하둡 시스템 정보의 이상탐지를 위한 시각화*

양석우[1]°, 손시운[1], 길명선[1], 문양세[1], 원희선[2]
[1]강원대학교 컴퓨터과학과, [2]한국전자통신연구소
e-mail: {seakwoo, ssw5176, gils, ysmoon}@kangwon.ac.kr, hswon@etri.re.kr

Visualization of Anomaly Detection in Hadoop System Information

Seokwoo Yang[1]°, Siwoon Son[1], Myeong-Seon Gil[1], Yang-Sae Moon[1], and Hee-Sun Won[2]
[1]Dept. of Computer Science, Kangwon National University, [2]ETRI

요 약

본 논문에서는 하둡 환경에서 시스템 정보의 이상탐지를 위한 시각화 기능을 설계 및 구현한다. 제안한 이상탐지 시각화 기능은 크게 세 단계로 구분된다. 먼저, 각 노드로부터 시스템 로그 데이터(캐시 및 메인 메모리)를 수집하여 하이브(Hive)에 저장한다. 그리고 저장한 데이터에 3-시그마 규칙을 적용하여 이상탐지를 수행한 후 관계형 데이터베이스에 적합하도록 재가공한다. 마지막으로, 스쿱(Sqoop)을 통해 RDBMS(MariaDB)에 이상탐지 결과를 저장하고, DHTMLX 차트 라이브러리를 사용하여 이를 시각화한다. 시각화 결과, 로그 데이터의 이상탐지와 데이터간의 상관관계를 직관적으로 이해할 수 있게 되었다.

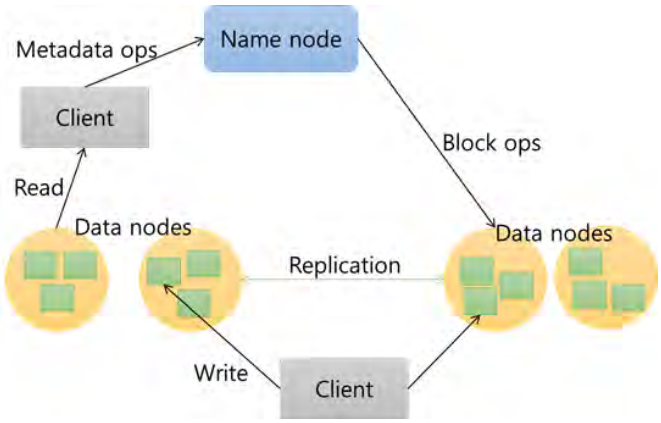
1. 서론

시각화[1]는 데이터 자체나 데이터 분석 결과를 이미지나 차트, 다이어그램, 애니메이션 등의 기법을 사용하여 시각적으로 표현하는 방법이다. 이러한 시각화 기술은 인문, 자연 과학, 교육, 정보공학 등 여러 분야에서 사용되고 있으며, 특히 데이터 마이닝 분야에서는 SNA(social network analysis), 클러스터링, 시계열 매칭 등의 마이닝 결과를 효과적으로 표현하는데 많이 사용되고 있다. 본 논문에서는 시각화를 통해 이상탐지(anomaly detection)[2]에 대한 분석 및 처리 결과를 보다 직관적으로 판단하고자 한다. 이를 위해 본 논문에서는 하둡(Hadoop)[3] 환경에서 시스템 정보(캐시 및 메인 메모리의 로그 데이터)의 이상탐지를 위한 시각화 도구를 설계하고 구현한다. 이상탐지는 기계학습과 데이터 마이닝의 중요한 문제로서 여러 응용 분야에서 다양한 형태로 연구되고 있다. 또한, 하둡 환경은 다수의 서버로 구성된 분석 시스템으로, 각 분석 서버의 비정상 활동에 대한 신속한 탐지가 필요하다. 따라서, 본 논문에서는 하둡 환경에서 수집한 로그 데이터를 하이브(Hive)[4]에 저장하고, 저장된 로그 데이터에 3-시그마 규칙[5]을 적용하여 이상탐지를 수행한다. 그리고 분석된 이상탐지 결과를 DHTMLX[6] 차트 라이브러리를 통해 시각화하여 보다 직관적으로 그 결과를 확인하고자 한다.

* 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [14-000-05-001, 스마트 네트워킹 핵심기술개발]

2. 관련 연구

하둡 분산 파일 시스템(Hadoop Distributed File System: HDFS)[7]은 빅데이터 분석을 위해 구축된 분산처리 시스템이다. HDFS는 그림 1과 같이 네임노드(name node)와 데이터노드(data node)로 구성되어 있다. 네임노드는 파일 시스템에 대한 메타데이터를 보유하며, 네임스페이스와 클라이언트를 제어한다. 데이터노드는 보통 클러스터 내의 노드당 하나씩 존재하며, 클라이언트는 저장되고자 하는 파일을 분할하여 각 데이터노드에 블록 단위로 저장한다. 본 논문에서는 이러한 HDFS에서의 이상탐지분석을 위해 네임노드와 데이터노드에서 일정 주기(실험에서는 1분 주기)로 캐시 메모리와 메인 메모리 로그 데이터를 수집한다.



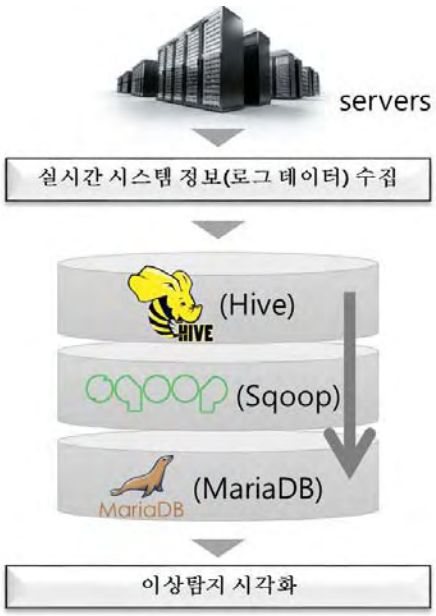
(그림 1) 하둡 분산 파일 시스템(HDFS)의 구조.

최근 데이터의 양이 증가하면서 대용량의 데이터, 즉 빅데이터를 활용한 분석이 활발히 이루어지고 있다. 또한, 분석 결과를 보다 직관적으로 이해하기 위해 다양한 시각화 방법이 연구되고 있다. 참고문헌[8]에서는 윤곽선 이미지 기반의 시계열 매칭 결과를 다양한 차트로 제공하였다. 구현된 시각화 도구는 질의 시계열과의 윤곽선 이미지 매칭, 이동평균 변환 매칭, 정규화 변환 매칭 등의 결과를 폴라 차트, *k*-NN 차트 등으로 표현하였다. 또한, 시계열 매칭 결과를 원본 이미지와 함께 제공함으로써, 이미지 도메인과 시계열 도메인의 매칭 결과를 직관적으로 표현한다. 이상탐지에 대한 시각화 연구로는, Janetzko 등[9]의 전력 소비량에 대한 이상탐지 분석 결과의 시각화가 있다.

3. 하둡 시스템 정보의 이상탐지 시각화 설계 및 구현

3.1. 이상탐지를 위한 시각화 설계

본 논문에서 제안하는 하둡 환경의 시스템 이상탐지를 위한 시각화는 그림 2와 같이 실시간 데이터 수집, 데이터 저장, 이상탐지 분석 및 MariaDB[10]를 통한 시각화로 구성된다. 이상탐지 과정을 자세히 설명하면 다음과 같다. 먼저, 하둡 환경에서 각 노드의 시스템 정보를 실시간으로 수집하여 하이브에 저장한다. 그리고 저장된 데이터에 3-시그마 규칙을 적용한 후 결과를 스쿱(Sqoop)을 통해 MariaDB에 저장한다. 이후 MariaDB에 저장된 데이터를 DHTMLX의 차트 라이브러리를 활용하여 시각화를 수행한다. 여기에서, 이상탐지의 분석 데이터를 스쿱을 통해 MariaDB로 다시 저장하는 이유는 하이브에 저장된 데이터를 바로 사용할 경우, 하이브가 질의를 맵리듀스(Map-Reduce)[11]로 처리한 후 결과를 반환하여 매우 많은 처리 시간 오버헤드를 발생시키기 때문이다. 따라서, 이러한 중간 과정이 없는 MariaDB를 사용하여 분석 데이터를 저장한다.

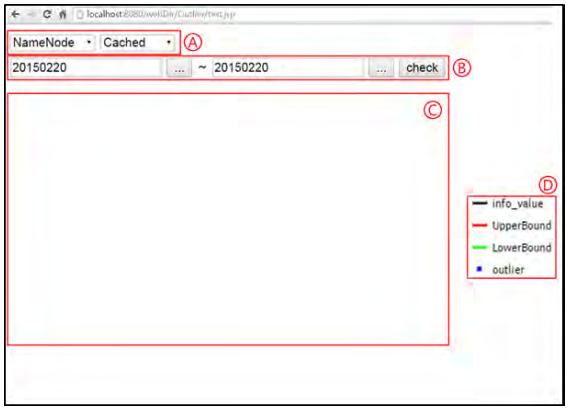


(그림 2) 이상탐지 시각화 전체 프레임 워크.

이상탐지를 위해 본 논문에서는 3-시그마 규칙과 이동평균 변환 기법을 사용한다. 먼저, 3-시그마 규칙이란, 대부분의 실제 값들은 평균에서 ±3 표준편차 범위 이내에 존재한다는 통계학적 규칙을 배경으로 이 범위를 벗어나는 값을 이상치로 판단하는 방법이다. 다음으로, 이동평균 변환 기법은 일정구간의 데이터에서 평균을 계산하는 기법이다. 본 논문에서는 3-시그마 규칙에서 사용하는 평균 값을 이동평균 값으로 사용하므로, 이동평균 구간에 따라 이상탐지 성능이 크게 달라진다. 본 논문에서는 시스템 로그 데이터를 1분 간격으로 수집하며, 이동평균 구간을 30분으로 설정하였다.†

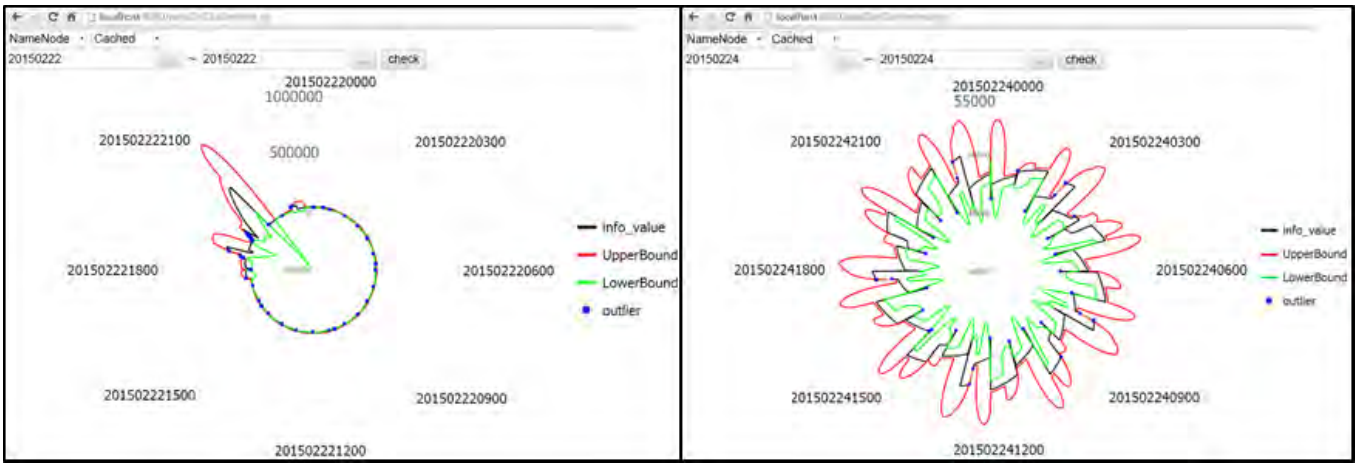
3.2. 이상탐지 시각화 도구 구현

본 논문에서는 총 네 대의 동일 서버를 사용하여 로그 데이터를 수집하였다. 서버의 하드웨어 플랫폼은 Intel® Core™ i3-4150 CPU @ 3.5GHz, 450GB HDD, 4GB RAM이며, 소프트웨어 플랫폼은 CentOS-6.6-x64 Linux 운영체제를 설치하여 수행하였다. 로그 데이터 저장을 위한 하둡 프레임워크는 하나의 네임노드와 세 개의 데이터노드로 구성되어 있으며, 모두 Hadoop 버전 2.6.0을 사용하였다. 시각화에 사용된 데이터는 네 대의 서버에서 추출한 시스템 정보로, 메인 메모리(RAM) 사용량과 CPU 캐시(cache) 메모리 사용량을 1분마다 수집하여 하이브에 저장하였다. 그리고 하이브에 저장된 로그 데이터로부터 HiveQL을 사용하여 이상탐지를 수행 후, 이 결과를 스쿱을 통해 MariaDB에 저장하였다. 그림 3은 로그 데이터의 이상탐지 시각화를 위한 초기 화면이다. 화면에서 Ⓐ는 서버의 호스트 명(NameNode, DataNode01, DataNode02, DataNode03)과 해당 서버에서 시각화 하고자 하는 로그 데이터를 선택하는 부분이고, Ⓑ는 시각화를 원하는 기간을 설정하는 부분이다. 또한 Ⓒ는 실제 로그 데이터를 폴라 차트로 표현할 영역이고, Ⓓ는 시각화된 차트의 범례를 나타낸다.



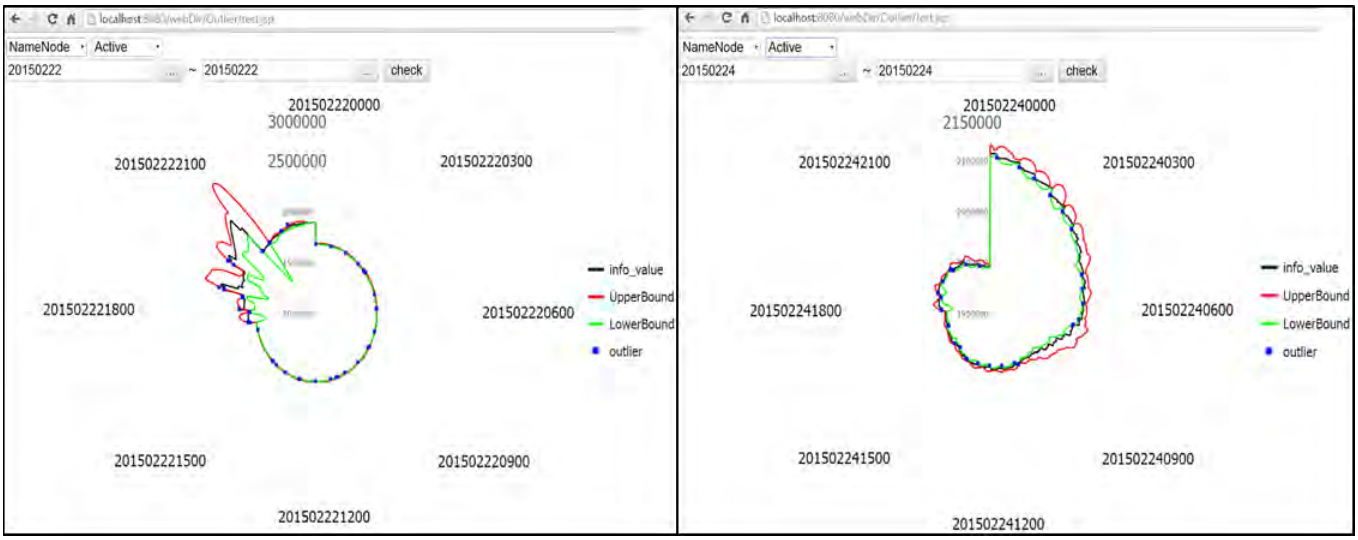
(그림 3) 이상탐지 시각화의 초기 화면.

[1] 이동평균 성능은 데이터의 특성(주기)에 따라 다르게 설정해야 한다. 본 논문에서는 수집한 데이터를 사용하여 10 분, 30 분 60 분을 이동평균 구간으로 각각 테스트 해 보았으며, 이중 이상 탐지가 가장 잘 나타나는 30 분을 이동평균 구간으로 고정하였다.



(a) 네임노드의 작업이 있을 경우. (b) 네임노드의 작업이 없을 경우.

(그림 4) 캐시 메모리 사용량에 대한 시각화 결과.



(a) 네임노드의 작업이 있을 경우. (b) 네임노드의 작업이 없을 경우.

(그림 5) 메인 메모리 사용량에 대한 시각화 결과.

그림 4는 캐시 메모리에 대한 시각화 결과이다. 먼저, 그림 4(a)는 2015년 2월 22일 하루 동안 네임노드의 캐시 메모리 사용량을 시각화한 결과이다. 그림을 보면, 시간의 흐름에 따라 캐시 메모리에 어떠한 변화가 있는지를 알 수 있다. 그림에서 12시 방향은 자정(0시)을 나타내며, 6시 방향은 정오(12시)를 나타낸다. 또한, 점으로 표시된 부분은 캐시 메모리의 이상치를 탐색한 결과이다. 먼저, 자정부터 18시까지는 캐시 메모리의 변화가 거의 나타나지 않는데, 이는 네임노드가 기본 작업(운영체제 등 시스템 소프트웨어) 이외에는 다른 작업을 수행하지 않기 때문이다. 반대로, 18시 이후에는 네임노드가 작업을 수행하기 때문에 그림과 같이 캐시 메모리 사용량의 변화가 생겼다. 그리고 다시 21시부터는 네임노드의 작업이 중단되어 캐시 메모리의 사용량이 줄어들고 변화 또한 점차 줄어들게 된다. 그림을 보면, 매 시간마다 이상 탐지가 발생하는 것을 알 수 있다. 여기에서 발생하는 이상탐지는 크게 두 부분으로 나눌 수 있다. 먼저, 캐시 메모리가 많이 사용되었을 경우(18시 - 21시)에는 3-시그마 규칙을 적용하였을 경우의 이상탐지 범위를 벗어나 실제로 이상탐지가 발생한 경우이고, 반대로 캐시 메모리의 변화가 없는 경우(21시 - 18시)에는 캐시 메모리 수집 시 매 시간마다 캐시 메모리를 정리하는 `sync; echo 3 > /proc/sys/vm/drop_caches` 리눅스 명령어를 수행하여 이상탐지로 판단된 경우이다.

다음으로, 그림 4(b)는 2015년 2월 24일 하루 동안 네임노드의 캐시 메모리 사용량에 대한 시각화 결과이다. 그림을 보면, 그림 4(a)와는 다르게 특정한 주기만 있을 뿐, 전체적으로는 큰 변화가 없는 것을 알 수 있다. 여기서, 특정 주기가 나타나는 그림 4(a)에서 설명한대로, 캐시 메모리를 매 시간 정리하기 때문이다. 따라서, 그림과 같이 캐시 메모리 사용량이 일정한 주기를 보이게 되며, 이러한 상태에서 발생한 이상탐지는 실제로는 이상이 아니라 정상으로 판단된다. 그런데, 그림 4(a)에서는 그림 4(b)와 같은 일정한 주기가 보이지 않는데, 이는 그림 4(a)와 4(b)의 y-축 스케일(1:10)이 다르기 때문이다. 만약, 두 그림의 y-축 스케일을 동일하게 적용하면, 그림 4(a)의 네임노드 휴지시간(idle time, 21시 - 18시)은 그림 4(b)와 동일한 형태를 띄게 된다.

그림 5는 그림 4와 동일한 날짜에 사용한 메인 메모리를 각각 시각화한 결과이다. 먼저, 그림 5(a)를 보면, 그림 4(a)와 유사한 경향을 보인다. 즉, 네임노드가 작업을 수행하는 시간(18시 - 21시)에 3-시그마 규칙에 의해 이상이 탐지된 것을 알 수 있다. 또한 휴지시간에서의 이상탐지는 앞서 설명한 것과 동일하게 정상으로 간주한다. 다음으로, 그림 5(b)는 그림 4(b)와는 조금 다른 경향을 보이는 것을 알 수 있다. 하지만, 이는 *y*-축 스케일 문제로, 실제 캐시 메모리 비율과 메인 메모리를 같은 스케일로 정규화 한다면, 그림 5(b) 또한 그림 4(b)와 유사한 경향을 보이게 된다.

4. 결론 및 향후 연구

본 논문에서는 하둡 환경에서 시스템 정보의 이상탐지를 위한 시각화 기능을 설계 및 구현하였다. 시각화는 마이닝 결과를 직관적으로 이해하는데 매우 중요한 요소이다. 제안한 이상탐지의 시각화는 하둡 환경에서 각 노드의 로그 데이터를 활용하여 이상탐지를 분석했다는 점에서 매우 실용성 있는 연구라 사료된다. 로그 데이터 이상탐지 결과를 폴라 차트로 표현함으로써, 보다 직관적으로 분석 결과를 이해할 수 있었다. 향후 연구로는 네트워크 트래픽 데이터, CPU 점유율 등의 이상탐지 분석과 장기간 수집된 로그 데이터를 기반으로 이상탐지를 수행하고 그 결과를 시각화하여 분석할 예정이다.

참고문헌

[1] Wikipedia, http://en.wikipedia.org/wiki/Visualization.

[2] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” University of Minnesota, Tech. Rep., 2007.

[3] C. Lam and J. Warren, “Hadoop in Action,” Manning Publications, 2010.

[4] Apache Hive, https://hive.apache.org.

[5] 3-sigma rule, http://ko.wikipedia.org/wiki/68-95-99.7_rule.

[6] DHTML, http://www.dhtmlx.com.

[7] HDFS, http://hadoop.apche.org/hdfs.

[8] 문성우, 이상훈, 김범수, 문양세, “시계열 데이터 기반 윤곽선 이미지 매칭의 시각화 도구,” 한국정보과학회 동계학술발표회, pp. 243-245, 12월 2014.

[9] H. Janetzko, F. Stoffel, S. Mittelstädt, and D. A. Keim, “Anomaly Detection for Visual Analytics of Power Consumption Data,” *Computers and Graphics*, Vol. 38, pp. 27-37, Feb. 2014.

[10] MariaDB, http://mariadb.org.

[11] J. Dean and S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters,” *Communications of the ACM*, Vol. 51, No. 1, pp. 107-113, Jan. 2008.