논문 15-40-03-14 The Journal of Korean Institute of Communications and Information Sciences '15-03 Vol.40 No.03
http://dx.doi.org/10.7840/kics.2015.40.3.541

학내 망 자원 효율화를 위한 빅 데이터 트래픽 분석

안 현 민◆, 이 수 강*, 심 규 석**, 김 익 한***, 진 서 훈°, 김 명 섭°

Big-Data Traffic Analysis for the Campus Network Resource Efficiency

Hyun-min An◆, Su-kang Lee*, Kyu-seok Sim**, Ik-han Kim***, Seo-hoon Jin°, Myung-Sup Kim°

요 약

급하게 일어나는 인터넷의 활성화는 그 어느 때보다 효율적인 엔터프라이즈 망 운영 방안을 필요로 하고 있다. 효율적인 망 운영을 위해서는 장기간의 트래픽 분석을 통해 망의 특성을 정확히 반영한 운영 정책 적용이 필요하다. 하지만 기존에는 급격하게 증가하는 장기간 트래픽 데이터의 처리가 불가능했고, 다양한 분석 결과를 낼 수 없는 단기간 분석만 이루어졌다. 최근 빅 데이터 분석 플랫폼과 도구의 개발로 인해 장기간 트래픽 분석이 가능하게 되었고, 이를 이용해 망의 특성을 정확히 반영할 수 있는 장기간 트래픽 분석을 통한 엔터프라이즈 망 자원 효율화 방안이 요구되고 있다. 본 논문에서는 엔터프라이즈 망에서 발생한 장기간의 트래픽을 수집하고 저장 및 관리하는 방안에 대해 제안한다. 또한 분류기준을 정의하였으며, 수집된 빅 데이터 트래픽을 각 분류 기준으로 분류한 뒤 다각적인 통계 분석을 통해 망 자원을 효율화 하는 방안을 제안한다. 제안하는 방법을 학내 망에 적용하여 실험하였으며, 통계 분석 결과 시간과 공간, 그리고 사용목적에 따라 Quality of Service(QoS)정책을 달리 적용해야 함을 확인하였다.

Key Words : Enterprise network, Big data traffic, Statistical analysis, Long-term traffic, Network policy

ABSTRACT

The importance of efficient enterprise network management has been emphasized continuously because of the rapid utilization of Internet in a limited resource environment. For the efficient network management, the management policy that reflects the characteristics of a specific network extracted from long-term traffic analysis is essential. However, the long-term traffic data could not be handled in the past and there was only simple analysis with the shot-term traffic data. However, as the big data analytics platforms are developed, the long-term traffic data can be analyzed easily. Recently, enterprise network resource efficiency through the long-term traffic analysis is required. In this paper, we propose the methods of collecting, storing and managing the long-term enterprise traffic data. We define several classification categories, and propose a novel network resource efficiency through the multidirectional statistical analysis of classified long-term traffic. The proposed method adopted to the campus network for the evaluation. The analysis results shows that, for the efficient enterprise network management, the QoS policy must be adopted in different rules that is tuned by time, space, and the purpose.

※ 본 연구는 BK21플러스사업(No.T1300572) 및 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단 차세대정보컴퓨팅기술개발사업(2010-0020728) 및 2012년 정부(교육과학기술부)의 재원으로 한국연구재단(2012R1A1A2007483)의 지원을 받아 수행되었습니다.
◆ First Author : Dept. of Computer and Information Science, Korea University. kusuk007@korea.ac.kr, 학생회원
° Corresponding Author : Dept. of Applied Statistics, Korea University. seohoon@korea.ac.kr, 정회원
° Corresponding Author : Dept. of Computer and Information Science, Korea University. tmskim@korea.ac.kr, 종신회원
* Dept. of Computer and Information Science, Korea University. sukanglee@korea.ac.kr, 학생회원
** Dept. of Computer and Information Science, Korea University. kusuk007@korea.ac.kr, 학생회원
*** Dept. of Applied Statistics, Korea University, ihregal@korea.ac.kr
논문번호 : KICS2014-12-500, Received December 29, 2014; Revised March 6, 2015; Accepted March 16, 2015

Ⅰ. 서 론

네트워크 관리의 목적은 네트워크 자원을 최대한 활용하여 사용자에게 목적에 맞는 서비스를 신속하게 제공하는 것이다. 이를 위해 네트워크 관리자들은 적절한 네트워크 정책을 수립하여 관리 대상 네트워크에 적용한다[1,2].

제한된 네트워크 환경하에서 급하게 일어나는 인터넷의 활성화는 그 어느 때보다 효율적인 엔터프라이즈 망 운영 방안을 필요로 하고 있다. 인터넷 활용도가 높아지고 있으며 인터넷에 의존하는 업무/학습 등이 증가함에 따라 서비스 품질을 보장하는 QoS(Quality of Service) 정책의 중요성이 강조되고 있다. 하지만 업무 혹은 학습과 무관한 응용에 관련된 트래픽으로 인해 서비스 불만과 대역폭 확장에 대한 요구가 발생하고 있다. 악의적 목적으로 인해 발생되는 비정상 트래픽의 양도 꾸준히 증가하여 네트워크 자원 가용률을 감소시키고 있다. 예를 들어, 현재 본교에서는 원활한 네트워크 서비스를 지원하기 위해 방화벽, IDS / IPS (Intrusion Detection / Prevention System) 등과 같은 고가의 관제 시스템을 운영하고 있으며 이를 유지 보수하기 위해 많은 비용이 지출되고 있다. 하지만 실제 구성원들이 느끼는 질적 향상은 미비한 실정이다. 이러한 문제에는 무조건적인 고가의 장비 충당이 아닌, 근본적인 해결책이 필요하다. 망 트래픽에 대한 다각적인 분석을 통해 망에 특화 된 효율적인 QoS정책을 추진해야 하며, 설치된 장비의 운영률과 효율성을 판단하여 향후 네트워크 장비 확충 시 객관적인 근거 자료를 마련해야 한다.

정확한 트래픽 분석을 위해서는 장기간 망에서 발생한 모든 구성원들의 빅 데이터 트래픽을 분석할 필요가 있다. 엔터프라이즈 망의 특성 상 그 사용자 군(학내망의 경우 학생, 연구자, 교직원)이 뚜렷하게 구분되며 사용자군별로 네트워크 사용 목적이 다르기 때문이다. 때문에 사용자군별 트래픽 특징을 반영한 분석이 필요하다. 또한, 엔터프라이즈 망은 일정에 따라 독특한 트래픽 발생 형태를 가지므로 최소한 1년 이상의 트래픽을 수집하여 분석해야 정확한 결과를 도출 할 수 있다. 이를 위해 장기간의 트래픽을 수집 및 저장, 분석하기 위해서는 빅 데이터 기반의 파일 시스템과 이에 특성화 된 분석 알고리즘의 적용이 필요하다. 대상이 되는 엔터프라이즈 망의 사용자군별, 사용 목적별, 특정 일정별 등 다각적인 분석을 통해 해당 망에 대해 특화 된 QoS정책을 수립 할 수 있고, 이를 통해 망 관리 자원을 최적화 시킬 수 있다.

이에 본 연구에서는 장기간의 트래픽을 수집하여 기초 데이터를 생성하고, 빅 데이터 기반의 고급 통계분석을 통해 학내 네트워크 자원의 효율화 방안을 도출하는 방법을 제안한다. 제안하는 방법을 학내 망에 적용하였으며, 학내 망에서 3년간 발생한 트래픽을 통해 실험하였다.

본 논문의 구성은 다음과 같다. 서론에 이어, 2장에서 관련 연구에 대해 조사하고, 3장에서 제안하는 빅 데이터 트래픽의 수집 및 통계 기반 분석의 기초 데이터 생성 방안에 대해 기술한다. 4장에서는 제안하는 데이터 저장 및 관리 방안에 대해 기술하고, 5장에서 통계 기반 트래픽 분석과 그 결과에 따른 학내 망 자원 효율화 방안에 대해 기술한다. 마지막으로 6장에서 결론과 향후 연구에 대해 언급한다.

Ⅱ. 관련 연구

현재의 인터넷은 방송과 통신의 융합, 다양한 유·무선 네트워크의 결합, 다양한 서비스 및 응용 프로그램들의 개발, 사용자 요구사항의 다양화로 그 사용량이 폭발적으로 증가하고 있으며, 이러한 요구를 수용하기 위하여 현재의 인터넷은 고속화되고 있다. 이로부터 발생하는 트래픽의 양은 기하급수적으로 증가하는 추세이다. 이러한 네트워크 트래픽의 급격한 증가는 전통적인 트래픽 분석 방법으로 처리하기 어렵게 하고 응용 판별과 같은 단순한 분석만이 가능하였다[3-5]. 이러한 결과는 네트워크 관리를 위한 충분한 정보를 제공할 수 없는 상황이다[6].

트래픽 분석은 현재까지 고속 네트워크에서 발생하는 대용량의 트래픽에 대한 실시간 처리, 장시간 저장, 다양한 분석 등과 같은 기능들을 포함하는 종합적인 분석이 미흡한 실정이다. 현재의 트래픽 분석 시스템들이 장기간의 대용량 트래픽에 대한 다각적인 분석을 하지 못하는 이유는 크게 세 가지로 정리할 수 있다. 대용량의 장기간 트래픽 데이터에 대한 효과적인 저장 방법의 부재와 개별적으로 개발되고 구축된 다양한 분석 방법론들의 효과적인 통합 방법의 부재, 그리고 장기간에 걸쳐 축적된 대용량의 트래픽 데이터로부터 다양한 분석 정보의 효율적인 추출방법의 부재가 그것이다. 트래픽 분석은 분석 관점에 따라 크게 경향 분석(Trend Analysis), 집단 분석(Point Analysis), 계층 분석(Layer Analysis), 장애 분석(Event Analysis)의 네 가지로 나눌 수 있다.

경향 분석은 시간대를 기준으로 트래픽의 추이를 분석한다. 대표적인 분석 시스템으로는 MRTG(Multi

Router Traffic Grapher)를 이용한 시간대 별 추이 분석 시스템[7]이 있다. SNMP(Simple Network Management Protocol)를 이용하여 네트워크 장비로부터 트래픽 정보를 수집하여 RRD(Round-Robin DB)[8]의 형태로 테이블을 구축하여 시간대별 트래픽 정보를 제공한다. RRD를 이용하여 시간의 흐름에 관계없이 저장 공간의 크기를 일정하게 유지하면서 다양한 시간 추이 그래프(일, 주, 월, 연간)를 제공한다. 그러나 트래픽의 총량에 대한 분석 정보만을 제공하며 집단 분석을 위한 호스트, 서브넷 별 분석 정보와 계층 분석을 위한 응용, 프로토콜 별 분석 정보를 제공하지 못하기 때문에 구간 별 트래픽 특성을 파악하기 어려운 단점이 존재한다.

집단 분석은 공간을 기준으로 트래픽의 추이를 분석한다. 일정 주기 동안의 트래픽 통계 수치와 그래프, 다운로드와 업로드 트래픽이 높은 호스트 별, 서브넷 별로 바이트, 패킷, 플로우의 양과 백분율로 제공한다. 또한 특정 호스트에 대한 관리를 위하여 해당 호스트에 대한 자세한 트래픽 추이 정보도 제공한다. 이러한 분석 시스템은 호스트, 서브넷, 빌딩을 기준으로 정보를 제공하기 때문에 집단 분석에 강점을 보인다. 하지만 Point 별로 발생하는 트래픽의 양에 대한 분석 정보만을 제공하기 때문에 발생하는 트래픽이 어떤 응용에 의해서 발생하는지, 혹은 정상/비정상 트래픽에 의해서 발생하는 트래픽인지에 대한 분석은 불가능하다.

계층 분석은 트래픽을 구성하는 네트워크 계층별로 트래픽의 특성을 분석한다. 대표적인 분석 시스템인 Bro[9], nTop[10] 등은 트래픽을 응용 계층에서 분석하여 네트워크 내에서 발생하는 응용 프로그램의 종류를 분석한다. 응용 레벨 분석을 통해 특정 링크에서 발생하는 모든 트래픽을 응용 프로그램 또는 프로토콜 단위로 분석하고, 응용별 트래픽 발생 현황을 시간대 별로 제공한다. 이러한 분석 시스템은 응용 프로그램 및 프로토콜 별 정보를 제공하여 트래픽의 특성을 파악하는 Layer Analysis에 대해 강점을 갖지만, 호스트 또는 구간 별로 발생하는 트래픽의 특성을 분석하기 위한 정보를 제공하지 못하며, 비정상 트래픽도 응용 계층의 트래픽으로 분석하기 때문에 비정상 트래픽에 의해서 발생하는 트래픽의 특성을 반영하지 못하는 단점이 존재한다.

장애 분석은 비정상 트래픽과 같이 특정한 event에 의해 발생하는 네트워크 트래픽을 분석한다. 대표적인 event 분석 시스템인 Snort[11]는 침입 탐지 규칙을 기반으로 오버플로우, 포트스캔, CGI공격 등의 다양한 공격의 탐지가 가능하다. 또한, 각종 이벤트에 대하여 로그 정보와 통계 정보를 제공하여 비정상 트래픽에 대한 발생 현황 파악이 가능하다. Snort와 같은 장애 분석 시스템은 사전에 정의된 공격 트래픽에 대한 현황 분석 정보를 제공할 수 있지만, 패킷의 수집과 분석이 동시에 이루어져야하기 때문에 분석 시스템의 부하가 커 event에 대한 분석 정보를 로그 형태로만 제공하며 장기간 트래픽에 대한 분석 추이를 제공하지 못하는 단점이 존재한다.

본 연구에서는 언급된 빅 데이터 트래픽 분석의 단점과 개별 관점으로의 분석의 단점을 해결하기 위해 수집된 빅 데이터 트래픽을 4가지 분석 관점 모두로 분석하고, 그 결과의 통계 분석을 통해 망에 특화된 자원 효율화 방안을 도출하였다.

Ⅲ. 빅 데이터 트래픽의 수집 및 기초 데이터 생성

본 장에서는 제안하는 빅 데이터 트래픽의 수집 및 기초 데이터 생성 방안에 대해 기술한다.

3.1 빅 데이터 트래픽 수집

효과적인 트래픽 분석을 위해서는 망 구성원의 개인 정보를 침해하지 않는 선에서 사용 가능한 트래픽 정보들을 최대한 사용해야 한다. 본 연구에서는 5-tuple(Source IP, Destination IP, Source Port, Destination Port, Protocol)이 고유한 양방향 패킷들을 집합시킨 플로우를 기초 데이터 단위로 하여, 플로우가 가지고 있는 주소 정보와 시간 정보, 단방향 플로우 별로 패킷 양과 바이트 양에 대한 정보와 기타 플래그 정보들을 사용하였다.

효과적인 망 자원 효율화 방안 모색을 위해서는 망 사용자들의 특성과 사용 패턴 등을 정확하게 분석해야 한다. 하지만 망 사용자들의 모든 트래픽을 수집하기 위해서 분석 대상 망과 인터넷 망을 연결하는 라우터에서 트래픽을 수집하게 되면 분석 가치는 없으나 분석하기 위한 오버헤드만을 늘리게 되는 트래픽들 또한 수집하게 된다. 이들을 제외해야 정확성 있게 망 사용자들의 특성과 사용 패턴을 모형화 할 수 있다. 때문에 수집된 트래픽을 분석 대상과 비대상으로 구분하는 전처리 과정이 필요하다. 본 연구에서는 트래픽 수집의 온전성, 사용 프로토콜, 통신 대역과 호스트의 활동 여부를 기준으로 각각 분석 대상 트래픽을 정의하고 비대상 트래픽을 제외하는 전처리 단계를 거쳐 기초 데이터를 생성하였다.

표 1은 실험 대상인 학내 망에서 수집된 전체 트래픽과 비정상 트래픽, 즉 전처리 대상 트래픽의 양을 나

표 1. 트래픽 카테고리와 용량
Table 1. The Traffic Category and The Volume

Traffic	Flow (10^6)	Packet (10^6)	Byte (10^9)
1. Total	45,844	3,091,318	2,050,787
2. Partially collected traffic	31 (10^1)	12,674 (10^1)	1,116 (10^3)
3. Traffic using minority protocol	2,150	879,020	77,432
4. Communication with special-use IP	7,670	745,126	428,192
5. Traffic of internal network hosts	640	29,044	17,789
6. Traffic of external network hosts	5,615	1,127,043	298,534
7. Traffic of inactive hosts	513	6,970	1,481
8. Target traffic	35,383	1,842,478	1,554,267

타낸다. 전체 트래픽은 학내 망과 인터넷 망을 연결하는 최상위 라우터에서 미러링을 통해 수집하였다.

1번 전체트래픽에서 2번부터 7번까지 트래픽을 제외한 8번이 분석 대상 트래픽이다. 2번의 트래픽 수집 관련 전처리는, 온전히 수집되지 않은 플로우를 제거하는 과정이다. TCP 트래픽 중 3-way handshake과정의 패킷들과 3/4-way handshake 과정의 패킷들이 온전히 수집되지 않은 플로우들을 제거한다. 2번 관련 플로우는 매우 적은 양을 포함하기 때문에 타 값들과는 다른 자릿수를 나타낸다. 3번의 프로토콜 관련 전처리는, 네트워크 계층에서 IP 프로토콜과 전송 계층에서 TCP, UDP를 사용하지 않는 플로우들을 제거하는 과정이다. 이는 사용자의 의지가 들어있지 않은 트래픽이나 너무 적은 발생으로 인해 분석 가치가 없는 트래픽을 제외하기 위한 것이다. 4, 5, 6번은 통신 대역에 관련된 전처리이다. 외부 망과 외부 망과의 통신의 경우엔 분석 범주를 벗어난 것으로, 수집에서의 문제가 있던 것이며 내부 망과 내부 망과의 통신 트래픽은 그 양도 매우 적을뿐더러 망 자원에 가하는 부담이 거의 없기 때문에 분석 시간과 노력에 대비해 분석 효율이 미미하다. 그리고 다른 목적을 위해 예약된 IP 대역과 통신하는 트래픽을 제외한다. 예약된 IP 대역으로는 Loopback 전용 IP 대역, 내부 망과 외부 망 간 통신이 가능한 무선 IP대역이 아닌 개인용 무선 IP 대역, 그리고 Multicast를 위해 전송되는 IP 대역을 의미한다. 이는 사용자의 사용 의도와 상관없이 발생하는 트래픽을 제거하여 효과적으로 학내 망 자원 효율화 방안을 도출하기 위함이다. 7번의 호스트 활동 관련 전처리는 실제 사용자가 사용하고 있지 않은 단말에서 발생하는 트래픽을 제거하는 과정이다. 사용자에게 할당되어있는 IP 이더라도 사용자에 따라 활동이 없는 경우가 존재한다. 주말동안, 혹은 휴가 기간 동안에는 해당 호스트에서 네트워크 활동이 전혀 없으나 실행중인 프로그램에서 기본적으로 발생하는 트래픽들이 있으며, 해당 단말이 종료되어 있더라도 포트스캔 등의 공격으로 인해 발생하는 트래픽도 존재한다. 이러한 트래픽들은 분석에 드는 시간과 노력에 대비해 효율이 좋지 않으며, 무엇보다 사용자들의 사용 패턴에 맞춰 망 자원 효율화 방안을 연구하는 본 연구의 목표에 부합하지 않는다. 2번부터 7번까지는 중복되는 데이터가 있을 수 있다.

표 2. 분석 대상 트래픽
Table 2. The Target Traffic

Traffic	Flow (10^6)	Packet (10^6)	Byte (10^9)
Wired Traffic	28,311	1,408,982	1,162,839
Wireless Traffic	7,072	433,496	391,428

표 2는 분석 대상 트래픽을 유무선으로 나누어 그 양을 나타낸 것이다. 비정상 트래픽을 제한 분석 대상 트래픽은 전체 트래픽에서 플로우 기준으로 총 77%이며 패킷, 바이트 기준으로는 총 60%와 76%를 차지한다. 분석 목표에 맞지 않은 이들을 제함으로써 분석 효율과 타당한 결과를 얻을 수 있다. 트래픽은 실험 대상 망인 학내 망에서 2011-01-01부터 2013-12-31일까지 수집한 트래픽이다.

3.2 기초 데이터 생성

효과적인 망 자원 효율화 방안을 모색하기 위해서는 트래픽을 다양한 분류 속성으로 분류한 뒤 분류 결과와 트래픽 정보를 토대로 한 통계 분석을 통해 망에 특화 된 정책 도출이 필요하다. 이를 위해서 먼저 트래픽을 각 서비스와 메타 데이터를 기준으로 분류하였다. 트래픽의 서비스 분류는 여러 응용 트래픽 분류 방법론 중, 개인 정보 침해를 최소로 하며 높은 정확도와 분석률을 가지고 트래픽을 분류 할 수 있도록 헤더 정보 기반의 분류 방법론과 트래픽 상관관계 기반의 분류 방법론을 사용하였다.

그림 1은 실험 대상인 학내 망 트래픽의 여러 분류 속성을 나타낸다. 분류 속성은 트래픽을 분류하는 기본

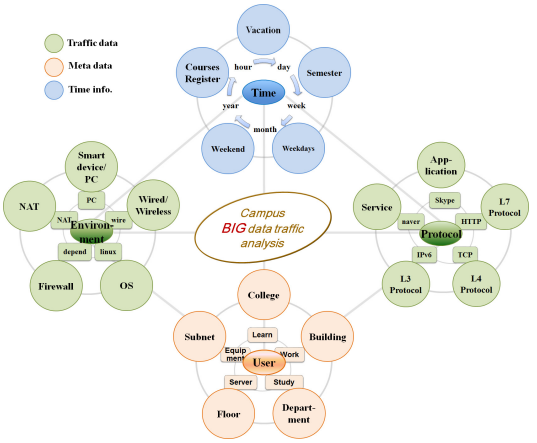

그림 1. 분류 카테고리
Fig. 1. The Classification Categories

단위로, 엔터프라이즈 망에서의 트래픽 분류 속성은 정의하기에 따라 매우 다양하게 존재한다. 본 연구에서는, 학내 망 트래픽 분석에 중요하게 사용될 수 있는 속성들인 학사 일정, 사용 프로토콜과 사용 서비스, 단말의 접속 환경과 사용자 정보 카테고리 별로 분류 속성을 정의하여 트래픽 분류에 적용했다. 트래픽 분류에서 중요한 문제로 분류 범위가 있다. 망 사용자들의 개인정보 보호를 위해 페이로드 정보를 제외한 헤더 정보만을 분류 범위로 삼아야 한다. 본 연구에서는 트래픽의 헤더 정보와 상관관계에 기반을 두는 분류 수행하였고 이 때문에 100%의 정확도와 100%의 분석률을 보장하지는 못하였다. 95%이상의 정확도가 검증 된 시그니쳐를 사용하였으며, 이를 이용해 80% 가량의 분석률을 달성하였다. 정확도를 저해시키는 원인은 동적 포트나 임의의 포트 사용이 가능한 P2P 응용이다. 해당 응용 프로그램을 이용하면 사용자가 임의의 포트 번호를 적용시킬 수가 있는데, 사용자가 적용시킨 포트 번호가 1000번대 이하인 잘 알려진 포트 번호와 겹칠 수 있다. 이러한 경우, 일반 사용자들만으로 통신이 이루어져 서버와 클라이언트의 구분이 확실치 않고, 서버의 IP정보도 적용할 수 없는 P2P 응용에 적용되는 헤더 정보 기반 분류의 정확도는 낮아질 가능성이 있다. 때문에 본 연구에서는 헤더 정보기반의 분류를 적용하기 전, 트래픽 상관관계에 기반하여 BitTorrent 트래픽 분류를 먼저 수행한다. 이후 헤더 정보를 이용하여 트래픽들을 분류하고, 마지막으로 아프리카TV 트래픽 분류를 위해 BitTorrent와는 다른 방법의 트래픽 상관관계 기반 분류를 수행한다. 이러한 트래픽 분류 순서는 분류의 정확도를 최우선으로 한다.

Ⅳ. 데이터 저장·관리 및 분석

본 장에서는 3장에서 생성된 기초 데이터를 저장·관리하고 분석하기 위해 제안하는 시스템 구성과 스키마 구성에 대해 기술한다.

4.1 빅 데이터 트래픽 저장·관리 방안

빅 데이터 분석은 기존 데이터베이스 관리 도구가 수집, 저장, 분석 및 관리 할 수 있는 범위를 넘어서는 데이터셋으로부터 가치를 추출하고 결과를 분석하는 기술이다[12]. 따라서 빅 데이터 기반의 분석을 제공하는 플랫폼이 필요하다. 본 연구에서는 빅 데이터 트래픽을 저장하고 관리하기 위한 플랫폼으로 대량의 자료를 처리할 수 있는 컴퓨터 클러스터를 통해 분산 응용 프로그램을 지원하는 하둡(Hadoop) 사용을 제안한다[13]. 하둡은 분산 저장소와 연산 기능을 모두 제공하는 플랫폼이다. 하둡은 데이터 저장을 위한 하둡 분산 파일 시스템(Hadoop Distributed File System: HDFS)[14]으로 구성된 마스터-슬레이브 아키텍처(마스터-데이터 노드)와 연산을 위한 맵리듀스[15]로 이루어진다. 하둡은 기본적으로 대용량 데이터셋의 데이터 파티셔닝(분할)과 병렬 처리에 맞게끔 설계됐다. 하둡의 저장 공간과 연산 능력은 하둡 클러스터에 호스트를 추가함에 따라 늘어나고, 수천 개의 호스트를 클러스터에 추가해 페타바이트 크기의 데이터까지 처리할 수 있다. 저렴한 스케일 아웃이 가능하며 분산 처리를 통해 빠른 분석이 가능하고, 단순한 데이터 모델과 오프라인 배치 프로세싱에 최적화 되어 있는 하둡은 장기간의 빅 데이터 트래픽을 분석하기 위한 최적의 플랫폼이다.

하지만 하둡에는 몇 가지 단점이 존재한다. 첫 번째는 보안에서의 단점이다. 기본적으로 하둡에 존재하는 유일한 보안 기능은 HDFS 파일과 디렉터리 레벨의 소유권 및 권한뿐이다. 악의적인 사용자가 다른 사람의 신원을 훔쳐 하둡 시스템에 침입 할 수 있는 것이다. 때문에 하둡 시스템이 설치 된 리눅스 머신에서 보안을 책임져야 한다. 권한이 있는 IP에서의 접속만을 허가하며, 사용자 계정 또한 정해진 인원에게만 할당해야 한다. 두 번째는 HDFS가 가지는 단점이다. HDFS는 작은 파일을 효율적으로 처리하지 못하며 투명한 압축 방식이 존재하지 않는다. 따라서 파일의 기본 크기를 충분하게 하여야 하며, 본 연구에서는 512MB 이상의 기본 파일 크기를 제안한다. 세 번째는 맵리듀스가 가지는 단점이다. 맵리듀스는 배치 기반의 아키텍처로써, 실시간 데이터 접근이 필요한 사례에는 적합하지 않다. 맵리듀스는 무공유 아키텍처이며, 전역 동기화나 수정

가능 데이터의 공유 같은 작업을 사용할 수 없다. 하지만 망 자원 효율화 방안 모색을 위한 빅 데이터 트래픽의 통계 분석은 실시간성을 요하지 않으며, 전역 동기화 혹은 데이터 공유를 요구하지 않는다. 맵리듀스에는 또한 프로그래밍 초보가 접하기에는 진입장벽이 높다는 단점이 있다. 양질의 분석을 위해서는 통계 전공자들의 직접적인 분석이 필요한데 자바 기반의 맵리듀스 언어는 비 컴퓨터 전공자들이 접근하기가 쉽지 않다. 때문에 분석가들이 맵리듀스 언어를 직접적으로 사용하지 않도록 하이브를 사용하여야 한다. 마지막으로 하둡 생태계의 버전 호환성 문제가 있다. 향후 상위 버전들이 계속 나오겠지만, 현재로서는 하둡 1.1버전과 하이브 0.10버전을 사용하면 하둡과 하이브, 그리고 Rhive의 연동이 자유롭다.

본 연구에서는 또한 분류 된 트래픽을 분석하기 위한 통계 분석 툴로써 R을 사용한다. R은 데이터 분석 및 결과를 그래픽으로 표현하기 위한 통계 프로그래밍 언어로써 R을 활용하면 통계적 분석 및 예측적 분석, 데이터 마이닝, 데이터의 시각화를 수행할 수 있다. 전통적으로 통계 분야에서 많이 사용되어온 R은 금융, 생명 과학, 제조업, 소매업 등 다양한 분야 에서 응용할 수 있다. 하지만 R은 단일 노드에서 동작한다는 점에서 빅 데이터를 분석해내지 못한다는 단점이 있다. 때문에 R은 클러스터를 통해 빅 데이터를 분석해낼 수 있지만 고급 통계 분석 방법을 적용하지 못한다는 단점을 가지고 있는 하둡과는 상호보완적이다. 고급 통계 분석이 가능한 R을 하둡과 연동시켜 빅 데이터를 분석하는 방안에 대한 연구는 많이 이루어졌으며 그 연동 방안 또한 많이 제안되었다. 그 중 RHive를 통해 R과 하둡을 연동시키는 것이 가장 쉽고 빠른 방법이다. 표 3은 하둡과 R을 연동시키는 방안 중, RHive와 다른 방안을 비교한 것이다.

표 3. 하둡과 R의 상호연동 방법
Table 3. The Hadoop-R Interworking Scheme

Method	RHive	RHipe, RHadoop
R Basic Function	Provided	Need to implement by programmer
SQL	SQL based HiveQL provided	Not provided
MapReduce Abstraction	Help programmer with abstraction framework	Programmer need to understand about MapReduce
Distributed Programming	Possible without MapReduce knowledge	To do, MapReduce knowledge required

4.2 시스템 및 스키마 구성

제안하는 빅 데이터 트래픽 저장 및 관리 시스템은 그림 2처럼 기초 데이터 생성부와 하둡 클러스터 부분으로 시스템을 나눌 수 있다. HDFS 클라이언트는 하둡 클러스터 외부에 설치하는 것이 일반적이지만 하둡의 네임노드에 설치하여 사용할 수도 있다.

기초 데이터 생성부는 하둡 클러스터 외부에서 수집된 트래픽을 분류한 뒤 하둡으로 데이터를 입력하는 모듈이다. 데이터는 HDFS 클라이언트와의 연동을 통해 하둡 클러스터에 저장된다. 실제 데이터는 하둡의 데이터 노드들에 저장되고, 네임노드는 이들에 대한 관리를 수행한다. HDFS 클라이언트에는 하둡 생태계의 여러 프로젝트들이 설치될 수 있으며, 분석가는 HDFS 클라이언트에 접속하여 하둡에 저장되어 있는 데이터들을 분석한다.

제안하는 시스템에서 하둡 클러스터에 저장한 데이터를 관리하는 프로그램은 하둡 생태계 프로젝트 중 하나에서 개발된 하이브(Hive)이다. 하이브는 RDBMS 같은 인터페이스를 제공하여 사용자들이 쉽게 다가갈

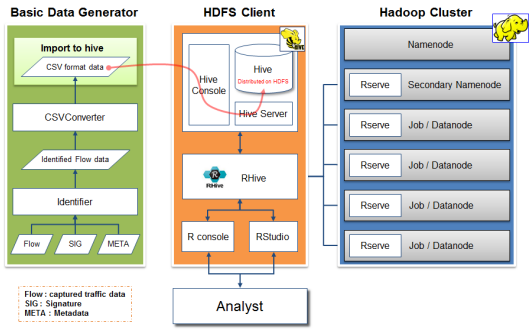

그림 2. 시스템 구조
Fig. 2. The System Structure

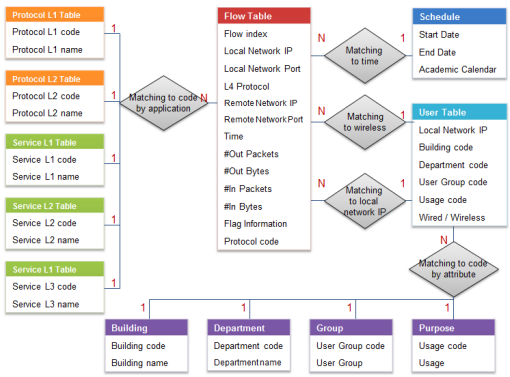

그림 3. 하이브 테이블의 E-R 다이어그램
Fig. 3. The E-R Diagram of Hive Tables

수 있도록 개발되었다. 하이브를 효과적으로 이용하기 위해 그림 3에서 보는 바와 같이 12종류의 테이블을 포함하는 E-R 다이어그램 기반의 스키마를 구성하였다. 실제 통계 분석에는 RHive를 통해 R을 사용하여 하이브와 연동하였다.

V. 통계 기반 빅 데이터 트래픽 분석

본 장에서는 빅 데이터 트래픽간의 관계·패턴·규칙 등을 찾아내고 모형화하여 관리자의 의사 결정을 돕는 통계 기반 트래픽 분석 방법에 대해 기술한다. 그리고 그 결과로 인해 도출 된 망 자원 효율화 방안에 대해 기술한다. 실험은 학내 망의 특징에 맞춰 학사 일정과 학습에 무관한 트래픽을 대상으로 이루어졌다.

5.1 학내 망 트래픽의 시공간적 사용 행태

본 절에서는 특정 시공간과 관련하여 학내 망 자원 효율화 방안 제시를 목적으로 한 실험에 대해 기술한다. 이를 위해 교내에서 발생하는 유선 학내 망 네트워크 트래픽의 시공간적 분포를 확인하고 분석하였다.

그림 4는 수강신청기간과 비수강신청기간 동안 각 건물 별로 발생한 한 시간 단위의 트래픽 양의 평균을 막대그래프로 나타낸 것이다. 가로축은 각 건물을 나타내며 세로축은 트래픽 발생량(GB 단위)을 나타낸다. 건물별로 좌측의 막대가 수강신청기간의 트래픽 발생량, 우측의 막대가 비수강신청기간의 트래픽 발생량이다. 비수강신청기간에 비해 수강신청기간 동안 학생들의 트래픽 발생량이 증가한 건물은 B, E, G, H 건물이며, E건물에서 가장 큰 증가량을 보인다.

그림 5는 수강신청기간과 비수강신청기간에 학내 E건물에서 발생하는 트래픽 양을 사용자군별로 나타낸 것이다. 두 기간 동안 교직원과 연구원들의 트래픽 발생량은 거의 동일하며, 학생들의 트래픽 양에서 큰 변화를 보이고 있다. 또한, 수강신청 기간 학생들의 트래픽의 90%이상이 본교 사이트인 점과 그림 4의 결과를

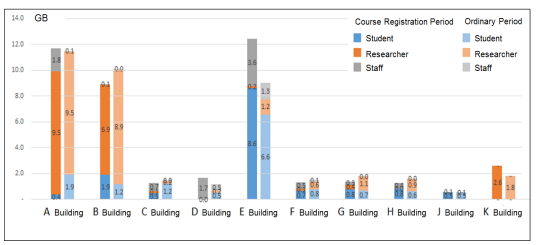

그림 4. 두 기간의 건물 별 시간당 평균 트래픽 발생량
Fig. 4. The Average Traffic Volume per Hour in Two Different Period

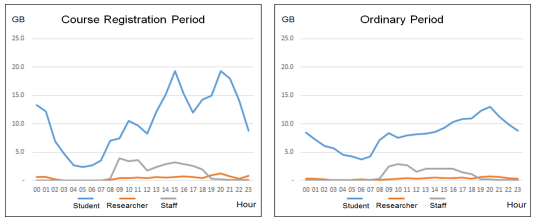

그림 5. 두 기간의 E건물 시간당 트래픽 발생량
Fig. 5. The Average Traffic Volume per Hour of E Building in Two Different Period

종합해보면, 많은 학생들이 E건물에 모여 수강신청을 진행한 것을 알 수 있다. 수강신청은 한 학기동안 수강할 강의들을 신청하는 것으로 학생들에게 매우 중요한 문제이기 때문에 해당 기간 내에 다른 건물들의 대역폭을 줄이고, E건물의 트래픽 발생량을 최대로 보장해 주어야 한다. 하지만 수강신청과는 무관한 연구원들의 트래픽 또한 보장해주기 위해서는 P2P 트래픽을 발생시키는 토렌트와 같은 서비스의 트래픽들을 우선적으로 제어해야 한다.

5.2 클러스터링을 이용한 토렌트 사용자 분류

본 절에서는 학습과 무관하며 많은 대역폭을 차지하는 토렌트 트래픽을 분석하여 학내 망 자원 효율화 방안을 제시하는 것을 목적으로 한 실험에 대해 기술한다. 먼저 토렌트 트래픽 데이터를 추출하고 요일별, 시간별 사용량을 이용하여 클러스터를 생성한다. 그 후 클러스터간의 특징을 추출하여 특별한 주의가 필요한 클러스터와 IP를 파악한다. 해당 정보를 이용하면 대역폭 조절이 필요할 때 최우선순위의 제어 대상 트래픽을 지정할 수 있다.

클러스터링에 앞서 WSS(Total Within Sum of Squares)와 CH(Calinski-Harabasz Index)를 사용하여 클러스터 개수를 추정하였다. WSS는 클러스터의 중심점 (centroid)과 클러스터를 이루는 데이터포인트간의 거리를 합한 숫자로, 클러스터의 개수를 정하는 척도 중 하나이다. 클러스터의 개수가 늘어나면서 WSS가 줄어들기 때문에 WSS가 급격히 줄어드는 값을 찾아야 한다. 반면, CH는 클러스터간의 분산과 전체 데이터 포인트 분산의 비율이다. CH값이 클수록 클러스터 내부의 분산은 작아지며, 클러스터 외부와의 분산이 커지므로 CH값이 클 때 클러스터링의 분별력이 높아진다.

그림 6은 k-means 클러스터링을 이용하여 k값(클러스터 개수)의 변화에 따른 WSS값과 CH값의 변화를 나타내는 그래프이다. 가로축은 k값을, 세로축은 WSS, CH값을 각각 나타낸다. 그림 6을 보면 k=3일 때 최대

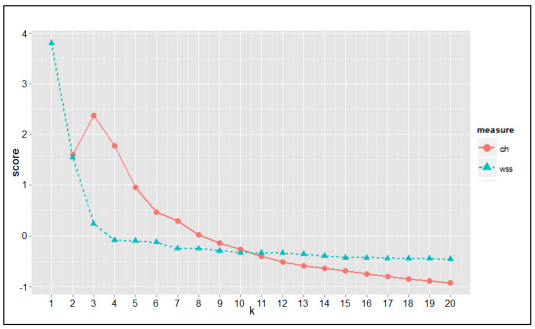

그림 6. K-Means를 통한 CH, WSS값의 비교
Fig. 6. Comparing of CH, WSS Using K-Means

CH값을 가지며 WSS값의 급격한 감소도 줄어든다. 하지만 학교 전체 사용자를 3개의 클러스터로 나누는 것은 클러스터간의 특징을 잘 반영하지 못할 수 있다. 따라서 CH값의 감소율은 줄어들고, WSS값의 감소율은 증가하는 k=7일 때와 비교했다.

그림 7, 8은 각각 k=3일때와 k=7일 때 클러스터의 적합성을 PCA(Principal Component Analysis)를 통해 시각화 한 것이다. PCA는 4차원 이상의 데이터를 2차원 공간으로 투영하여 시각화해준다. 그림 7과 8을 비교해보면, k=7인 경우 k=3일 때 보이는 특징을 간직한 채로 각 클러스터를 더욱 세분화 해주는 것을 볼 수 있다. 즉, k=7인 경우 더 구체적으로 사용자들을 분류하고 있다. 따라서 k값을 7로 정하였다.

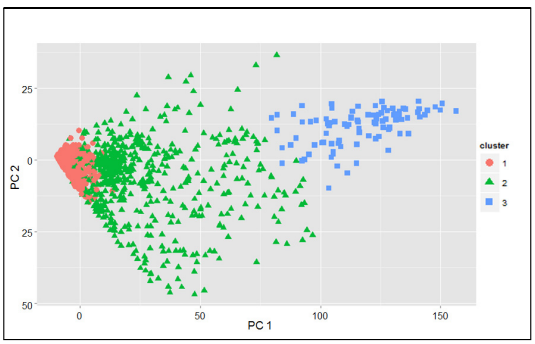

그림 7. k=3일 때 PCA를 이용한 유효성 시각화
Fig. 7. Validity Visualization Using PCA when k=3

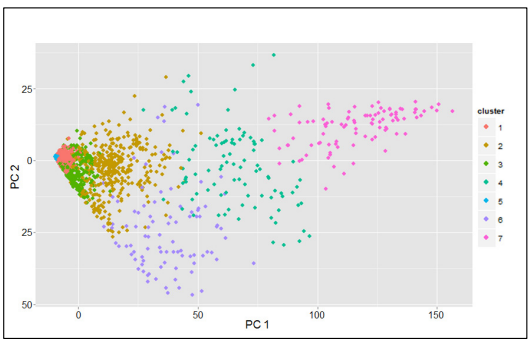

그림 8. k=7일 때 PCA를 이용한 유효성 시각화
Fig. 8. Validity Visualization Using PCA when k=7

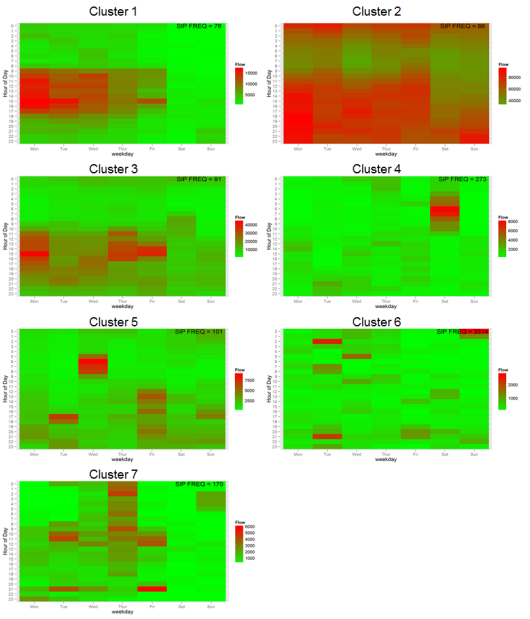

그림 9. 각 클러스터의 BitTorrent 사용 패턴
Fig. 9. BitTorrent Usage Pattern of Each Cluster

그림 9는 각 클러스터별 토렌트 사용 패턴을 Heatmap으로 나타낸 것이다. 가로축은 요일, 세로 축은 시간을 나타내며 색이 붉어질수록 많은 플로우 개수가 발생한 것이다. 분석 결과 대다수의 학내망 IP에서는 토렌트를 거의 사용하지 않는 것으로나타났다. 학내 망 전체 트래픽의 30%에 달하는 토렌트 사용량을 볼 때, 클러스터 2에 속하는 소수의 토렌트 헤비 유저가 학내 망 전체 대역폭에 끼치는 영향을 알 수 있다. 따라서, 개별 토렌트 헤비 유저들의 사용 시간과 특성을 파악하여 수강신청등과 같은 대역폭 보장이 필요한 학사 일정이 있을 때 대상에 대한 타이트한 제어 정책을 적용하여 대역폭을 확보하여야 한다.

Ⅵ. 결 론

본 논문에서는 빅 데이터 트래픽의 수집, 기초 데이터 생성, 데이터의 저장 및 관리 방안에 대해 제안하였다. 또한 분류 기준을 정의하고 분류 된 트래픽의 통계 분석을 통해 대상 망에 특화 된 자원 효율화 방안 모색을 제안하였다. 실험은 학내 망을 대상으로 진행하였으며, 3년간 학내 망에서 발생한 트래픽을 수집하고 상기

정의한 분류 기준으로 분류한 뒤, 통계 분석을 통해 수강신청 기간에 적합한 QoS 정책을 모색하였고 해당 QoS정책 적용을 우선시 해야 할 토렌트 헤비 유저들을 분류하였다.

분석 결과 수강신청 기간 동안 학내 망의 트래픽은 건물과 사용자군, 그리고 사용목적에 따라 발생량이 달라지며, 때문에 건물 및 사용 목적에 따라 다른 QoS정책 적용이 필요함을 알 수 있었다. 실험 결과를 통해 제안하는 방법을 이용한 통계 분석은 대상 망에 특화된 QoS정책을 도출해 낼 수 있음을 알 수 있었다.

향후 연구로는 유/무선 트래픽의 다양한 분석을 통해 대상 망의 특성에 맞는 유/무선 네트워크 QoS정책을 모색하고, 트래픽의 서비스 별 분석률을 높이는 방안과 분류 기준을 더욱 세분화하여 정책 적용이 달라질 수 있는 기준을 찾는 연구를 할 계획이다. 더 나아가, 사용자 별로 사용하는 서비스 분석을 통해 각 구성원들의 업무 및 수업의 참여도/집중도 분석을 할 계획이다.

References

[1] Y. Wang, Y. Xiang, W. L. Zhou, and S. Z. Yu, “Generating regular expression signatures for network traffic classification in trusted network management,” *J. Network Comput. Appl.,* vol. 35, pp. 992-1000, May 2012.

[2] B. Park, Y. Won, J. Chung, M. S. Kim, and J. W. K. Hong, “Fine-grained traffic classification based on functional separation,” *Int. J. Network Management,* vol. 23, pp. 350-381, Sept. 2013.

[3] C. S. Park, J. S. Park, and M. S. Kim, “Automatic Payload Signature Generation System,” *J. KICS*, vol. 38B, no. 08, pp. 615-622, Aug. 2013.

[4] J. H. Choi, J. S. Park, and M. S. Kim, “Processing speed improvement of HTTP traffic classification based on hierarchical structure of signature,” *J. KICS,* vol. 39B, no. 04, pp. 191-199, Apr. 2014.

[5] J. S. Park, S. H. Yoon, and M. S. Kim, “Performance improvement of the payload signature based traffic classification system using application traffic locality,” *J. KICS,* vol. 38B, no. 7, pp. 519-525, Jul. 2013.

[6] S. Lohr, *The age of big data*, New York Times, 11, 2012.

[7] T. Oetiker, “Monitoring your IT gear: the MRTG story,” *IT Professional*, vol. 3, no. 6, pp. 44-48, 2001.

[8] RRDtool, Available at: http://oss.oetiker.ch/rrdtool/.

[9] Bro, Available: http://www.bro.org/.

[10] Ntop, Available: http://www.ntop.org/.

[11] Snort, Available at: http://www.snort.org.

[12] B. H. Hong and H. J. Joo, “*A study on the monitoring model for traffic analysis and application of big data,*” 2013.

[13] S. P. Huang and G. E. Meng, “Research on the application of hadoop platform in the big data processing,” *Modern Computer*, vol. 29, no. 4, 2013.

[14] Hadoop, Available: http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html.

[15] A. D. Sarma, F. N. Afrati, S. Salihoglu, and J. D. Ullman, “Upper and lower bounds on the cost of a map-reduce computation,” *Very Large Data Bases(VLDB) Endownment*, pp. 277-288, Riva del Garda, Italy, 2013.

안 현 민 (Hyun-min An)

2012년 : 고려대학교 컴퓨터 정보학과 졸업

2014년 : 고려대학교 컴퓨터 정보학과 석사

2014년~현재 : 고려대학교 산업기술연구소 연구원

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

이 수 강 (Su-kang Lee)

2014년 : 고려대학교 컴퓨터 정보학과 졸업
2014년~현재 : 고려대학교 컴퓨터 정보학과 석사과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

심 규 석 (Kyu-seok Sim)

2014년 : 고려대학교 컴퓨터 정보학과 졸업
2014년~현재 : 고려대학교 컴퓨터 정보학과 석사과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

김 익 한 (Ik-han Kim)

2009년 : 고려대학교 정보통계학과 학부 졸업
2014년-현재 : 고려대학교 정보통계학과 석사과정 재학중
<관심분야> 빅데이터, 데이터마이닝

진 서 훈 (Seo-hoon Jin)

1994년 : 고려대학교 통계학과 석사 졸업
1998년 : 고려대학교 통계학과 박사 졸업
2007년~현재 : 고려대학교 응용통계학과 교수
<관심분야> 빅데이터, 다변량자료분석, CRM, 데이터마이닝

김 명 섭 (Myung-Sup Kim)

1998년 : 포항공과대학교 전자계산학과 졸업
2000년 : 포항공과대학교 컴퓨터공학과 석사
2004년 : 포항공과대학교 컴퓨터공학과 박사
2006년 : Post-Doc. Dept. of ECE, Univ. of Toronto, Canada
2006년~현재 : 고려대학교 컴퓨터정보학과 부교수
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 멀티미디어 네트워크