하둡 빅데이터 처리 플랫폼 기반의 교육전산망 자원 효율화 연구

정우석, 이수강, 심규석, 김성민, 김명섭
고려대학교

{hary5832, sukanglee, kusuk007, gogumiking, tmskim}@korea.ac.kr

Research on Hadoop Big Data Processing Platform based Education Network Resource Efficiency

Jung Woo-Suk, Lee Su-Kang, Sim Kyu-Seok, Kim Sung-Min, Kim Myung-Sup
Korea Univ.

요 약

효율적인 망 운영을 위해 장기간의 트래픽 분석을 통한 망의 특성을 정확히 반영하는 정책 적용이 필요하다. 빅데이터 분석 플랫폼과 도구의 개발을 통해 장기간 트래픽 분석이 가능해 졌고, 이를 활용한 엔터프라이즈 망 자원의 효율화 방안이 요구 되고 있다. 본 논문에서는 엔터프라이즈 망에서 발생한 장기간의 트래픽을 수집하고 저장 및 관리하는 방안에 대해 제안한다. 또한 분류기준을 정의하였으며, 수집된 빅데이터 트래픽을 각 분류 기준으로 분류한 뒤 다각적인 통계적인 분석을 통해 망 자원을 효율화 하는 방안을 제안한다. 제안하는 방법은 학내 망에 적용하여 실험하였으며, 통계 분석 결과 시간과 공간, 그리고 사용목적에 따라 QoS 정책을 달리 적용해야 함을 확인하였다.

Ⅰ. 서 론

네트워크 관리는 네트워크 자원을 최대한 활용하여 사용자에게 목적에 맞는 서비스를 제공하는 것을 목표로 한다. 이를 위해 네트워크 관리자들은 적절한 네트워크 정책을 수립하여 대상 네트워크에 적용한다. [1]

현재 본교에서는 원활한 네트워크 서비스를 지원하기 위해 방화벽, IDS/ IPS 등과 같은 고가의 관제 시스템을 운영하고 있고, 이를 유지 보수하기 위해 많은 비용이 지출되고 있다. 하지만 지출에 비해 구성원들이 느끼는 질적 향상은 미비한 수준이다. 이러한 문제를 해결하기 위해서는 망 트래픽에 대한 다각적인 분석을 통해 망에 특화 된 효율적인 QoS 정책 추진이 필요하다. 또한, 설치된 장비의 운영률과 효율성 판단을 통해 향후 네트워크 장비 확충 시에 필요한 객관적인 근거 자료 마련이 필요하다.

엔터프라이즈 망의 특성상 사용자 군(학생, 연구자, 교직원)이 뚜렷하게 구분되고, 사용 목적이 다르다. 따라서 사용자 군에서 나타나는 트래픽의 유형별 특징을 반영한 분석이 필요하다. 장기간의 트래픽 수집, 저장 그리고 다각적인 분석을 통하여 해당 망에 특화 된 QoS 정책 수립과 망 관리 자원 최적화를 할 수 있다.

본 연구에서는 장기간의 트래픽을 수집하여 기초 데이터를 생성하고, 빅데이터 기반의 고급 통계분석을 통해 학내 네트워크 자원의 효율화 방안을 도출하는 방법을 제안하고, 이를 학내 망에서 3 년간 발생한 트래픽을 통해 실험하였다.

본 논문의 구성은 다음과 같다. 2 장에서 관련 연구에 대해 조사하고, 3 장에서는 빅데이터 트래픽 수집, 저장 및 관리방안을 제안한다. 4 장에서는 빅데이터 트래픽을 통계 기반으로 분석하고, 마지막으로 5 장에서는 결론과 향후 연구를 언급한다.

Ⅱ. 관련 연구

인터넷이 고속화 됨에 따라 발생하는 트래픽의 양 또한 기하급수적으로 증가하였다. 전통적인 트래픽 분석 방법으로는 현재 인터넷에서 발생하는 대용량의 트래픽에 대한 응용판별과 같은 단순한 분석만이 가능하다. [2, 3] 또한, 고속 네트워크에서 발생하는 대용량의 트래픽에 대한 실시간 처리, 장시간 저장, 다양한 분석 등 다양한 기능을 포함한 종합적인 분석이 미흡한 실정이다. 현재의 트래픽 분석 시스템들이 장기간의 대용량 트래픽에 대한 다각적인 분석을 하지 못하는 이유는 크게 세가지로 정리 된다. 장기간의 대용량 트래픽 데이터에 대한 효율적인 저장 방법의 부재와 개별적으로 개발 및 구축된 다양한 방법론들의 효과적인 통합 방법의 부재, 그리고 장기간 동안 축적된 대용량 트래픽 데이터로부터 다각적인 분석 결과를 추출할 수 있는 방법의 부재가 그것이다.

본 논문에서는 언급된 빅데이터 트래픽 분석의 단점을 해결하기 위해 수집된 빅데이터 트래픽을 경향 분석, 집단분석, 계층분석 그리고 장애분석의 4 가지 관점을 통해 분석하고, 그 결과를 통해 망에 특화된 자원 효율화 방안에 대해 기술한다.

Ⅲ. 빅데이터 트래픽 수집・저장 및 관리

망 사용자들의 모든 트래픽을 수집을 하게 되면 분석 가치는 없고 시스템의 오버헤드만을 늘리는 트래픽 또한

수집하게 된다. 이를 해결하기 위해 본 연구에서는 각각 분석 대상 트래픽을 정의하고 비대상 트래픽을 제외하는 전처리 작업을 진행하였다.

효과적인 망 자원 효율화 방안을 모색하기 위해서는 트래픽을 다양한 분류 속성으로 분류하고, 분류 결과와 트래픽 정보를 토대로 한 통계 분석을 통해 망에 특화된 정책 도출이 필요하다. 이를 위해 본 연구에서는 트래픽을 각 서비스와 메타 데이터를 기준으로 헤더 정보 기반의 분류 방법론과 트래픽 상관관계 기반의 분류 방법론을 사용하여 분류하였다.

그림 1 은 제안하는 빅데이터 트래픽 저장 및 관리 시스템이다. 빅데이터 트래픽을 저장하고 관리하기 위한 플랫폼으로 하둡(Hadoop)을 사용한다. 하둡 클러스터에 저장한 데이터를 관리하기 위한 관리 프로그램으로는 하이브(Hive)를 사용하였다. 또한, 분류 된 트래픽을 분석하기 위한 통계 분석 툴로는 R 을 사용 하였고, 실제 통계 분석에는 RHive 를 통해 R 을 사용하여 하이브와 연동하였다.

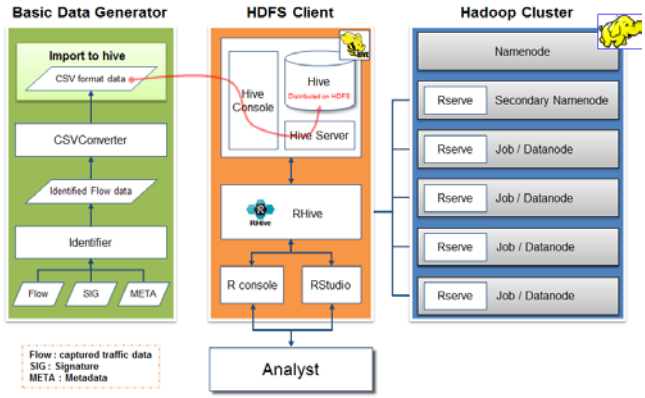

그림 1. 빅데이터 트래픽 저장 및 관리 시스템 구조

Ⅳ. 통계 기반 빅데이터 분석

그림 2 는 수간 신청 기간과 비수강 신청기간에 학내 A 건물에서 발생하는 트래픽 양을 사용자군별로 나타낸 것이다. 두 기간 동안 교직원과 연구원들의 트래픽 발생량은 동일하며, 학생들의 트래픽 양에서 큰 변화를 보이고 있다. 또한, 수간 신청 기간 학생들의 트래픽 양에서 큰 변화를 보이고 있다. 수강신청은 학생들에게 매우 중요한 문제이기 때문에 해당 기간 내에 토렌트와 같은 서비스의 트래픽들을 우선적으로 제어해야 한다.

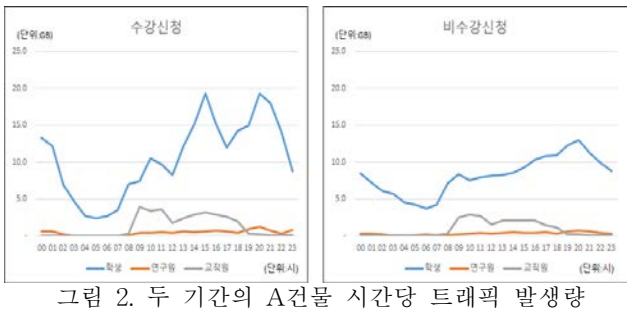

그림 2. 두 기간의 A건물 시간당 트래픽 발생량

그림 3 은 학습과는 무관하며 많은 대역폭을 차지하는 토렌트 트래픽의 각 클러스터별 사용 패턴을 Heatmap 으로 나타낸 것이다. 가로축은 요일, 세로축은 시간을 나타내며 색이 붉을수록 많은 플로우가 발생한 것이다. 분석 결과 대다수의 학내망 IP 에서는 토렌트를 거의 사용하지 않는 것으로 나타났다. 학내 망 전체 트래픽의 30%에 달하는 토렌트 사용량을 볼 때, 클러스터 2 에 속하는 소수의 토렌트 헤비 유저가 학내 망 전체 대역폭에 큰 영향을 미침을 알 수 있다. 따라서, 개별 토렌트 헤비 유저들의 사용 시간과 특성을 파악하여 대역폭 보장이 필요한 학사일정이 있을 때 적절한 제어 정책을 통해 대역폭을 확보하여야 한다.

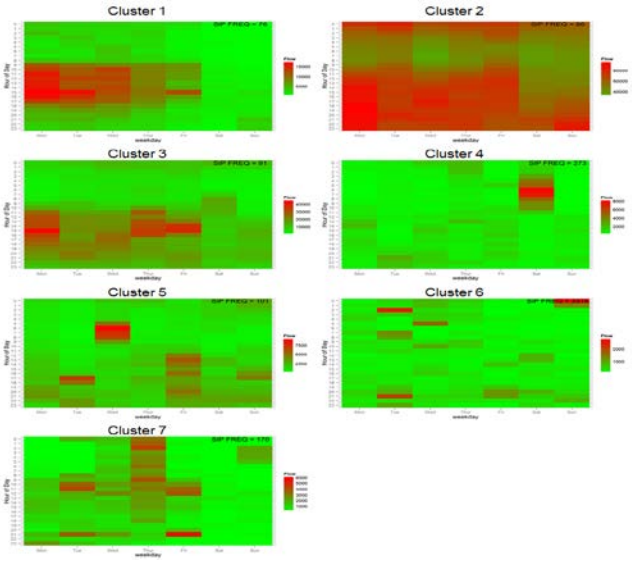

그림 3. 각 클러스터의 BitTorrent 사용 패턴

Ⅴ. 결론 및 향후 연구

본 논문에서는 빅데이터 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. 분석 결과 수강 신청 기간 동안 학내 망의 트래픽은 사용자군과 사용 목적에 따라 발생량이 달라지며, 이에 따라 사용 목적에 따라 다른 QoS 정책이 필요함을 알 수 있었다. 또한, 실험 결과를 통해 제안하는 방법을 이용한 통계 분석은 대상 망에 특화된 QoS 정책을 도출해 낼 수 있음을 알 수 있었다.

향후 연구로는 유/무선 트래픽의 다양한 분석을 통해 대상 망의 특성에 맞는 유/무선 네트워크 QoS 정책을 모색하는 연구를 진행할 예정이다.

ACKNOWLEDGMENT

본 논문은 교육과학기술부의 재원으로 한국연구재단의 지원을 받아 수행된 BK21 플러스 사업의 연구 결과임 (No. T1300572)

참 고 문 헌

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

[2] Lohr, S. (2012). The age of big data. New York Times, 11.

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