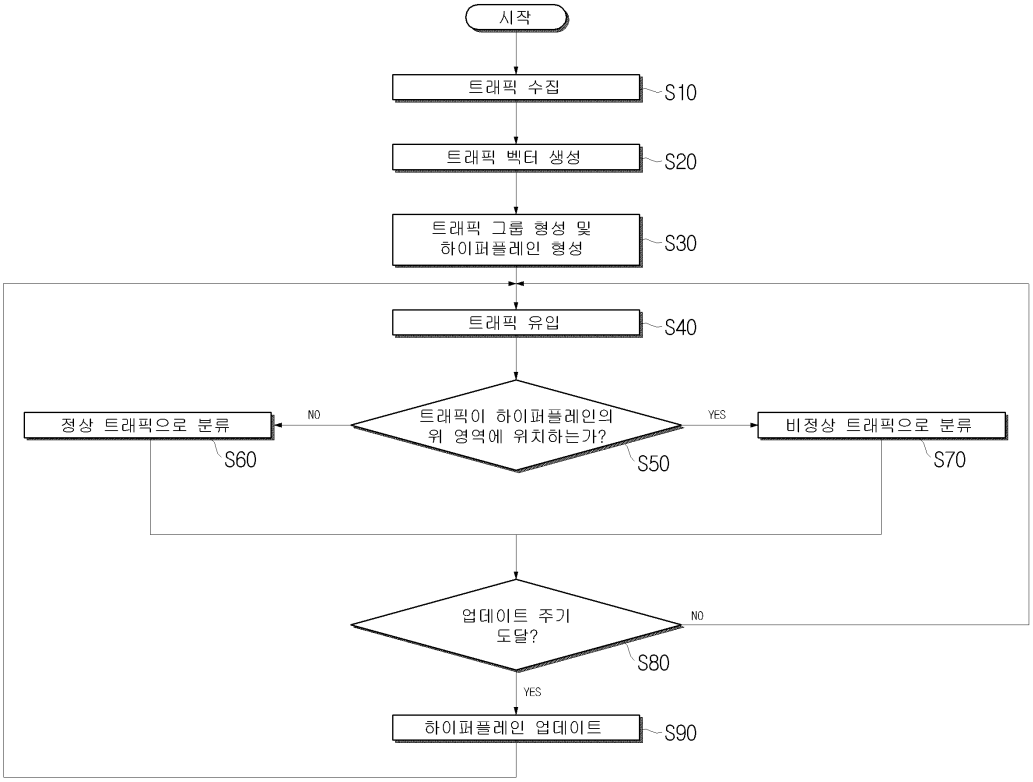
[0137] 도 8은 기존 하이퍼플레인과 업데이트된 하이퍼플레인의 비교도이다.

도면

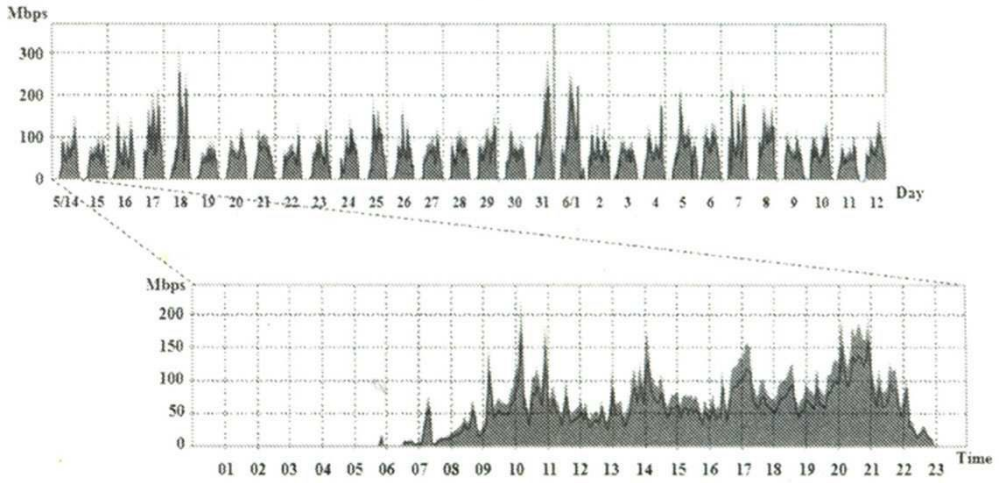
도면1


11
네트워크
10
17
19
트래픽 모니터링 수단
라우터
트래픽 캡쳐부
트래픽 샘플링부
트래픽 저장 및
모니터링부
13
15
20
분석 수단
지장부
25
21
분류부
(MS-SQL)
수학적 분석부
23
탐지수단
제어수단
30
40

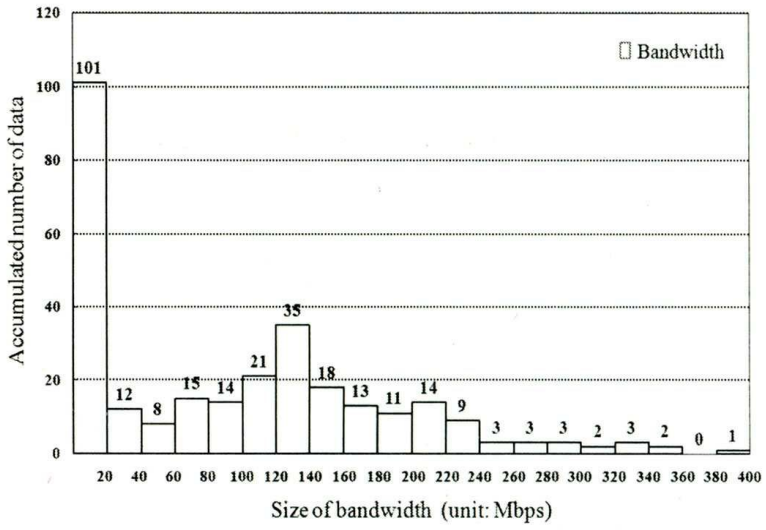
도면2


시작
트래픽 수집
S10
트래픽 벡터 생성
S20
트래픽 그룹 형성 및
하이퍼플레인 형성
S30
트래픽 유입
S40
정상 트래픽으로 분류
NO
트래픽이 하이퍼플레인의
위 영역에 위치하는가?
YES
비정상 트래픽으로 분류
S60
S50
S70
업데이트 주기
도달?
NO
S80
YES
하이퍼플레인 업데이트
S90

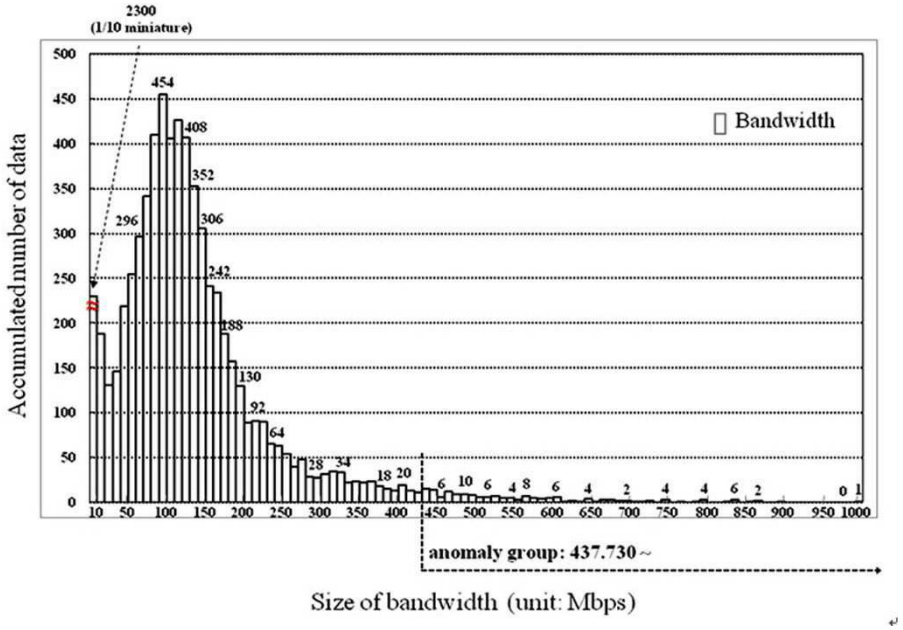
도면3


Mbps
300
200
100
0
5/14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 6/1 2 3 4 5 6 7 8 9 10 11 12
Day
Mbps
200
150
100
50
0
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Time

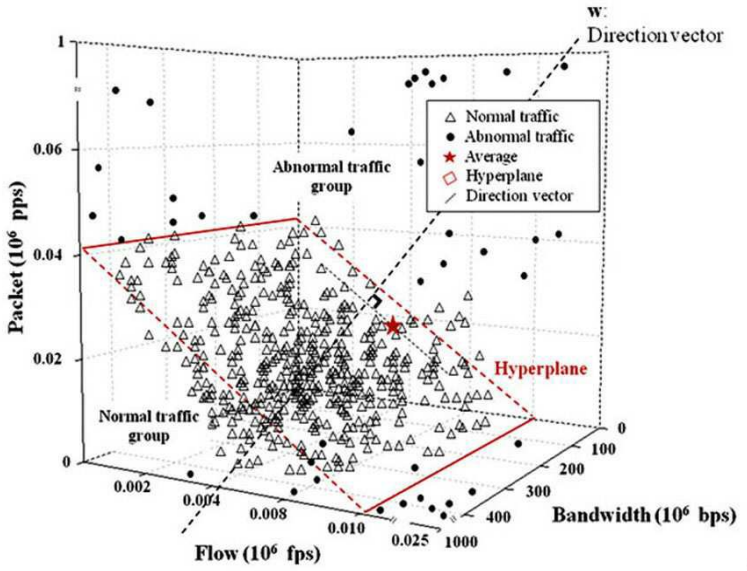
도면4


Bandwidth
Accumulated number of data
120
100
80
60
40
20
0
101
12
8
15
14
21
35
18
13
11
14
9
3
3
3
2
3
2
0
1
20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400
Size of bandwidth (unit: Mbps)

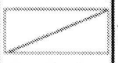
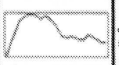
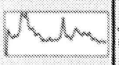
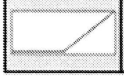
도면5


2300
(1/10 miniature)
Accumulated number of data
Bandwidth
anomaly group: 437.730 ~
Size of bandwidth (unit: Mbps)

도면6


w:
Direction vector
Normal traffic
Abnormal traffic
Average
Hyperplane
Direction vector
Abnormal traffic group
Normal traffic group
Hyperplane
Packet (10⁶ pps)
Flow (10⁶ fps)
Bandwidth (10⁶ bps)

도면7

abnormal traffic pattern	abnormal traffic type	dangerous level	amount of traffic	duration	interface of backbone router in Korea univ. network
	Jump (TrafficPps)	Critical	403.958	2008-07-02 21:37:00~ 2008-07-02 21:37:00	163.152.16.17 - HanaPark_16.17 (HanaPark_16.17) 163.152.16.47 - Vl217
	Jump (TrafficPps)	Critical	1160.620	2008-07-02 21:14:00~ 2008-07-02 21:35:00	163.152.16.17 - HanaPark_16.17 (HanaPark_16.17) 163.152.16.47 - Vl217
	Jump (TrafficPps)	Critical	901.967	2008-07-02 21:14:00~ 2008-07-02 21:35:00	163.152.16.17 - HanaPark_16.17 (HanaPark_16.17) 163.152.16.17 - Vl216
	Jump (TrafficPps)	Critical	1402.500	2008-07-02 20:12:00~ 2008-07-02 21:11:00	163.152.16.17 - HanaPark_16.17 (HanaPark_16.17) 163.152.16.17 - Vl216
	Jump (TrafficPps)	Critical	1200.720	2008-07-02 20:12:00~ 2008-07-02 21:11:00	163.152.16.17 - HanaPark_16.17 (HanaPark_16.17) 163.152.16.47 - Vl217
	Profile (HighPps)	Critical	1.696	2008-07-02 21:10:00~ 2008-07-02 21:10:00	163.152.16.17 - HanaPark_16.17 (HanaPark_16.17) 163.152.16.47 - Vl217

도면8