고정 IP-port 를 이용한 인터넷 트래픽 분석

윤성호*, 박진완*, 오영석*, 김명섭*
*고려대학교 컴퓨터정보학과
e-mail : {hiption, pakjw84, 840105, tmskim}@korea.ac.kr

Internet Traffic Analysis using Fixed IP-port

Sung-Ho Yoon*, Jin-Wan Park*, Young-Seok Oh*, Myung-Sup Kim*
*Dept. of Computer and Information Science, Korea University

요 약

인터넷의 대중화로 인해 네트워크 트래픽이 급증하였다. 따라서 효과적인 네트워크 관리를 위한 트래픽 응용 별 분석의 중요성은 매우 커지고 있다. 효과적인 트래픽 응용 별 분석을 위하여 여러 방법론이 제안 되었지만, 아직 완벽하게 트래픽을 분석하는 방법론은 존재하지 않는다. 본 논문에서는 기존의 여러 트래픽 분석 방법론의 단점을 보완한 고정 IP-port 을 이용한 인터넷 트래픽 분석 방법론을 제안한다. 하나의 서비스를 고정적으로 제공하는 고정 IP-port 을 찾아 낸 다면 효과적으로 트래픽을 분석 할 수 있다. 본 논문에서는 고정 IP-port 를 효과적으로 추출하는 방법론을 제시하며, 실제 학내 트래픽에 적용한 결과를 보인다.

1. 서론

대용량의 인터넷 회선이 보편화되고 인터넷 서비스를 이용하는 사용자가 급격히 증가 함에 따라 네트워크 트래픽이 급증하였다. 이는 전통적으로 사용되는 WWW, FTP, e-mail 등의 인터넷 서비스뿐 아니라 통합된 음성망 서비스, 멀티미디어 파일의 스트리밍 서비스, P2P(peer-to-peer) 파일 공유, 게임 등의 멀티미디어 서비스를 제공하는 네트워크 기반의 응용 프로그램이 더욱 다양하게 개발됨에 따른 것이다. 이로 인하여 네트워크 트래픽의 개별적인 용량이 증가하며, 전체적인 트래픽량이 계속해서 증가하고 있다. 따라서 네트워크 관리를 위해 네트워크 트래픽의 모니터링 및 분석의 중요성이 증대되고 있다[1,2].

네트워크 트래픽 분석이란 관리대상 네트워크의 트래픽을 수집하여 해당 네트워크의 상태를 확인하는 것이다. 이렇게 분석된 트래픽 정보는 네트워크 관리 및 보안 등에 큰 이점을 가지며, 네트워크를 사용하는 사용자의 경향 분석에도 사용되는 등 광범위하게 활용될 수 있다. 이러한 트래픽 분석은 프로토콜별, 서비스 품질(quality of service : QoS)별, 응용별 등 다양한 기준에 의해 수행될 수 있다.

응용 별 트래픽 분석이란, 네트워크에서 발생하는 트래픽을 수집하여 각각의 트래픽을 발생 시킨 응용을 알아내는 것이다. 응용 별 분석은 실제 트래픽을 발생 시키는 응용을 기준으로 하는 것이기 때문에 네트워크 관리 측면에서 큰 이점을 가진다. 예를 들어 특정 응용을 차단하거나 대역폭을 조절 할 수 있다.

응용 별 트래픽 분석을 위하여 well-known port, signature 생성, 머신러닝 등 많은 방법론이 제시 되었다. 하지만 기존의 방법론들은 큰 제한을 가지고 있다. well-known port 방법은 우리가 알지 못하는 응용이나 동적 port 을 사용하는 응용에 대해서는 분석하지 못한다. signature 생성 방법은 분석 시점 이전에 signature 를 생성해 놓아야 할 뿐만 아니라 payload 가 없거나 암호화된 패킷의 경우에는 적용할 수 없다. 머신러닝 방법 또한 분석 시점 이전에 학습을 해야 하고 실시간 분석이나 새로운 응용에 대한 대처가 어렵다는 제한이 있다.

본 논문에서는 앞선 방법론의 단점을 보완한 고정 IP-port 을 이용한 인터넷 트래픽 분석 방법론을 제안한다. 고정 IP-port 란, 하나의 서비스를 고정적으로 제공하는 서버의 IP-port 을 의미한다. 단순히 IP-port 정보만을 이용하기 때문에 payload 정보가 불필요하고 실시간 트래픽을 분석하여 업데이트하기 때문에 선행학습 또한 필요하지 않는 장점을 가진다. 또한 각각의 호스트가 사용하는 응용의 종류를 알 수 있기 때문에 호스트 기반 상관 관계 방법론에 유용하게 사용될 수 있다.

본 논문의 구성은 다음과 같다. 2 장에서는 관련연구를 3 장에서는 고정 IP-port 의 추출 방법, 4 장에서는 추출 알고리즘을 제시한다. 5 장에서는 실험 및 결과를, 마지막 6 장에서는 결론 및 향후 연구를 제시한다.

2. 관련연구

2.1 TMA(Traffic Measurement Agent)

고정 IP-port 를 추출하기 위해서는 각 서버 IP-port 와 통신하는 트래픽이 어떤 응용에서 발생된 것인지를 알아야 한다. 본 논문에서는 실시간으로 발생하는

이 논문은 2007 년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2007-331-D00387)

트래픽의 응용을 확인하기 위해 TMA (Traffic Measurement Agent)[3]를 사용하였다. TMA 는 종단 호스트에서 실행 되며, 해당 호스트의 소켓정보를 주기적으로 수집하여 지정된 서버로 제공한다. 모든 트래픽은 소켓에서 발생되며 소켓 정보는 소켓을 생성한 프로세스 정보를 포함하기 때문에 각 호스트에서 발생한 트래픽이 어떤 응용에서 발생되었는지를 확인할 수 있다. <표 1>은 TMA 가 제공하는 정보들이다.

<표 1> TMA 정보

0	1	2	3	4
Local IP				
Remote IP				
Local Port			Remote Port	
Protocol	State		Process ID	
Process Path (32 bytes)				
Process Name (32 bytes)				

2.2 검증 방법

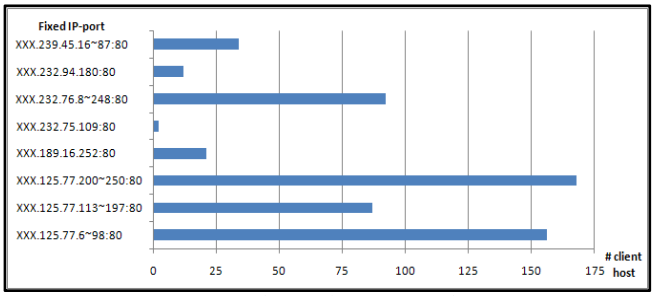
본 논문에서는 추출한 고정 IP-port 를 검증하기 위해 기존의 트래픽 분석 논문에서 제시하는 검증 방법과 다른 검증 방법을 제시한다. BLINC[4]에서는 분석한 결과를 검증하기 위하여 페이로드 기반 분석 방법을 사용하였다. 또한 Constantinou et. al.[5]는 알려진 포트 기반의 분석을 사용하였다. 즉, 분석을 올바르게 했느냐를 확인하기 위해 100%로 확신 할 수 없는 또 다른 트래픽 분석 방법을 사용하였다. 이러한 검증 방법은 트래픽 검증 방법으로 적합하지 않다.

본 논문에서는 정확한 검증을 위하여 TMA 를 이용하였다. 검증이란, 발생 시킨 응용을 알고 있는 정답 트래픽과 분석한 결과를 비교하는 것이다. TMA 는 트래픽을 발생하는 호스트에 직접 동작하기 때문에 해당 트래픽을 발생 시킨 응용을 확인할 수 있다. 따라서, 우리가 분석한 트래픽과 TMA 를 통해 알 수 있는 정답 트래픽을 비교 함으로써 검증한다. 정확한 검증을 위하여 본 시스템은 고정 IP-port 생성 시스템에서 사용하는 TMA 호스트와 검증에서 사용하는 TMA 호스트를 분리하여 사용하였다

3. 고정 IP-port

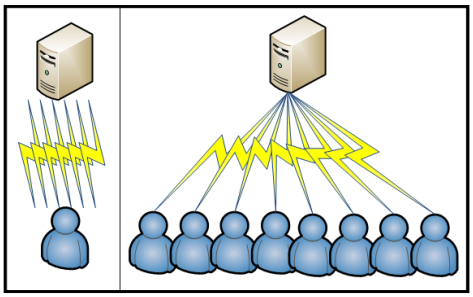
고정 IP-port 란, 일정 기간 동안 단일 서비스를 제공하는 서버의 IP-port 를 뜻한다. 또한 특정 응용을 사용할 시 반드시 접속해야 하는 서버를 뜻한다. 예를 들어 사용자 인증(로그인)이 필수적인 응용의 경우 인증을 위해 통신 하는 서버가 고정 IP-port 인 것이다. 이러한 고정 IP-port 를 각 응용마다 찾아내 관리 할 수 있다면 각 호스트 마다 사용하고 있는 응용을 알아 낼 수 있다. 이러한 정보는 호스트 기반 상관 관계 방법론에 중요한 정보가 된다.

고정 IP-port 를 확인하기 위하여 7 일간 학내에서 발생하는 트래픽을 TMA(Traffic Measurement Agent)기반 분석 방법[3]으로 분석하였다. 2.1 장에서 설명한 바와 같이 모든 트래픽은 응용(process)에 의해 생성된 소켓에서 시작하고 종결됨으로 TMA 정보와 플로우(인터넷과 연결된 백본 스위치에서 수집)를 비교하면 해당 트래픽을 발생 시킨 응용을 알아낼 수 있다. TMA 기반 분석 방법은 해당 호스트의 소켓정보를 기반으로 하기 때문에 TMA 가 설치된 호스트에서 발생한 트래픽을 100%의 신뢰도로 분석할 수 있다.



(그림 1) gom.exe 의 고정 IP-port 과 Client host 수

(그림 1) 은 7 일간 동영상 스트리밍을 서비스하는 gom.exe 프로세스를 사용할 때 발생되는 트래픽 중 다른 프로세스와 단 한차례도 "충돌" (하나의 서버 IP-port 가 둘 이상의 서로 다른 프로세스와 통신하는 경우) 하지 않은 서버 IP-port(고정 IP-port)와 해당 서버와 통신한 client host 들의 수를 나타낸 것이다. 즉, 응용마다 고정으로 한 서비스만 제공하는 고정 IP-port 를 가지고 있는 것을 확인할 수 있다. 또한 고정 IP-port 들이 특정 C 클래스 영역에 모여 있는 것을 알 수 있다. BLINC[4]에서는 이러한 IP 의 무리를 "farms" 라는 용어로 설명한다. 같은 도메인 안에서 로드분산을 위해 여러 서버들 사용하는데 이러한 서버들을 "farm" 이라고 한다.



(그림 2) 빈도 중심과 Client 중심의 고정 IP-port

고정 IP-port 는 두 가지로 구분 할 수 있다. (그림 2)와 같이 단일 client 와 통신하지만 지속적으로 같은 응용으로 통신하는 것과 여러 client 들과 같은 응용으로 통신하는 경우가 있다. 전자는 비록 사용자는 적지만 고정된 서버를 사용하는 응용의 경우이고 후자는 사용자가 많은 응용의 경우이다. 각 경우에 따라 고정 IP-port 의 추출 방법은 달리한다.

- 빈도(frequency) 중심의 추출 방법
  단일 호스트가 특정 서버와 2 차례 이상 통신할 경우 해당 서버를 고정 IP-port 라 한다.
- Client 중심의 추출 방법

2 개 이상의 client 들과 동일한 응용으로 통신하는 서버를 고정 IP-port 라 한다.

비록 위의 두 가지 방법 중 하나를 만족 하더라도 다른 응용과 충돌 할 경우 고정 IP-port 에서 삭제한다.

4. 알고리즘

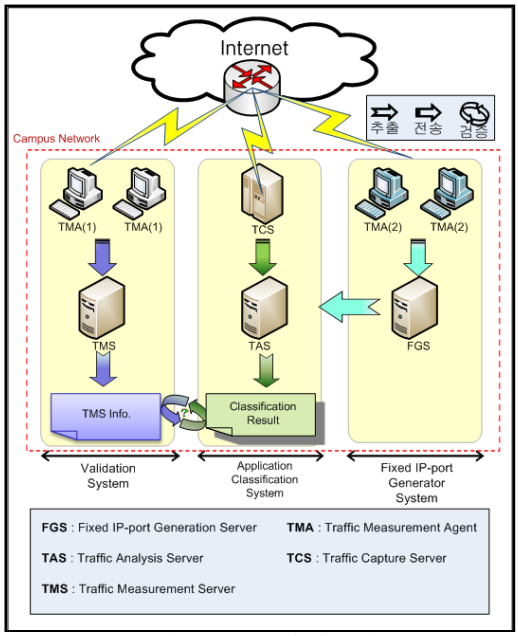
고정 IP-port 를 추출하기 위한 시스템은 다음과 같은 요소를 반영하여 한다.

•Coverage: 최대한 많은 응용에서 고정 IP-port 를 찾아내어야 한다. 응용마다 동작하는 방식이 다르기 때문에 대부분의 응용에 적용되는 고정 IP-port 생성 알고리즘을 만들어야 한다.

•Accuracy: FP(false positive)와 FN(false negative)를 최소화하는 정확한 고정 IP-port 를 찾아야 한다. 고정 IP-port 는 상관 관계 방법론의 첫 단계이기 때문에 100% 신뢰도를 가지고 있어야 한다.

•Overhead: 고정 IP-port 생성시 필요한 IP-port 들의 정보를 최소화 시켜 파일 I/O 시 발생되는 overhead 를 최소화 한다. 고정 IP-port 를 생성하기 위해서는 과거의 정보를 계속 유지하고 있어야 한다. 무한정 모든 IP-port 의 정보를 저장하는 것은 고정 IP-port 생성의 가능성을 높여주긴 하지만, 실시간 트래픽 분석 시 많은 부하를 발생시킨다. 따라서 고정 IP-port 로 결정될 가능성이 높은 로그들만 유지 하여야 한다.

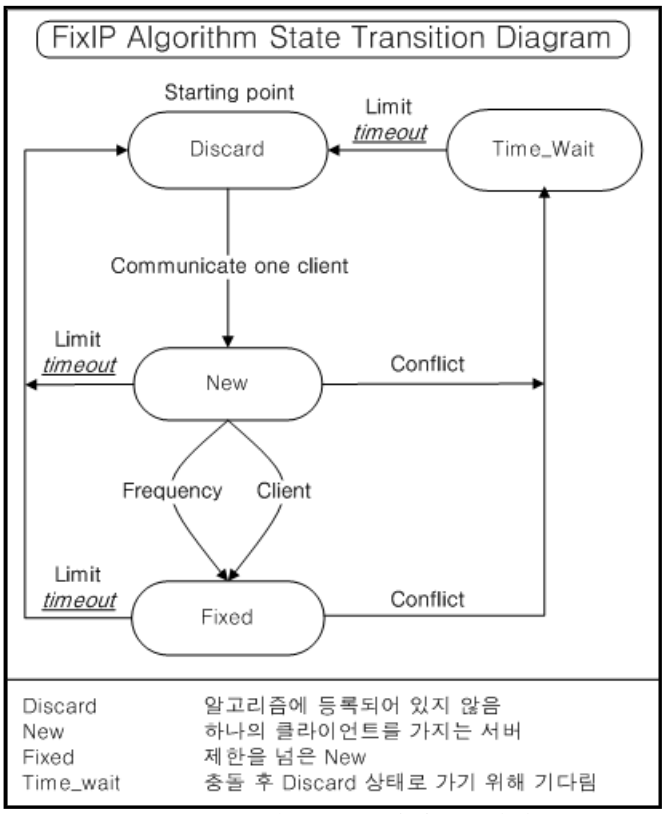
고정 IP-port 를 추출하기 위해 학내 네트워크에 시스템을 구축하였다. (그림 3)는 고정 IP-port 추출 시스템 구조를 나타낸다.



(그림 3) 고정 IP-port 추출 시스템 구조

본 시스템은 크게 고정 IP-port 생성 시스템, 응용별 분석 시스템, 그리고 검증 시스템으로 구성된다.

고정 IP-port 추출 시스템에서는 TMA 정보와 고정 IP-port 알고리즘을 이용하여 고정 IP-port 를 생성한다. 응용 별 분석 시스템에서는 학내 네트워크와 인터넷 사이에서 수집한 플로우 단위의 트래픽을 고정 IP-port 를 이용하여 분석한다. 이렇게 분석된 결과는 고정 IP-port 생성에 관여하지 않은 학내 TMA 정보와 비교 함으로써 검증한다.



(그림 4) 고정 IP-port 추출을 위한 State Transition Diagram

고정 IP-port 를 추출하기 위해 (그림 4)와 같은 State Transition Diagram 을 고안하였다. 트래픽에 나타나는 모든 IP-port 들은 위에 제시된 4 가지 상태 중 하나의 상태에 속하게 된다. 즉, 최초 TMA 로그의 모든 IP-port 들은 알고리즘에 등록되어 있지 않은 Discard 상태에 속한다. 실시간으로 발생되는 트래픽을 분석하여 하나의 client 와 통신하는 특정 응용 서버 IP-port 를 New 상태에 속하게 한다. 그 이후, 3 장에서 설명한 2 가지 조건에 속하는 IP-port 의 상태를 Fixed 로 한다. Fixed 상태의 레코드들이 본 논문에서 제안하는 고정 IP-port 들이다.

New 와 Fixed 상태에서는 Limit Time 과 충돌을 체크하여 무한히 과거의 정보를 유지시키는 것과 한 상태에서 영원히 머무르는 것을 방지 하였다. 또한, Time_wait 상태에서는 고정이 아닌 IP-port 즉, 여러 응용에 의해 사용되는 IP-port 들이 등록과 동시에 충돌이 일어나 시스템의 overhead 가 증가되는 것을 방지하기 위해 만들었다.

State Transition Diagram 에 따라 각 IP-port 의 상태를 변경하는 것이 본 알고리즘의 핵심이다. 고정 IP-

port 알고리즘은 다음과 같은 슈도코드(pseudo code)에 의해 수행된다.

```
1:  procedure FixedIP_Extraction
2:    TT ← TMA Table
3:    IPT ← IP-port Table
4:    for each record in TT do
5:      search the record in IPT
6:      if already registered in IPT then
7:        update;
8:      else
9:        move to New;
10:   end for
11:   for each record in IPT do
12:     Check State of the record
13:     if State == New then
14:       if conflict then
15:           move to Time_Wait;
16:       else if communicate two more clients then
17:           move to Fixed;
18:       else if communicate four more times then
19:           move to Fixed;
20:       else if over time limit then
21:           move to Discard;
22:     else if State == Fixed then
23:       if conflict then
24:           move to Time_Wait;
25:       else if over time limit then
26:           move to Discard;
27:     else if State == Time_Wait then
28:       if over time limit then
29:           move to Discard;
30:    end for
```

본 알고리즘은 크게 두 부분으로 구분된다. 새로운 IP-port 를 New 상태로 등록 시키거나 이미 등록된 IP-port 의 정보를 업데이트(충돌여부, client 개수, 통신 횟수, 마지막 사용 시간 등)하는 부분과 업데이트된 정보에 따라 상태를 변화 시켜주는 부분이다.

5. 실험 및 결과

본 논문에서 제안한 고정 IP-port 추출 방법론을 실제 학내 트래픽에 적용하였다. 12 일간 고정 IP-port 를 생성하고 추출된 고정 IP-port 를 검증하기 위해 1 일간 추출된 고정 IP-port 를 이용하여 트래픽을 분석하였다.

<표 2> 분석 결과

Completeness	28.8%(flow)
Accuracy	95.39%(flow)
Overhead	231705(fixed)/238543(total)
Coverage	346(fixed)/354(total)

총 354 개의 응용의 정보를 입력 받아 346 개의 응용에서 고정 IP-port 를 찾아냈다. 또한 추출된 고정 IP-port 를 실제 네트워크 트래픽을 대상으로 분석한 결과 <표 2>와 같은 결과를 나타내었다.

대부분의 응용에서 정확한 고정 IP-port 를 찾아 내었다. Coverage 에서 제외된 응용은 특정 서버를 사용하지 않는 경우와 실험 기간 중 오직 한차례 트래픽이 발생한 경우이다. 대부분의 로그가 고정 IP-port 이므로 고정 IP-port 가 아닌 데이터를 유지 시키기 위한 overhead 는 매우 낮다.

고정 IP-port 를 사용하지 않는 P2P Transfer 의 경우, 본 시스템으로 분석하지 못하므로 낮은 Completeness 을 가진다. 하지만 서론에서 설명한 바와 같이 정확한 고정 IP-port 을 시작으로 여러 알고리즘을 적용한다면 만족할 만한 결과를 얻을 수 있다.

또한 생성한 고정 IP-port 를 분석한 결과, <표 3>과 같이 “farm”을 형성하는 IP-port 대역을 찾을 수 있었다.

<표 3> 고정 IP-port “farm” 의 예

IP	port	service	process
xxx.234.239.175 ~ 217 xxx.234.240.135 ~ 252	5004	nateon	nateonmain
xxx.125. 77. 93 ~ 250 xxx.232. 76. 29 ~ 148	80	gretech	gom
xxx. 38.137. 11 ~ 48	80	ahnlab	acaas
xxx.153. 8. 51 ~ 67	80	estsoft	albncollector
xxx.234.243.132 ~ 171	80	cymini	skcbgm

6. 결론 및 향후 연구

네트워크에서 발생한 트래픽을 응용 별로 분석하는 것은 네트워크 관리 측면(특정 응용의 차단 혹은 대역폭 제어)에서 많은 이점을 가진다. 응용 별 분석을 위하여 여러 방법론이 제안 되었지만 실시간 분석시스템에 적용하기에는 다소 어려운 점을 가지고 있다. 본 논문에서는 실시간 트래픽 분석을 위해 고정으로 하나의 응용만 서비스하는 고정 IP-port 를 생성하는 시스템을 구축하였다. 이렇게 생성된 IP-port 들은 상관 관계 방법론에 중요한 정보를 제공한다. 앞으로 본 시스템을 발전 시켜 호스트 기반 상관 관계 방법론을 완성 시키고 또한 응용 별 트래픽 분석에 있어 가장 기본적인 응용을 정의하는 작업과 정확한 검증 네트워크를 구성하는 작업을 병행 할 계획이다.

참고문헌

[1] Myung-Sup Kim, Young J.Won, James Won-Ki Hong, “Application-Level Traffic Monitoring and an Analysis on IP Networks”, ETRI Journal Vol. 27, No.1, February 2005.

[2] S. Sen, J. Wang, “Analyzing peer-to-peer traffic across large networks”, Internet Measurement Conference (IMC), Proc. Of the 2nd ACM SIGCOMM Workshop on Internet measurement, pp 137-150, 2002.

[3] 윤성호, 노현구, 김명섭, "TMA(Traffic Measurement Agent)를 이용한 인터넷 응용 트래픽 분류", 2008 년 제 29 회 정보처리학회 춘계학술발표대회 (KIPS), 대구, 경일대학교, May. 17, 2008, 제 15 권 제 1 호, pp.946-949.

[4] T. Karagiannis, K.P apagiannaki and M.F aloutsos, “BLINC: Multilevel Traffic Classification in the Dark,” in Proc. of ACM SIGCOMM, August 2005.

[5] F. Constantinou and P. Mavrommatis. Identifying known and unknown peer-to-peer traffic. In IEEE International Symposium on Network Computing and Applications (NCA), pages 93–102, Cambridge, MA, USA, July 2006.