HTTP 헤더를 이용한 서버별 분류

(Server-Side Classification with the HTTP Header Information)

진창규*, 김명섭**,최미정*°
*강원대학교 컴퓨터과학과, **고려대학교 컴퓨터정보학과
*{jinchanggyu, mjchoi}@kangwon.ac.kr, **tmskim@korea.ac.kr

요 약

최근에는 언제 어디서든 인터넷에 연결할 수 있는 네트워크 환경이 만들어졌다. 원하는 정보를 얻고 싶을 때에는 언제든지 스마트폰, 타블릿 PC 등을 사용해 웹에 접속하여 정보를 얻을 수 있다. 또한 트위터, 페이스북 등 소셜 네트워크를 통하여 많은 사람들과 정보를 손쉽게 공유할 수 있다. 그러나 손쉬운 정보획득은 부작용을 야기한다. 청소년들은 유해 사이트에 너무나 손쉽게 노출되고, 스팸 메일로 인하여 원치 않는 정보들을 받기도 한다. 이러한 이유로 네트워크 관리와 네트워크의 원활한 서비스를 위해 어떤 정보들이 어떤 사이트에서 발생하는지에 대한 파악 또한 중요해지고 있는 시점이다. 따라서 본 논문에서는 HTTP 트래픽의 서버별 분류 방법을 제시한다. 서버별 분류를 통해 추후에 HTTP 트래픽에 대한 사이트별 제어를 수행할 수 있다. 제안된 방법을 학내 네트워크에서 발생하는 HTTP 트래픽을 수집하고 분류함으로써 그 타당성을 검증한다.

Keywords: HTTP, Server-specific Traffic Classification, Site-level

1. 서론

최근에 인터넷의 많은 정보들은 웹을 통하여 공유되고 있다. 또한 스마트폰의 보급과 소셜 네트워크의 발전으로 인하여 언제 어디서든 원하는 정보를 웹을 통하여 요청하고, 공유할 수 있는 환경이 되었다. 네트워크의 발전과 정보의 증가로 많은 정보들을 우리는 손쉬운 방법으로 얻을 수 있지만 이에 대한 부작용도 존재하고 있다. 청소년들은 성인물에 노출이 되고, 스팸메일 등을 통하여 원치 않는 정보들을 접하고 있다. 또한 회사에서는 업무 시간에 증권 사이트에 접속하는 등의 업무의 효율성을 저해하는 사례가 발생한다.
이러한 이유들로 인하여 서비스 관리자 측면에서는 어떤 내용의 정보들이 어떤 사이트에서 검색되는지에 대한 파악이 중요해지고 있다. 유해 사이트나 제어가 필요한 사이트에 대한 블록(block)을 수행하기 위해서는 HTTP 트래픽의 사이트별 분류가 필요하다. 또한, 해당 사이트에 광고를 하고자 하는 광고주의 입장에서는 해당 사이트에 얼마나 많은 접속이 일어나는지를 파악하는 것이 필요하다.

본 논문에서는 선행 연구에서 수행한 HTTP 트래픽의 어플리케이션별(클라이언트 측면) 분석[1][7]에 이어서 어떤 서버에서 HTTP 트래픽이 발생하는지에 대해 확인하기 위해, 다차원분석 방법 중에 하나인 서버측에서 발생하는 트래픽을 기준으로 서버별(사이트 측면) 분류를 제안하고자 한다. 또한 한 사이트 내에서도 다양한 사이트가 존재한다. 예를 들면 네이버의 경우 blog, mail, search, dic, maps 등의 다양한 사이트가 존재하며, 각 사이트별 분류에 대한 세부 분석이 필

이 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단(2009-0090455)의 지원 및 2011년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(20110014032)

° 교신저자: 최미정 (mjchoi@kangwon.ac.kr)

요하다. 서버별 분류에서 Top10 리스트에 대해 세부 분류를 수행하고자 한다.

본 논문은 다음과 같은 순서로 구성된다. 2장에서는 관련 연구에 관하여 설명하고, 3장에서는 시스템 설계 및 구현에 대해 설명한다. 4장에서는 학내망에서 발생하는 트래픽을 대상으로 제안하는 시스템을 구축하고 분석결과를 통해 타당성을 증명한다. 마지막으로 5장에서는 결론 및 향후 과제에 대해서 기술한다.

2. 관련연구

최근까지 HTTP 트래픽을 분류하고자 하는 연구는 계속 진행되고 있다. 대부분은 HTTP 트래픽을 생성하는 어플리케이션에 대한 분석 연구이고, HTTP 서버 사이트를 분류해 주는 사이트가 있다. 어떤 종류의 어플리케이션 프로그램들이 HTTP 프로토콜을 사용하여 통신하는지 정확히 나열한 연구는 아직 없다. 하지만 HTTP 프로토콜을 사용하는 어플리케이션 프로그램들을 그룹 지어 구분하는 연구는 이루어지고 있다.

Wei lie의 논문[3]은 HTTP 패킷의 헤더 분석을 통해 기본적인 Web browsing 이외에 Web mail, Web application, file download, multimedia, messenger 들의 세부적인 항목으로 분류한다. 하지만 어플리케이션을 그룹으로 묶어서만 분류할 뿐, 어플리케이션별 분류가 이루어지지 않기 때문에 각 어플리케이션 프로그램별 제어를 적용하기는 적합하지 않다.

K. Kim의 논문[2]은 HTTP 트래픽의 내용을 분석하여 HTTP 프로토콜을 사용하는 응용을 분류한 논문이다. 이 논문은 HTTP 트래픽의 헤더가 아니라 페이로드의 내용을 분석하여, 애플리케이션을 쇼핑(shopping), 성인물(adult), 주식(stock), 커뮤니티(community), 게임, 음악, 영화, 웹메일, 교육, 뉴스&웹으로 총 10가지로 나누어 분석하고 있다. 논문 제목[2]에서 나타나듯이 일본어로 된 홈페이지에 대해서만 수행한 것이다. 각 분류 별로 10~15개의 키워드를 추출하여 사전을 만들고, 페이로드 중에서 일본어로 특정 키워드만 추출하여 사전의 키워드들과 비교 분석하여 HTTP 트래픽을 항목별로 나누었다. 키워드 매칭에 대해서는 HTML 페이지에서 첫 번째 나온 단어로 매칭한 경우와 가장 많이 사용된 키워드로 매칭된 경우 분류결과에 약간의 차이를 보였다. 그러나 페이로드 분석을 하여 키워드 기반으로 응용 프로그램을 분류하는 시도는 어느 정도의 정확도를 보여주고 있다. 하지만 시스템이 분류를 처리하는 오버헤드에 대한 언급이나 키워드 분석의 정확도에 대한 고려가 부족하여, 실시간 분류 시스템으로 개발하기는 어렵다.

또한 사이트별 분석 방법의 연구 중, 사이트에 대해 접속 순서를 보여주는 www.rankey.com[5]과 같은 사이트가 있지만, 등록한 사이트에 대한 순위만 제공하며, 특정 사용자의 접속 내용만을 이용하여 순위를 매기는 샘플링 기법을 쓰는 등 아직 미흡한 실정이다. 또한 의미 있는 분석이 되기위해서는 전체 사이트별 순서뿐 아니라 각 사이트 별로 세부적인 분석도 제공되어야 한다. 예를 들면, 다음 네이버의 경우 전체 네이버 사이트 접속 트래픽 양 분만 아니라 블로그, 메일, 뉴스, 검색 등 세부적으로 나누어 검색 할 필요가 있다.

본 연구에서는 사이트에서 발생하는 트래픽을 기준으로 분석하고, 각 사이트 중 세부분류가 필요한 사이트에 대해서는 세부분류가 가능한 시스템을 구현한다.

3. 시스템 설계 및 구현

본 장에서는 논문에서 제안하는 서버별(사이트 측면) 분류 기준과 분류된 사이트 중 Top10 리스트의 세부 분류 방법을 기술하고 시스템 설계와 구현에 대해 기술한다.

3.1 트래픽 모니터링 시스템: KU-MON 시스템

트래픽 모니터링 시스템으로는 MRTG, Ntop 등이 있다. MRTG는 표준 네트워크 관리 프로토콜인 SNMP를 이용하여 트래픽 모니터링 및 관리를 위해 사용되는 툴이다. 그래프를 포함한 웹페이지를 제공하여 모니터링의 용이성을 제공하지만 트래픽 양의 변화를 제공할 뿐, 상세한 트래픽 분석을 제공 하는 데는 부족함이 많다. Ntop은 flow를 이용하여 실시간으로 네트워크 트래픽을 모니터링해주는 시스템이다. 하지만 과거의 트래픽 정보를 확인할 수 없으므로 장기간의 트래픽을 분석하여 얻을 수 있는 트래픽 패턴 등의 정보를 얻을 수 없다.

본 논문에서는 위에 제시한 모니터링 시스템의 단점을 보안한 KU-MON 시스템[4]을 사용한다. KU-MON 시스템은 특정 호스트의 전체 대역폭에 대한 차지 비율, 어플리케이션별 트래픽 분류와 세부적인 정보를 제공하기 위해 네트워크의 각 호스트를 중점으로 한 처리에 유용한 flow 기반의

실시간 모니터링 시스템이다.

KU-MON의 시스템 구축환경은 다음 그림 1과 같다. 그림 1의 위치에서 Packet Collector 가 패킷을 수집한다. 수집된 패킷은 Flow Generator에서 5-tuple을 기준으로 해당 플로우를 생성한다. 그리고 Flow Identifier에서 플로우를 식별한다. 그리고 HTTP로 분석된 flow를 제안하는 분석 시스템에서 분석한다.

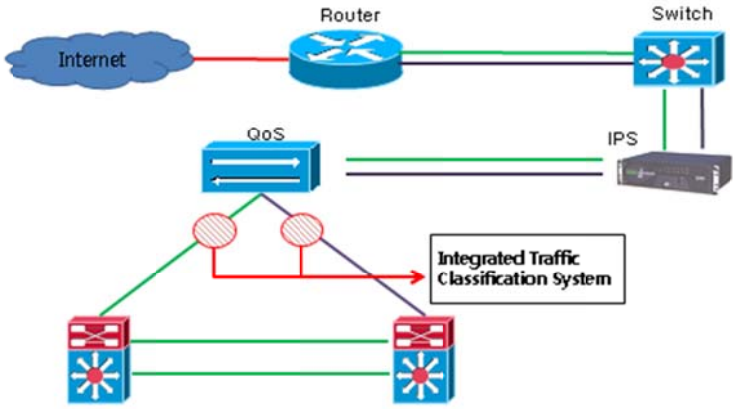

그림 1. 시스템 구축 환경

3.2 서버별(사이트 측면) 분석 알고리즘

본 장에서는 HTTP 트래픽의 서버별 분석 알고리즘과 서버에 따른 세부 분석인 호스트별 분석 알고리즘에 대해 기술한다.

3.2.1 도메인 추출방법

HTTP의 서버측에서 발생하는 트래픽을 확인하기 위해서는 HTTP의 헤더의 HOST 필드를 기반으로 분석한다. 전체적인 도메인의 구성은 기본적으로 다음과 같이 구성된다. 예를 들어, mail.naver.com의 경우 mail이 호스트 네임(서브도메인)이고, naver.com가 도메인네임이다. 호스트네임(서브도메인)과 도메인네임을 합쳐 전체도메인이라고 말한다. 전체도메인의 구성 절차[6]는 그림 2와 같다.

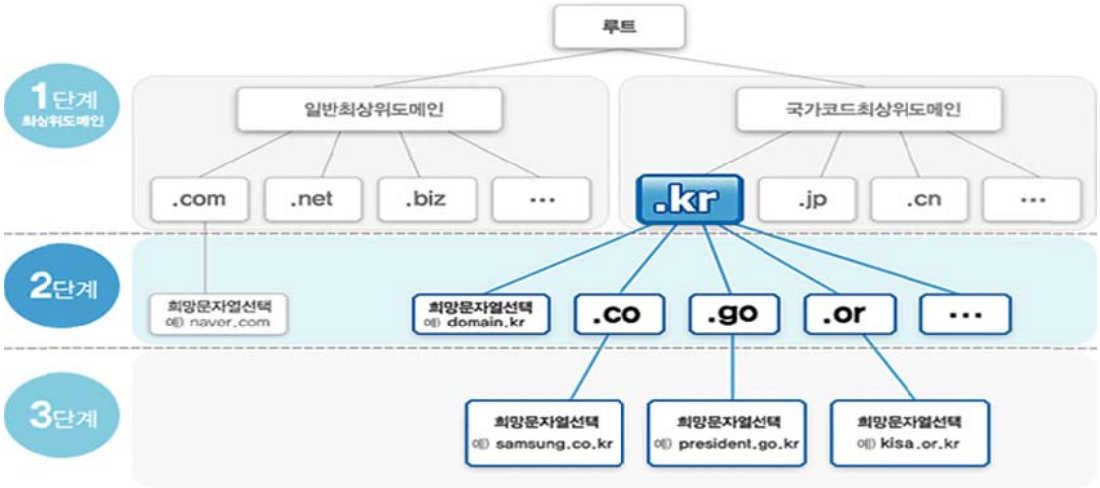

그림 2. 전체도메인 구성 절차

도메인의 구성은 일반최상위도메인과 국가코드최상위도메인을 선택하는 1단계의 선택을 하게 된다. 1단계 선택에서 일반 최상위 도메인에서는 .com .net .biz 등 사이트의 특성에 맞는 도메인을 선택하게 되고, 국가 코드최상위 도메인에서는 .kr .jp 와같이 국가별 선택을 하게 된다. 최상위 도메인의 선택한 후 naver, yahoo등 사이트가 원하는 희망 문자열을 통하여 도메인 네임을 선택 하게 된다. 문자열을 선택한 후 필요에 따라 호스트네임(서브도메인)을 선택하여 전체 도메인을 구성한다. 본 논문에서는 이러한 전체 호스트의 구성특성에 따라 호스트를 각각의 특성 별로

분석을 한다. 아래 그림 3은 KU-MON시스템의 flow_pkt_twoway 의 구조이다. 위에서 언급한 봐와 같이 HTTP의 요청 헤더의 HOST필드를 이용한다. HOST필드는 도메인 네인 시스템(DNS, Domain Name System)의 도메인 네임을 이용해 인터넷 호스트를 지정하는 필로서, HTTP 1.1에 포함 되어야 하는 유일한 필수 헤더이므로 서버 별 분석에 용이한 특성을 지니고 있다.

```
[redacted] : 49818 --  6 -- [redacted]   :    80 [ D]
                                                    > 53.78->53.79 [ 0.01sec] [     7p        611b] => [  0][SA FPU] [P H]
                                                    < 53.78->53.80 [ 0.02sec] [     9p       7967b] => [  0][SA FPU] [P H]
                                                    stored pkt : [forward= 1] [backward= 6]
maxmake cl= 7
53.786 : [redacted] : 49818 --  6 --> [redacted]   :    80 =>  [pkt_len:   221]  [data_len:   163 = 163   ]
GET /pino/update/updater.txt HTTP/1.1
If-Modified-Since: Mon, 15 Nov 2010 10:48:28 GMT
If-None-Match: "1380185-1bf0-2e26f700"
Host: pino.peeringportal.co.kr


53.791 : [redacted]   :    80 --  6 --> [redacted] : 49818 =>  [pkt_len:  1518]  [data_len:  1460 = 1460  ]
HTTP/1.1 200 OK
Date: Sun, 26 Jun 2011 13:40:59 GMT
Server: Apache/2.2.3 (CentOS)
Last-Modified: Mon, 15 Nov 2010 10:48:29 GMT
ETag: "1c70003-1bf0-2e363940"
Accept-Ranges: bytes
Content-Length: 7152
Connection: close
Content-Type: text/plain; charset=EUC-KR
```

그림 3. flow_pkt_twoway 구조

다음 그림 4는 HOST별 분석을 위한 절차이다. HOST의 구성을 반대로 구성하여 호스트별 분석을 수행하는데, 처음에 HTTP의 헤더필드에서 HOST를 추출한다. 이때 사이트의 필요에 따라서 많은 서브 도메인들이 생성 되기도 한다. 많은 서브도메인 통해 분류하고 저장함으로써 각각의 사이트의 세부적인 분석이 가능하다. 하지만 여러 개의 서브도메인을 저장할 경우 많은 저장공간이 필요 할 수가 있다. 본 분석시스템에서는 1차 서브도메인까지만 저장을 하고 이를 통해 각각의 사이트 별 특성을 분석을 수행한다.

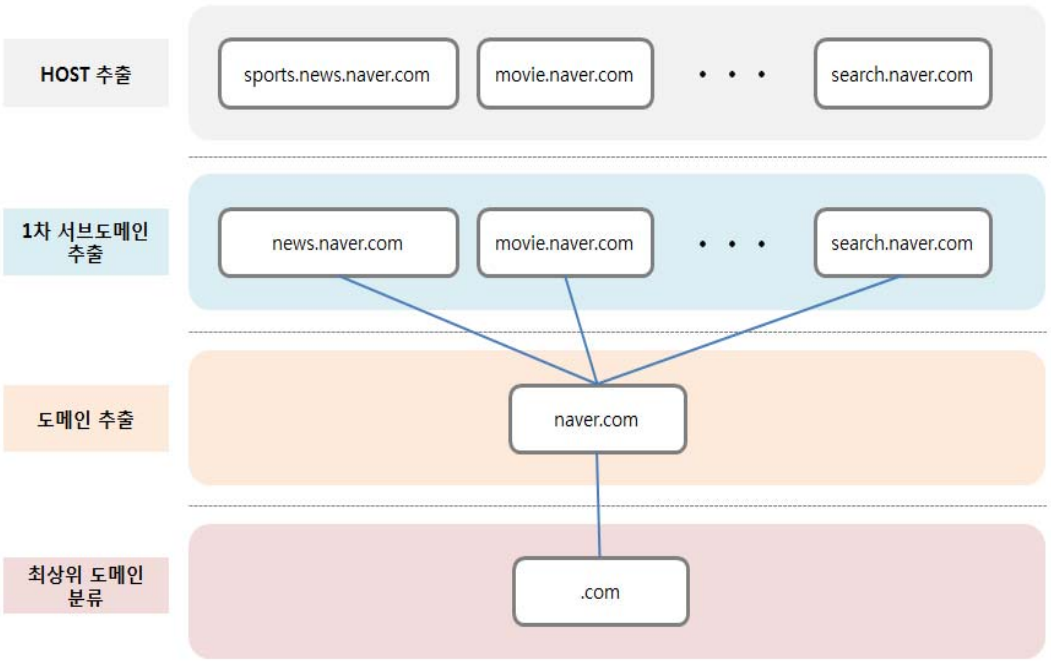

그림 4. HOST별 분석 절차

3.2.2 도메인에 따른 세부 분류

본 논문에서는 최초의 분석 결과로는 최상위도메인과 희망문자열(사이트네임)을 가지는 도메인네임으로 분류를 한다. 분석결과를 통하여 naver.com, daum.net에 접속하는 빈도와, 사용되는

트래픽에 대한 전반적인 모니터링이 가능하다. 하지만 naver.com, daum.net등 다양한 정보를 다루는 포탈 사이트의 경우 이러한 분석 만으로는 부족하다. 각각의 세부적인 특성을 알아야 할 필요가 있다. 각각의 세부적인 특성을 분석하기 위해서는 가지고 있는 서브도메인을 사용해야 한다. 아래 그림 5와 같이 naver.com 의 경우 호스트네임(서브도메인)을 통하여 세부적인 분석이 가능하다.

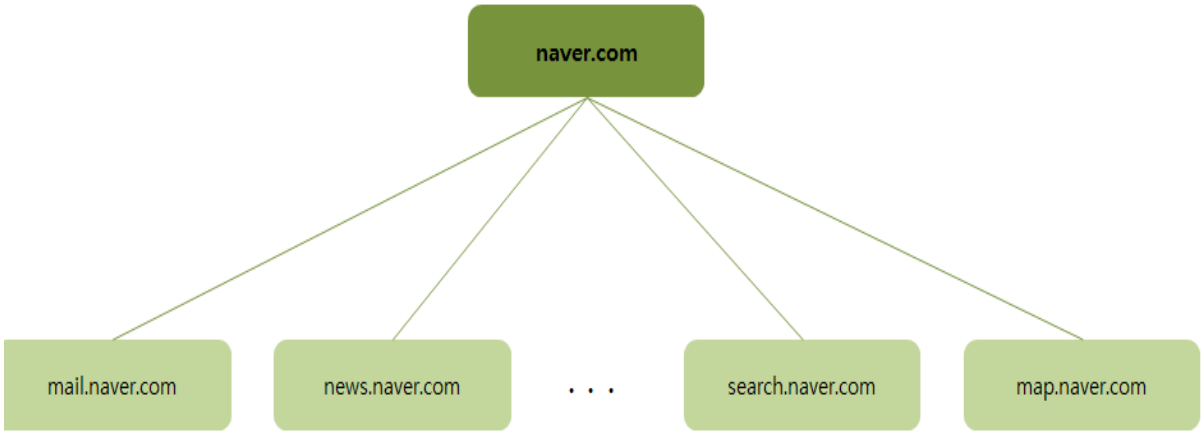

그림 5. naver.com의 세부적인 분류

예를 들면 naver.com 에는 그림 5에서 보는 것처럼 mail, news, search, map등 다양한 서비스들이 존재하고 있다. 그러므로 포탈사이트와 같은 다양한 서비스들을 다루는 사이트들에 대해서는 서브도메인을 통하여 각각의 특성 별 분석을 한다.

3.2.3 사이트 도메인네임 중복 처리

사이트별 분석 방법의 경우 naver.com, naver.net과 같이 동일한 사이트를 접근하지만 도메인의 이름은 다른 경우가 종종 나타난다. 위와 같은 경우 동일한 사이트이지만 도메인네임은 다른 주소를 갖고 있기 때문에 서로 다른 도메인으로 분류된다. 이와 같은 경우가 발생할 경우 헤더의 응답 정보를 통하여 분석을 한다.

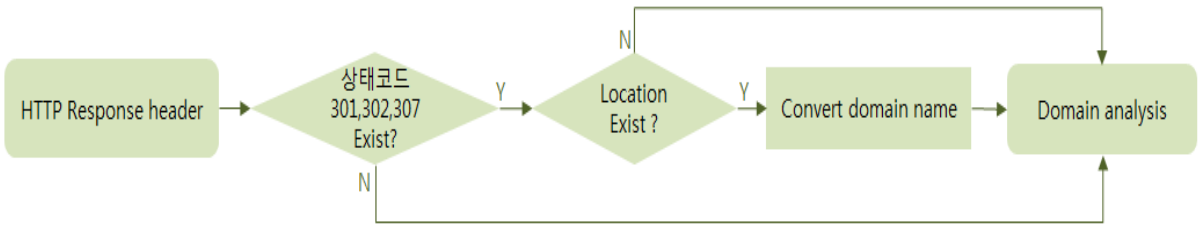

그림 6. 희망문자열의 중복처리

그림 6은 사이트 네임에 따른 중복 처리에 대한 절차를 나타낸다. HTTP Response 헤더의 상태코드 3xx 의미는 리다이렉션을 나타내어준다. 상태코드 301의 경우에는 요청한 자원이 새로운 URL로 영구히 이동한 것을 의미한다. 이 경우는 서버상의 파일이름을 변경하였거나 파일을 새로운 디렉토리로 이동했을 경우 처리할 때 사용하는 방법이고, 상태코드 302는 요청한 자원이 일시적으로 다른 URL을 사용하고 있다는 의미를 가진다. 그리고 마지막으로 307은 302와 같은 코드로서 이전 HTTP버전에서 발생했던 302와 관련된 혼란을 없애기 위해 만든 코드이다. 301,302,307과 같은 상태코드가 응답헤더에 존재 할 경우 Location 정보를 이용하여 도메인을 변화하여 분석을 수행한다.

아래의 그림 7은 naver.net에 대한 요청/응답에 대한 헤더의 정보이다. 클라이언트가 naver.net으로 접속을 요청하였을 경우 응답헤더에서는 302상태코드를 응답헤더로 전송하여 준다. 302상태코드는 요청한 자원을 일시적으로 다른 URL로 변경 시켜 준다. 이 때 변경되는 URL은 Location

의 정보인 www.naver.com 이다. 이를 통하여 naver.net 으로 요청된 HOST 정보를 www.naver.com 으로 변경하여 분석을 수행 하게 된다. 이를 통하여 naver.com 과 naver.net 이 동일한 사이트로 접속이 되지만, 서로 다른 도메인으로 분석을 하지 않음으로써, 분석의 정확도를 향상 시키고, 분석의 일관성을 유지한다.

```
GET / HTTP/1.1
Accept: text/html, application/xhtml+xml, */*
Accept-Language: ko-KR
User-Agent: Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0)
Accept-Encoding: gzip, deflate
Host: naver.net
Connection: Keep-Alive

HTTP/1.1 302 Found
Date: Sun, 13 Nov 2011 13:37:13 GMT
Server: Apache
Location: http://www.naver.com/
Vary: Accept-Encoding
Content-Encoding: gzip
Content-Length: 184
Connection: close
Content-Type: text/html; charset=iso-8859-1
```

그림 7. "naver.net" 에 대한 요청/응답 헤더

4. 분석결과

본 장에서는 위에서 제시한 분류 방법을 수집된 학내망 트래픽에 적용하여 분석한 결과를 제시한다. 표 1는 2011년 10월 19일 ~ 21일까지의 학내망 트래픽을 수집한 내용이다. 총 트래픽 중에 HTTP트래픽은 Flow는 약 23% 차지하고 있으며 Byte, Packet은 약 17% 정도를 차지 하고 있다.

		Day1	Day2	Day3
Flow	HTTP	12,198,258 **(22.54%)**	10,941,744 **(23.06%)**	8,123,946 **(24.94%)**
	Total	54,108,763	47,448,599	32,567,711
Byte	HTTP	759,977,352,296 **(16.08%)**	605,966,673,063 **(16.31%)**	409,208,149,494 **(20.02%)**
	Total	4,726,351,924,553	3,715,707,440,675	2,044,135,035,188
Packet	HTTP	863,349,035 **(16.78%)**	697,294,264 **(17.10%)**	477,189,069 **(20.81%)**
	Total	5,146,579,273	4,077,787,425	2,293,329,669

표 1. 2011년 10월 19일 ~ 21일까지의 학내망 트래픽

4.1 서버 별 분석결과

해당 트래픽들을 본 논문에서 제안하는 시스템에 적용하였을 경우 포털 사이트의 비중이 높게 차지 하고 있었다. 아래 그림 8은 도메인 별 분석결과 중 상위 10개의 분포도이다. naver,com, daum.net와 같은 포털 사이트가 많은 부분을 차지하고 있다. 포털 사이트가 많은 트래픽을 차지하는 이유로는 인터넷 익스플로러, 사파리, 모질라 등 대부분의 초기 사이트 설정이 포털 사이트로 설정되어 있다. 또한 학내망의 특성상 정보 검색이 많이 사용되기 때문에 포털 사이트의 비율이 많이 나타나고 있다. 그리고 korea.ac.kr이 2번째로 많은 비율을 차지하고 있다. 이 사이트는 학내망의 특성상 접속 빈도가 높게 나타나고 있다.

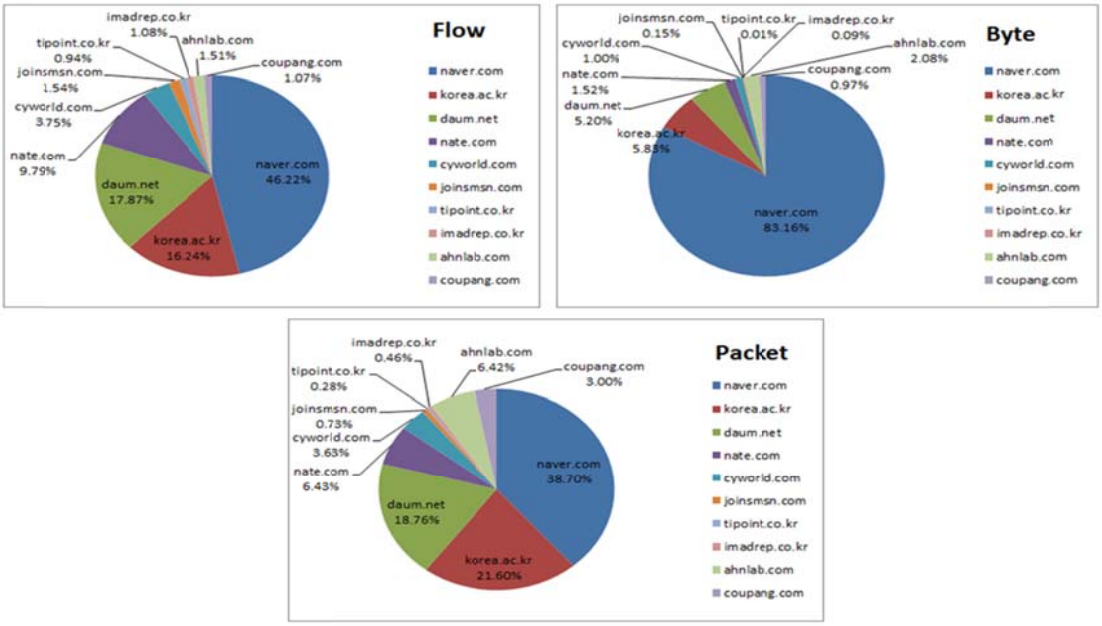

그림 8. Day1의 상위10개의 도메인의 비율

최근에 많이 사용되는 소셜커머스의 한 종류인 coupang.co.kr 의 경우 접속 뿐만 아니라 광고노출이 잦아서 많이 사용 된다. 그리고 cyworld.com 와 같은 커뮤니티 사이트 많이 사용되는 것을 확인 할 수 있었다. ahnlab.com 사이트도 보안프로그램의 특성상 많은 접속이 이루어 지고 있었다.

4.2 도메인의 세부분류

도메인을 기준으로 사이트 별 사용빈도 및 트래픽을 확인 할 수 있었다. 하지만 각각의 서비스를 좀 더 자세히 파악하기 위해서는 서브도메인을 통해, 각각의 서비스를 확인 할 필요가 있다.

	Sub_Domain	In_flow	Out_flow	In_byte	Out_byte	In_packet	Out_packet
1	search	97039 (10.78%)	90696 (15.03%)	347409095 (7.93%)	1846315237 (5.24%)	1769534 (8.18%)	2180897 (6.93%)
2	static	92087 (10.23%)	71841 (11.9%)	314743446 (7.19%)	1500143164 (4.25%)	1702188 (7.87%)	1938729 (6.16%)
3	www	52520 (5.83%)	20972 (3.47%)	138696544 (3.17%)	1085686286 (3.08%)	714321 (3.3%)	1040615 (3.31%)
4	imgshopping	49842 (5.54%)	27221 (4.51%)	95590163 (2.18%)	443220313 (1.26%)	650423 (3.01%)	683331 (2.17%)
5	cafe	49238 (5.47%)	19902 (3.3%)	230916602 (5.27%)	656771500 (1.86%)	619250 (2.86%)	788163 (2.5%)
6	imgnews	38867 (4.32%)	27834 (4.61%)	247522579 (5.65%)	1596252448 (4.53%)	1062217 (4.91%)	1480821 (4.7%)
7	lcs	37648 (4.18%)	25866 (4.29%)	130024676 (2.97%)	49584591 (0.14%)	362132 (1.67%)	356755 (1.13%)
8	blog	36140 (4.01%)	17767 (2.94%)	136019226 (3.11%)	624627170 (1.77%)	502894 (2.32%)	683677 (2.17%)
9	ad	27173 (3.02%)	18343 (3.04%)	78242600 (1.79%)	134714236 (0.38%)	271101 (1.25%)	280989 (0.89%)
10	nmv	26547 (2.95%)	21283 (3.53%)	72274496 (1.65%)	595721890 (1.69%)	491395 (2.27%)	623508 (1.98%)
11	sstatic	22532 (2.5%)	20430 (3.39%)	165319858 (3.77%)	1153039721 (3.27%)	880325 (4.07%)	1162629 (3.69%)
12	news	20804 (2.31%)	12685 (2.1%)	74959149 (1.71%)	1107366584 (3.14%)	528398 (2.44%)	872044 (2.77%)

그림 9. "naver.com" 의 세부분류

위의 그림 9는 naver.com 의 세부분류 결과이다. 호스트네임을 통하여 각각의 서비스 별로 분석을 수행을 할 수 있다. Search의 경우 검색, cafe의 경우 커뮤니티 그리고 ad 와 같은 경우는 광고로 각각의 서비스를 분류 할 수 있다. 그리고 나머지 호스트네임들 역시 호스트네임의 의미적인 판단으로 분석을 수행 할 수 있었다. 그러나 호스트네임을 통해서 의미적인 판단이 이루어지지 않은 경우에는 매뉴얼적인 방법을 통해 서비스의 종류를 파악하였다.

5. 결론 및 향후과제

본 논문에서는 서버측에서 발생하는 트래픽을 기준으로 분석을 하고 분석된 호스트네임을 통해 각각의 서비스 별로 분석을 수행하였다. 이를 통해 사이트별 접속 빈도 및 트래픽 양을 확인할 수 있었다. 그리고 사이트의 안에서도 어떤 서비스들이 많이 사용되는 지에 대한 분석이 가능하였다. 제안한 시스템에서 분석된 결과를 바탕으로 인터넷 광고주 입장에서는 접속빈도를 통해 좀 더 효율적인 광고도 가능할 뿐 아니라, 네트워크의 관리자에게는 스팸메일, 유해 사이트 차단 등 네트워크 관리에 적용하여 원활한 네트워크 서비스와 좀 더 안전한 네트워크 관리가 가능할 것으로 보인다.
향후 연구로는 호스트네임을 통하여 각각의 서비스별로 분석이 가능하지만 호스트네임의 의미적인 분석을 통해 판단할 수 없는 서비스들이 있다. 분석을 할 수 없는 호스트네임에 대해서 확인하여 분석을 정확성을 향상 시켜야 하며, 매뉴얼이 아닌 자동화는 연구도 필요하다. 그리고 www.rankey.com과 같이 사이트속성별(주제에 따른 그룹핑) 분류 방법을 적용하여 어느 속성에서 트래픽이 발생하는지에 대한 정보를 파악해야 한다. 그리고 이를 바탕으로 그룹핑 알고리즘을 통하여, 좀 더 쉽게 서버속성별 트래픽을 확인할 수 있는 연구가 필요하다.

참고 문헌

[1] 진창규, 김명섭, 최미정 “HTTP 응용 트래픽의 다차원 분석 방법", 한국통신학회 동계종합학술발표회, (KICS) 2011.
[2] K. Kim, B. Lee, T. Kwon, N. Ryo, K. Okamura, and Y. Lee, "Japanese Content classification of HTTP Traffic", DICOMO, Beppu, July 2009.
[3] Wei Li, Andrew W. Moore, Marco Canini, “Classifying HTTP traffic in the new age”, ACM SIGCOMM 2008, Poster, August 2008.
[4] 박진완, 박상훈, 김명섭, "Flow를 이용한 호스트 기반 트래픽 모니터링 및 분석", 통신학회 하계종합학술발표회, 라마다플라자호텔, 제주, July 2008, pp.197.
[5] 랭키닷컴, http://www.rankey.com
[6] Whois Search KISA, http://whois.nida.or.kr
[7] 최미정, 진창규, 김명섭, “HTTP 트래픽의 클라이언트측 어플리케이션별 분류”, 한국통신학회논문지 36권, 11호, 2011년 11월, pp. 1277~1284.

진 창 규

2011년 강원대학교 컴퓨터과학 학사 졸업

2011년 ~ 현재 강원대학교 컴퓨터과학과 석사과정

<관심분야> 트래픽 모니터링 및 분석, 네트워크 관리 및 보안

김 명 섭

1998년 포항공과대학교 전자계 산학과 학사

2000년 포항공과대학교 컴퓨터 공학과 석사

2004년 포항공과대학교 컴퓨터 공학과 박사

2004년~2006년 Post-Doc.Dept. of ECE, Univ. og Totonto, Canada

2006년 ~ 현재 고려대학교 컴퓨터정보학과 부교수

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 멀티미디어 네트워크

최 미 정

1998년 이화여자대학교 컴퓨터 공학과 학사

2000년 포항공과대학교 컴퓨터 공학과 석사

2004년 포항공과대학교 컴퓨터 공학과 박사

2004년 ~ 2005년 프랑스 INRIA 연구소 박사후 연구원

2005년 ~ 2006년 캐나다 워터루대학 컴퓨터과학부 박사후 연구원

2006년 ~ 2008년 포항공대 컴퓨터공학과연구 조교수

2008년 8월 ~ 현재 강원대학교 컴퓨터과학과 조교수

<관심분야> 트래픽 모니터링 및 분석, 미래 인터넷 자율 관리, M2M네트워크 및 서비스 관리