Journal of Korean Institute of Intelligent Systems, Vol. 23, No. 5, October 2013, pp. 392-399
http://dx.doi.org/10.5391/JKIIS.2013.23.5.392

빅데이터 특집 논문

빅 데이터 기반의 네트워크 패킷 분석 모델

The Model of Network Packet Analysis based on Big Data

최보민* · 공종환* · 한명묵**†
Bomin Choi, Jong-Hwan Kong, and Myung-Mook Han†

*가천대학교 컴퓨터공학과
† Department of Computer Engineering, Gachon University

요 약

IT 기술 발달 및 정보화 시대로 인해 우리 사회 전반에 걸쳐 많은 부분이 네트워크에 대한 의존도가 상당히 커지고 있다. 이는 다양한 정보 및 서비스 획득의 용이성을 제공해 주는 이점이 있는 반면에, 네트워크 침입자들로 하여금 더 많은 취약성의 루트를 제공할 수 있는 부정적 효과도 따르고 있다. 이는 네트워크 이용과 함께 증가한 패킷의 다양한 루트를 악용하여 네트워크의 연결된 시스템에 서비스 장애나 마비를 일으키는 악의적인 위협 및 공격 또한 함께 증가하고 있음을 의미하며 이러한 문제에 대한 해결책이 시급히 필요하다. 이에 보안 분야에서는 네트워크 패킷이나 시스템 로그 등을 수집하여 이를 분석하고 이러한 위협에 대응할 수 있는 다양한 보안 솔루션을 개발하고 있으나, 기존의 분석 방식들로는 점차 방대해져가고 있는 보안 데이터들을 처리하는데 데이터 저장 공간 부족 및 이에 따른 성능 저하와 같은 여러 문제점들이 발생하고 있다. 따라서 본 논문에서는 보안 영역 분야에서도 최근 이슈가 되고 있는 빅 데이터 기술을 적용하여 이러한 문제점들을 개선하는 모델을 제안한다. 즉, 대용량 데이터 저장 기술인 NoSQL을 통해 점차 방대해져 가는 패킷데이터를 수집하고, 분산 프로그래밍모델인 맵리듀스 기반의 K-means 클러스터링을 설계하여 네트워크 침입에 대한 특징 및 패턴을 추출 할 수 있는 분석모델을 제안하고 실험을 통하여 이에 대한 우수성을 입증하였다.

키워드 : 빅 데이터, NoSQL, 맵리듀스, K-means 클러스터링, 패킷 분석

Abstract

Due to the development of IT technology and the information age, a dependency of the network over the most of our lives have grown to a greater extent. Although it provides us to get various useful information and service, it also has negative effectiveness that can provide network intruder with vulnerable roots. In other words, we need to urgently cope with theses serious security problem causing service disableness or system connected to network obstacle with exploiting various packet information. Many experts in a field of security are making an effort to develop the various security solutions to respond against these threats, but existing solutions have a lot of problems such as lack of storage capacity and performance degradation along with the massive increase of packet data volume. Therefore we propose the packet analysis model to apply issuing Big Data technology in the field of security. That is, we used NoSQL which is technology of massive data storage to collect the packet data growing massive and implemented the packet analysis model based on K-means clustering using MapReudce which is distributed programming framework, and then we have shown its high performance by experimenting.

Key Words : Big Data, NoSQL, MapReduce, K-means Clustering, Packet Analysis

1. 서 론

접수일자: 2013년 8월 16일
심사(수정)일자: 2013년 9월 9일
게재확정일자 : 2012년 9월 16일
† Corresponding author
본 연구는 미래창조과학부가 지원한 2013년 정보통신·방송(ICT) 연구개발사업의 연구결과로 수행되었음.

컴퓨터와 스마트 기기의 발전에 따른 인터넷 접근 방식 수단의 다양화는 우리 삶에 많은 이점을 제공해 주고 있지만, 이는 또한 네트워크 공격자로 하여금 보다 다양한 공격 방법과 기술, 표적 대상 등 수 많은 보안 위협의 경로를 제공하고 있다[1]. 즉, 이는 ICT(Internet Communication Technology) 기술의 발달로 인터넷 이용이 잦아지면서 이로 인해 수많은 패킷이 발생하게 되고, 이와 비례하게 이를 악용하여 네트워크를 위협할 수 있는 정보 및 수단이 다양해지고 있음을 의미한다. 이와 같이 해마다 증가하고 있는 네트워크 위협 및 공격은 발생 빈도수 뿐 아니라, 그 유형이 점차 다양화 되어가고 있어 그 피해 규모가 점차 커지고

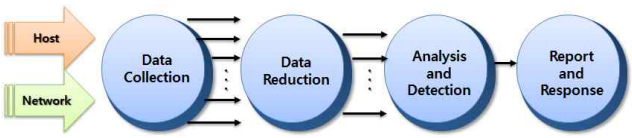


그림 1. 침입탐지 모델의 주요 기능 및 절차
Fig. 1. The Main Function and Process of Intrusion Detection Model

있다. 이처럼 빠르게 증가하고 점차 치명적으로 진화하고 있 보안 위협에 대응하기 위하여 많은 보안 전문가들은 방화벽, 침입 탐지 시스템 등과 같은 다양한 보안 솔루션을 제시하고 있다.

침입 탐지 모델의 경우 대게 트래픽이나 패킷과 같이 네트워크로부터 수집된 보안 관련 정보들을 분석하여 공격패턴을 추출하고 이를 기반으로 침입 또는 오용을 탐지할 뿐 아니라 침입에 대한 적절한 대응취하는 보안 솔루션 모델이다[2]. 그림 1은 침입탐지 모델의 주요 기능 및 절차를 나타낸 것이다. 가장 먼저 탐지 대상 시스템의 사용내역이나 통신에서 사용되는 패킷 등을 수집하여 침입 분석의 쓰임에 맞게 이를 가공하고, 이를 분석하여 산출된 결과를 가지고 시스템의 비정상적인 사용을 탐지 및 경우에 따라서는 대응하는 순으로 프로세스가 처리된다. 이러한 프로세스 과정 중 특히 분석단계는 가장 주요한 단계로서 이상행동을 보이는 패킷 데이터들의 특징이나 패턴을 추출하고 이를 바탕으로 침입들의 악의적인 성향을 판단할 수 있게 한다. 그러나 앞서 네트워크상에서 발생되는 정보의 양이 점차 증가하고 있음에 따라 수집되는 정보의 양 또한 방대해져 기존의 데이터 저장소로는 이를 담당하지 못할 뿐 아니라 이로 인한 성능적인 결함 또한 발생 시킬 수 있다. 이에 보다 빠르고 효율적으로 대응 할 수 있는 네트워크 보안 분석 기술이 시급히 요구되고 있다.

이에 본 논문에서는 대용량화 되어가고 있는 패킷 데이터들을 보다 효율적으로 수집하고 분석하기 위해 빅 데이터 기술을 적용할 것을 제안한다. 본 논문에서 제안하는 패킷 데이터 분석 모델에서는 NoSQL기반의 데이터 저장소를 통하여 데이터 저장 공간 부족 및 이에 따른 성능적 결함을 해결하고자 하였고, 수집된 대량의 패킷 데이터를 분산 프로그래밍 모델인 맵리듀스 기반의 K-means클러스터링의 구현을 통하여 보다 신속하게 분석을 수행하여 위협적인 침입에 빠르게 대응 할 수 있는 빅 데이터 기술 기반의 패킷 분석 모델을 제안한다.

본 논문의 구성으로 2장 에서는 관련 연구를 통한 이전 연구들에 대해 탐구하고 기존 연구의 한계점을 찾아내며, 제안하는 빅데이터 기술인 NoSQL과 맵리듀스 프로그래밍 기법에 대하여 소개한다. 또한 3장 에서는 제안하는 빅데이터 기술 기반의 네트워크 패킷 분석 모델을 소개하고, 4장에서는 이를 기반으로 실험한 내용을 바탕으로 제안하는 기법의 우수성을 입증하였다. 마지막으로 5장에서 본 연구에 대한 최종 결론과 향후 연구 과제를 제시하며 본 논문을 정리한다.

2. 관련 연구

2.1 연구 배경

해마다 증가하고 다양해지고 있는 공격을 탐지하고 이에 대응하기 위해 보안 데이터 분석에 대한 중요도가 점차 커지고 있다. 특히, 네트워크 패킷 데이터는 다양한 필드의 정보를 담고 있기 때문에 이를 분석하여 얻은 결과들은 침입에 대한 특징이나 패턴 등을 판단하는데 중요 척도가 될 수 있다. 기존의 패킷 분석은 주로 전문가를 통해 이루어져 왔다. 그러나 널리 알려진 공격의 경우 전문가의 선험적 지식을 통해 시그니처(Signiture) 생성이 가능하였지만, 변형되거나 새로운 공격유형이 많아지고 그 데이터의 양이 증가하게 되면서 기존의 사람이 하던 방식의 분석은 여러 가지 어려움이 동반하며, 많은 양의 분석 비용 또한 요구되는 단점이 있다[3]. 이에 다양한 종류의 기계학습(Machine Learning)이나 데이터 마이닝(Data Mining) 등의 인공지능(Artificial Intelligence) 기술들을 적용하여 분석과정의 자동화를 구축하고, 보다 신속하고 효율적으로 침입에 대응할 수 있는 기법들이 연구되고 있다. 즉, 인공지능 기술들의 도입으로 분석 프로세스를 자동화 하여 기존 공격 유형의 특징 뿐 아니라, 변형되거나 새로운 공격 유형의 특징들을 적은 비용으로도 보다 빠르게 추출해 낼 수 있게 되었다.

본 논문에서는 자동화된 패킷 분석처리에 있어 K-means 클러스터링 기법을 적용하였다. 기계 학습의 경우 대게 감독 모드(supervised mode) 기반의 알고리즘들로서, 이는 획득된 학습 데이터의 질에 따라 그 결과가 영향을 많이 받게 된다. 그러나 이는 경우에 따라서 변동 가능성이 다분한 네트워크 패킷 데이터의 특성상 효율적이지 못할 뿐 아니라 많은 비용이 요구될 수도 있는 단점이 있다. 이에 최근에는 비감독(unsupervised mode)모드로 운영되는 클러스터링을 채택하는 분석 기법들이 많이 제시되고 있다. 클러스터링의 경우 별도의 질 좋은 학습 데이터를 찾으려는 수고를 덜어주고, 이 과정을 생략함으로써 시간을 단축 시켜 줄 수 있으며, 새로운 유형의 침입 탐지 패턴 추출에 보다 유용하다는 장점이 있다[4].

그러나 기존의 클러스터링을 이용한 분석 기법들은 점차 빅 데이터화 되어가는 네트워크 패킷들의 데이터 저장 및 이에 따른 성능적 결함에 대한 문제가 발생할 수 있음을 간과하고 있다. 즉, 축적되는 데이터의 양이 많아질수록 저장 공간 문제 뿐 아니라, 이를 처리하기 위한 성능적인 문제가 발생할 수 있는데 이전 연구들은 이러한 사항들을 고려하지 못하고 있다. 한편으로는 이러한 문제에 대응하기 위하여 데이터 압축과정을 통한 클러스터링을 구현하여 성능적인 결함을 개선하려는 방안을 제시한 바가 있다[5]. 그러나 이러한 방안은 데이터 저장소 자체적인 문제는 해결해 주지 못하고 있으며, 별도의 압축 과정 또한 시그니처 데이터베이스를 최신의 상태로 유지하는 데에 어려움을 줄 수 있다.

2.2 패킷 데이터 수집을 위한 NoSQL기반의 데이터 저장소

기존 RDBMS 방식의 데이터 저장소는 풍부한 데이터 모델을 기반으로 다양한 응용에서 요구하는 기능을 지원하는 반면, IT 기술 발전에 따라 기하 급수적으로 늘어나고 있는 많은 양의 데이터들을 처리하는 데 비용이 높고 확장성이 떨어지는 문제점을 갖고 있다. 이러한 문제점을 해결하고자 등장하게 된 기술이 'Not Only SQL'의 약자로 정의되고 있는 NoSQL방식인데, 이는 데이터 모델링을 위한 고정된 데이터 스키마가 없다는 것이 가장 큰 특징이다[6]. NoSQL은 비-관계형 데이터베이스로서 Key값을 이용하여 다양한 형태의 데이터 접근 및 저장이 가능하며, 데이터 모

표 1. NoSQL 분류
Table 1. The Classification of NoSQL

Data Model	Description
key-value	· The simplest data model · Range queries in the database , it is not easy without explicitly support · The Application modeling on the key-value storage may be have complexity
Document Store	· Document can have different schema unlike RDBMS · As RDBMS, It is possible to describe the relationship between the records · Conceptually, it is very similar to RDBMS
Column Store	· It has advantageous structure that reads mainly associated data · In order to change one of the records, it needs to modify multiple locations · Because the values of same domain are in a row, so it is good at compression efficiency · It is beneficial to the range query

델의 유형으로는 컬럼(column), 값(Value), 문서(Document), 그래프(Graph) 형식을 기반으로 구분되고 있다. [표 1]는 NoSQL을 데이터 모델 별로 분류하여 그 특징을 설명하고 있다.

이와 같은 속성을 지닌 NoSQL은 점차 늘어나고 있는 대량의 패킷들을 분석하여 침입 패턴을 추출하는 데 이용할 경우 몇 가지 이점을 얻을 수 있는데 이는 다음과 같다. 가장 첫 번째로는 자체 Scale-out방식의 데이터 저장 공간을 확장하여 낮은 비용으로도 높은 성능의 구동을 가능하게 해 줄 수 있다. 뿐만 아니라, 고정된 스키마가 없이도 Key값을 이용하여 다양한 형태의 데이터에 접근이 가능하므로 수많은 필드를 지닌 패킷 데이터들에서 원하는 필드 값들을 추출할 수 있는데, 이는 기존의 RDBMS와 같이 복잡한 관계를 설정하지 않아도 되는 용이성이 있다. 따라서 이러한 NoSQL 방식의 데이터 저장소의 이점들은 패킷 추출을 위한 데이터의 수집 시 낮은 비용 대비 고 효율적으로 이용 데이터 셋을 구성할 수 있도록 해준다.

본 제안하는 시스템에서는 다양한 종류의 NoSQL 솔루션 중에서도 문서(Document) 지향 형식의 MongoDB를 채택하였다. 이는 MongoDB의 문서 지향 형식이 열 단위가 아닌 행단위로 데이터를 처리하기 때문에, 비교적 적은 시간 안에 많은 양의 패킷 데이터를 읽어드려 수집이 가능하기 때문이다. 또한, 각 공격 별 유형을 분석하는데 필요한 데이터 필드만을 간단히 추출해 낼 수 있는 (key, value)값 기반의 구조적 유연성이 생산 비용 또한 줄여 줄 수 있는 이점이 있기 때문이다. MongoDB의 데이터베이스 구성은

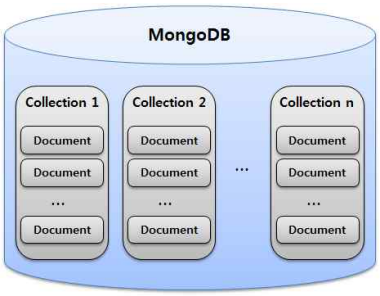


그림 2. MongoDB의 구조
Fig. 2. The Structure of MongDB

그림 2와 같은데, 기존의 RDMS에서 의미하던 테이블(Table)이 컬렉션(Collection)에 상응하고, 각각의 레코드(Record)들이 문서(Document)에 상응하는 구조이다.

2.3 기존 K-means 클러스터링의 맵리듀스 기법의 적용

클러스터는 같은 클러스터 내에서의 객체 집합은 유사성을 보이고, 다른 클러스터 간의 객체들의 성질은 상이성을 보이는 객체들이 집합을 의미함으로써, 데이터들 간의 전체적인 속성 및 패턴 또는 의미 있는 상관관계를 산출 해 낼 수 있다[7]. 이에, 앞서 2.1에서 언급되었듯이 클러스터링이 갖고 있는 이점에 따라 네트워크 보안을 위한 분석 분야에서 자주 사용되고 있다. 대표적으로 K-means 클러스터링 기법이 가장 대두되고 있는 방식인데, 이는 주어진 데이터를 특정 성질에 기초해서 k개의 군집으로 나누는 방법으로 표 2와 같이 동작 한다. K-means 클러스터링의 알고리즘적인 복잡도는 다음의 식 (1)과 같이 표현될 수 있는데, 이는 레코드의 개수(n)가 증가할수록 알고리즘의 수행 시간이 증가하여 속도가 느려짐을 알 수 있다[8].

$$\textit{Complexity of k-means clsutering algorithm} = k * n * O(\textit{distance metric}) * \textit{num(iterations)} \quad (1)$$

이러한 기존의 K-means 클러스터링 구조는 점차 증가하고 있는 대량의 패킷 데이터들을 분석하고 패턴을 추출하는데 성능적인 결함을 보일 수 있다. 이에 기존의 K-means 클러스터링 구조에 빅 데이터 환경에서의 대표적인 분산 프로그래밍 기법인 맵리듀스를 적용이 이루어지고 있다[9].

표 2. K-means 클러스터링의 수행 과정
Table 2. The Process of K-means Clustering

K-mean Algorithm
1: Select initial centroids among all points 2: While positions of centroids change 3: Find the points which are nearest to each centroid and make clusters having each centroid and corresponding points. 4: Select centorids agin in the latest clusters 5: end

표 3. map()과 reduce()의 구성
Table 3. The Composition of map() and reduce()

▫ map (in_key, in_ value) -> (out_key, intermediate _value) list ▫ reduce(out_key, intermediate_value list) -> out_value list

맵리듀스란 대용량 데이터를 병렬로 처리하기 위한 프로그래밍 모델로서, 맵(Map)과 리듀스(Reduce) 두 함수의 조합을 통해 분산/병렬 시스템을 운용하며 두 함수는 표 3과 같이 표현 될 수 있다[10]. 여기서 키(key)는 어떤 데이터에 대한 고유한 식별 값을 나타내고, 값(value)은 이 식별 값 각각 데이터의 해당 값을 나타낸다.
이러한 두 가지 이론적 배경을 종합하여 맵리듀스 기반의 K-means 클러스터링의 구현은 대량의 패킷 데이터들을 각각의 맵과 리듀스에 분산시켜 데이터양의 부담을 덜어줌으로써 성능적인 안정성을 보장할 수 있다. 또한 키-값 형식을 이용하여 원하는 데이터에 유연하게 접근할 수 있어 지속적으로 변화하는 공격 유형들이 갖는 속성들을 보다 빠르고 용이하게 반영 및 확장이 가능한데, 이러한 점은 시스템 설계의 생산비용 또한 덜어 주는 이점이 있다.

3.제안하는 방법

본 논문에서는 대용량화 되어가는 네트워크 패킷을 보다 효율적으로 수집 및 분석하여 나날이 증가하는 네트워크 위협 요인들을 보다 신속하고 안정적으로 탐지하는데 기여할 수 있도록 빅 데이터 기술 기반의 네트워크 패킷 분석 모델을 제안하고 있다. 그림 3은 본 패킷 분석 모델의 구성 및 처리 과정을 나타내고 있다. 제안하는 모델은 패킷 수집, 패킷 분석, 결과 추출의 순으로 프로세스가 수행되고 있으며, 3.1절은 수집 모듈에 관해 설명을 하고, 3.2절은 분석 모듈에 대하여 상세하게 설명하고 있다.

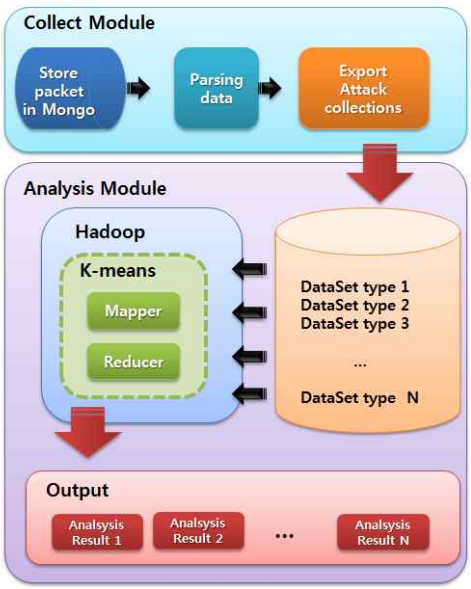


그림 3. 제안하는 패킷 분석 모델의 구성
Fig. 3. The Composition of Proposed Packet Analysis Model

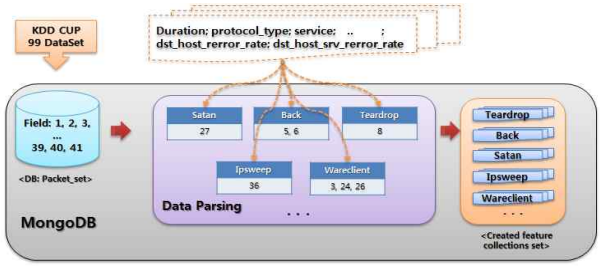


그림 4. MongoDB를 이용한 데이터 수집 및 파싱 과정
Fig. 4. The Process of Data Collecting and Parsing in Using MongoDB

3.1 NoSQL 기반의 패킷 수집 모듈

본 제안하는 패킷 분석 모델은 크게 NoSQL방식의 MongoDB와 Hadoop 두 가지의 프레임 워크로 구성되어 있으며, MongoDB 기반으로 구현된 수집 모듈과 Hadoop 기반으로 구현된 분석 모듈로 나누어 설명될 수 있다. 수집 모듈은 수집 모듈은 네트워크상에서 발생하는 대량의 패킷들을 수집하여 MongoDB에 저장함으로써 프로세스를 시작한다. 여기서 MongoDB의 이용은 다음과 같은 이유에서 사용한다. 첫 째로는, 빠른 읽기/쓰기 수행 능력이 대량의 패킷 데이터를 빠른 시간 안에 수집하는 데 이점이 있기 때문이고, 두 번째로는 문서 지향 형식의 유연한 구조는 각 공격 유형의 데이터 셋(DataSet) 마다 연관성 높은 특징 필드의 데이터들만을 용이하게 추출하여 구성할 수 있기 때문이다.

이러한 이점을 기반으로 데이터 가공단계인 Parsing 프로세스를 수행한다. 이는 수집된 패킷데이터들은 너무 많은 특징 필드(feature filed)를 포함하고 있기 때문에, 각 분석 대상의 공격 유형마다 갖는 특징 필드만을 별도로 추출하여 분석 대상 데이터 셋을 구성한다. 이는 시스템의 성능을 높이고, 개발 비용의 효율성을 높이기 위함이다. 본 논문에서는 KDD'99의 네트워크 패킷을 실험 데이터로 이용하였으며, 데이터 파싱 과정에서 이용되는 특징 필드들의 선별은 선 연구를 통해 산출된 자료를 기반으로 하고 있다[11]. 그림 4는 본 논문에서 제안하고 있는 수집 모듈의 처리 과정을 그림으로 도식화 한 것이다.

3.2 Map-Reduce 기반의 K-emans 클러스터링 분석 모듈

2.1절에서 언급하고 있는 클러스터링의 비감독(unsupervised mode)모드 특성의 이점을 바탕으로 K-means 클러스터링을 이용하여 분석을 시도하였다. 그러나 이전의 단순 K-means 클러스터링 기법으로는 점차 증가하고 있는 대량의 패킷들을 처리하는 데 성능적인 결함을 보일 수 있다. 이에 본 논문은 대용량 데이터를 위한 빅 데이터 기술을 적용하여 K-means 클러스터링 기반의 패킷 분석 프로세스의 성능을 개선한 것을 제안한다. 즉, 기존 K-means 클러스터링에 MapReduce 프레임워크를 적용하여 패킷 데이터 증가에 따른 알고리즘 복잡도를 줄여 프로세서의 성능을 개선하고자 한다.

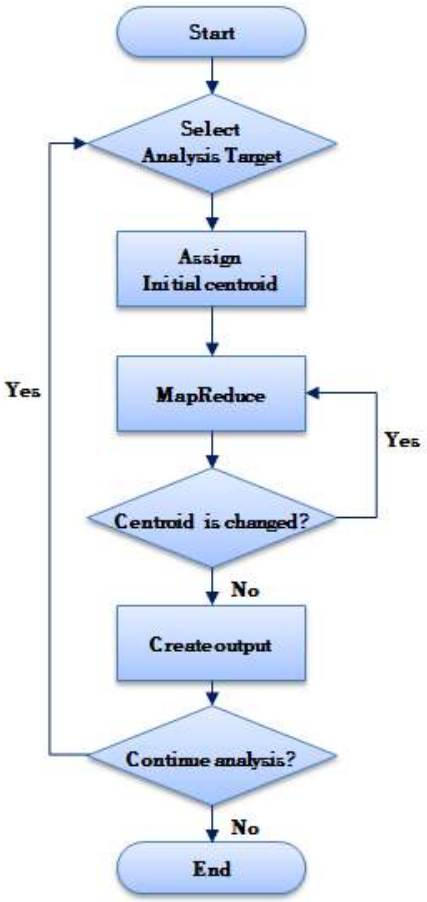


그림 5. 제안하는 모델의 분석 모듈 수행도
Fig. 5. The Work Process of Proposed Analysis Module

그림 5는 제안하고 있는 분석 모듈의 프로세스 처리 수행도이다. 분석 모듈은 특징을 추출해 내고자 하는 공격 대상의 Dataset type을 선택하면서 프로세스를 시작한다. 그 다음 클러스터 개수인 k개의 공통의 초기 중심점(Centroid)을 각 Map마다 할당해주고, Map() 함수는 사용자 정의에 의해 생성된 Map마다 전체 데이터를 나누어 초기 할당된 centroid를 기준으로 데이터들을 클러스터링을 수행한 후 거리 계산을 통해 새로운 중심점을 산출한다. 이에 각각의 Map()함수는 (cluster_id, data)의 key-value 값을 산출하게 되고, 이 값들은 Reduce()함수로 전달된다. 이에 Reduce()함수에서는 cluster_id에 따라 군집된 데이터들의 중심점들 위치 변화를 판단하고 더 이상의 변화가 없을 경우 최종 각 cluster_id에 따른 최종 중심점을 산출해낸다. 표 4는 분석모듈의 주요 단계인 Map()과 Reduce()함수가 중심점을 산출해 내는 과정의 소스코드이다. 이렇게 추출된 최종 중심점 값은 각 클러스터 집단 개체들이 갖고 있는 속성의 평균값으로 이는 분석 대상에 대한 주요 특징 또는 패턴을 추출해 낼 수 있는 주요 요소가 된다.

표 4. Map()과 Reduce() 함수의 중심점 계산 코드
Table 4. The Source code of computing centrioid source code in Map() and Reduce()

Compute centroid source code

```
while(centroid <= centroids) {
        if (center.distance(point) < minDist) {
          minDist = center.distance(point);
          minIndex = index;
        }
        index++;
      }
  }
```

4. 실험 및 결과

본 연구는 Intel Core i5 n3GHz의 CPU와 8G의 RAM을 사용하고 있는 Ubuntu-64bit의 운영체제에서 구현되었다. 실험에서 사용하고 있는 주요 두 가지 빅 데이터 프레임 워크인 MongoDB 2.2버전과 Hadoop 1.2.0버전을 이용하여 데이터를 수집하고, 높은 성능의 클러스터링을 구동시켜 공격 증가하고 있는 네트워크 패킷을 분석하는 실험을 구현하였다. 실험 데이터로는 KDD Cup 99 네트워크 패킷 데이터를 이용하였는데, KDD Cup 99는 총 41개의 특징 필드를 포함하고 있으며 정상인 연결기록과 비정상적인 연결기록 모두를 가지고 있는 데이터이다.

본 연구의 실험에서는 Teardrop, Back, Satan, Ipsweep, Wareclient 다섯 가지의 공격 유형을 대상으로 실험을 진행하였으며, 이는 KDD Cup 99 데이터 셋 안의 빈도수가 높은 공격 유형들을 선정한 것이다. 실제 KDD Cup 99의 전체 레코드수는 48,983,431개로 빅 데이터화된 패킷을 표현하기에 부족함이 있었으므로 공격 유형마다 그림 6과 같은 비율로 각 공격 유형의 패킷 데이터를 랜덤 하게 추출하고 이를 확장하여 총 100,000,000 개 레코드의 데이터 셋으로 재구성 하였다. 표 5는 선정된 공격 유형 별 패턴을 추출하는 데 연관성 있는 특징 필드를 정리해 놓은 것으로, 공격의 특징 collection들을 추출 시 key가 되고 있는 요소들이다. 실험은 다음과 같이 세 가지로 이루어 졌다.

[실험 1]은 제안하는 NoSQL 기반의 데이터 저장소가 기존의 RDBMS와 비교하여 발생하는 대량의 패킷 데이터를

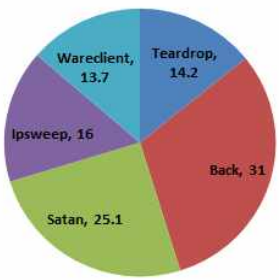


그림 6. 실험 데이터에 포함된 채택된 공격 비율
Fig. 6. The Proportion of Selected Attack Type in Test Data Set

표 5. 추출된 특징 필드
Table 5. The extracted feature field

Attack type	Feature fields
Teardrop	8: wrong_fragment
Back	5: src_bytes 6: dst_bytes
Satan	27: rerror_rate
Ipsweep	36: dst_host_same_rec_port_rate
Wareclient	3: service 24: srv_count 26: srv_serror_rate

얼마나 빠른 속도로 수집할 수 있는가를 알아보는 수집 성능 테스트이다. 각각 NoSQL 기반의 MongoDB와 RDBMS 기반의 MySQL을 통하여 100만, 500만, 1000만 건의 패킷 데이터를 수집하는 시간을 측정하여 비교하였다. [그림 7]은 이에 대한 결과를 그래프로 나타낸 것이며, 데이터의 양이 증가해 감에 따라 MySQL과 MongoDB 간의 속도차가 크게 벌어지는 것을 알 수 있다. 이는 대량의 패킷 데이터를 분석하기 위해 데이터를 수집할 시 NoSQL 기반의 데이터베이스가 뛰어난 성능을 보일 수 있음을 입증하는 바 이다.

[실험 2]는 데이터 레코드 숫자 증가에 따른 클러스터링의 밀집도를 통하여 공격 패킷 분석 시 다수의 트레이닝 개수를 확보할수록 질 좋은 분석 결과를 도출 할 수 있음을 보여주는 실험이다. [실험 2]의 데이터는 [실험 1]에서 수집한 41개의 특징 필드를 포함하고 있는 KDD Cup 99를 파싱(Parsing) 프로세서를 거쳐 공격 유형별로 선정된 관련 높은 특징필드 데이터만으로 구성된 5개의 공격 컬렉션을 기반으로 실험이 진행된다.

그림 8의 그래프 (a), (b), (c)는 다섯 가지 공격 유형 중 DoS공격 범주에 포함되는 Back공격에 대한 특징을 분석하기 위해 임의의 200개, 1000개, 10000개의 데이터를 추출하여 이를 K-means 클러스터링을 통해 분석한 결과이며, 이는 결과의 가시화를 위해 Weka 버전 3.7.9를 통해 출력하였다. v본 결과는 클러스터 개수를 임의로 k=5로 설정하여 산출한 결과이며 결과 (a)에서는 cluster 4, (b)에서는 cluster 0, (c)에서는 cluster 3이 Back 공격의 패킷 데이터를 가장 많이 포함하고 있다.

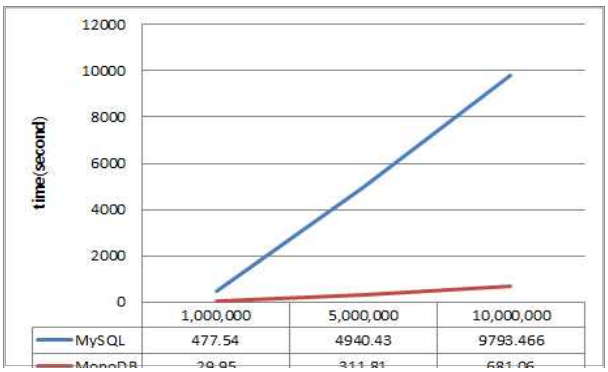


그림 7. 수집 테스트 결과
Fig. 7. The Result of Collect Test

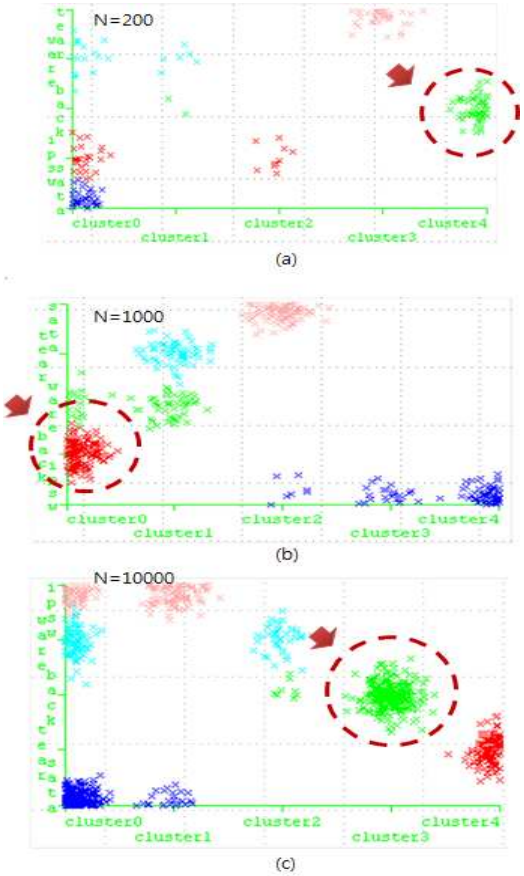


그림 8. 밀집도 테스트 결과
Fig. 8. The Result of Density Test

클러스터링 결과 레코드 N의 개체 수가 증가할수록 Back 공격에 해당되는 클러스터링의 밀집도 또한 높아지는 것을 볼 수 있다. 이는 선정된 특징 필드 '5: src_bytes',와 '6: dst_bytes'의 정보를 통해 올바른 클러스터링을 수행하고, 보다 농도 높은 Back공격의 특징을 추출해 낼 수 있음을 의미한다. 즉, 이는 데이터의 개수가 증가할수록 특징을 구성하는 개체 수가 많아져 침입 분석에 있어 보다 설득력 있는 정보의 추출이 가능함을 의미한다.

표 6. 제안하는 분석모듈의 성능 테스트 결과
Table 6. The result of analysis module's performance test

	10	100	1000	Average error
Existing Kmeans	1.07	14.29	40.22	2.69
Proposed Kmeans	4.02	6.79	29.39	0.511

(data unit: 10,000, unit time: second)

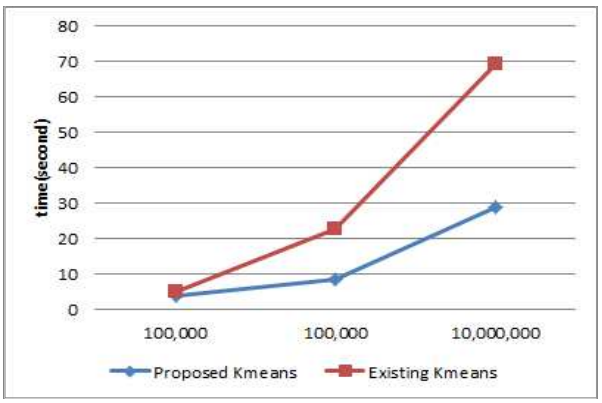


그림 9. 제안하는 분석모듈의 성능 테스트 결과
Fig. 9. The Result of Analysis Module's Performance Test

마지막으로, [실험 3]을 통하여 맵리듀스 기반의 K-means 클러스터링 분석 모듈의 속도를 측정하고, 기존 K-means 클러스터링 방식과 속도를 비교하여 제안하는 모델의 성능을 입증하였다. 실험 데이터로는 Back공격에 대한 특징을 추출하고자 임의의 10만개, 100만개, 1000만개의 패킷 데이터를 선정하였으며, 분석에 이용되는 필드는 [실험 2]에서와 마찬가지로 파싱된 특징필드를 사용하여 시스템 비용의 효율성을 높였다. 클러스터 개수는 임의로 k=5로 설정하였으며, 제안하는 방식의 K-means 클러스터링의 경우 4개의 map을 통하여 분산 처리되도록 모델을 설계하였다. 이를 수행한 속도를 측정한 결과는 표 6과 같이 산출되었으며, 이를 그래프로 나타낸 것이 그림 9이다. 표 6의 측정된 시간은 초단위로 구분되고 있으며 각각 10번씩의 수행결과에 대한 평균값을 기록한 것이다.

각각 데이터의 수가 작은 10만개의 데이터를 대상으로 클러스터링을 구동한 속도의 측정결과 기존 방법의 속도 측정치가 좀 더 빠르게 산출되었으나, 데이터의 개수가 증가할수록 제안하고 있는 K-means 클러스터링의 속도 측정결과가 우수하게 산출되었음을 알 수 있다. 또한, 기존 방법의 K-means 클러스터링의 성능 측정 산출결과 표준오차가 평균 약 2.69초로 다소 안정적이지 못한 결과를 나타내고 있다. 특히 기존 방식의 클러스터링을 1000만 건의 데이터를 대상으로 구동하였을 시 표준오차 값이 약 6.85초가 산출되었는데, 이는 데이터의 양이 증가할수록 클러스터링이 안정적이지 못하게 구동되고 있음을 의미한다. 반면, 제안하는 맵리듀스 기반의 K-means 클러스터링의 경우 1000만건의 데이터를 대상으로 구동 시에도 표준오차 약 1.11초로 보다 안정적으로 시스템이 구동되고 있음을 입증하였다.

5. 결론 및 향후 연구방향

본 논문에서는 증가하는 네트워크 패킷들로 인해 기존 보안 분석 모델들이 갖는 문제의 어려움을 자각하고, 이에 새로이 등장하고 있는 빅 데이터 기술을 이용하여 이러한 문제점을 해결할 것을 제안하였다. 즉, NoSQL 기반의 데이터 저장소를 통해 패킷 데이터 수집의 저장 공간 부족 문제를 해결하고, 맵리듀스 설계를 이용한 K-means 클러스터링 기반의 분석 모델을 통해 이를 분산 처리하여 이전 모델들의 성능적인 한계를 개선하고, 이에 대한 우수한 성능을 입증하였다.

그러나 본 분석 모델은 MongoDB의 맵리듀스 처리에 있어 섬세한 표현의 제약에 따라 Hadoop을 이용해 맵리듀스를 구현하였는데, 이는 데이터를 수집함과 동시에 분석을 처리하는데 한계를 가지고 있다. 따라서 향후에는 MongoDB 단일 프레임워크에서 맵리듀스 기반의 K-means 클러스터링을 통한 패킷 분석을 가능하게 하고, 더 나아가 이에 침입 탐지 시스템을 연동하여 실시간으로 빅 데이터화 된 패킷을 분석하고 여기서 발생하는 공격을 탐지하는 시스템을 연구하고자 한다.

References

[1] Dae-Soo Choi and Yong-Min Kim, "Big Data and Enterprise Security 2.0", *Journal of the Korean Institute of Information Scientists and Engineers(KIISE)*, vol. 30, no. 6, pp.65-72, Jun. 2012.

[2] Kim Hyun-Woo, Shin Seong-Jun, Lee Seung-Min, and Jeong Seok-Bong, "Network-based Intrusion Detection Scheme using Markov Chin Model", *Journal of Decision Science*, Vol. 20, No. 1, pp.75-88, Nov. 2012.

[3] Kim Sang Beom, "Reserach on development direction of network intrusion detection system", M.A., Yonsei University, 2008.

[4] Hansung Lee, Jiyung Song, Eunyoung Kim, Chulho Lee, and Daihee Park, "Adaptive Intrusion Detection System Based on SVM and Clustering", *Proceedings of KIIS Conference*, vol. 13, no. 2, pp. 237-242, Jun. 2005.

[5] Jong-Ha Ahn and Dae-Won Kim, "Compression-based Anomaly Detection using K-means Clustering, *Journal of the Korean Institute of Information Scientists and Engineers(KIISE)*, vol. 39, no. 8, pp. 605-612, Aug. 2012.

[6] Kyle Banker, *MongoDB in Action, Oreilly & Associates*, Aug. 2010.

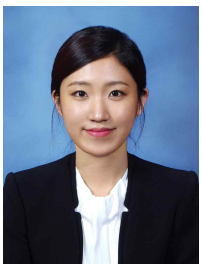
[7] Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Anomaly Detection : A Survey," *in ACM Computing Surveys*, vol.41 no. 3, Jul. 2009

[8] Kumar, Vipin, Pang-Ning Tan, and Michael Steinbach, *Introduction to data mining*, Addison-Wesley, 2005

[9] Zhao, Weizhong, Huifang Ma, and Qing He, *Cloud Computing*, Springer Berlin Heidelberg,

2009.
[10] Jeffrey Dean and Sanjay Ghemawat, “MapReduce: simplified data processing on large clusters,” *Communications of the ACM – 50th anniversary* , vol. 51, no. 1, pp. 107-113, 2008
[11] Olusola, Adetunmbi A., Adeola S. Oladele, and Daramola O. Abosede, “Analysis of KDD’99 Intrusion Detection Dataset for Selection of Relevance Features.” *Proceedings of the World Congress on Engineering and Computer Science.* vol. 1. pp. 20-22, Oct. 2010.
[12] Jaekwang Kim, KwangHo Yoon, Seunghoon Lee, Je-hee Jung, Jeehyong Lee, “A Slow Portscan Attack Detection and Countermove Mechanism based on Fuzzy Logic,” *INTERNATIONAL JOURNAL of FUZZY LOGIC and INTELLIGENT*, Vol.18, No.5, pp 679-684, 2008
[13] Kwee-Bo Sim, Jae-Won Yang, Dong-Wook Lee, Dong-Il Seo, Yang-Seo Choi, “Intrusion Detection System of Network Based on Biological Immune System,” *INTERNATIONAL JOURNAL of FUZZY LOGIC and INTELLIGENT*, Vol.12, No.5, pp 411-416, 2002
[14] Se-Yul Lee , Yong-Soo Kim, Kwee-Bo Sim, “A Study on Network based Intelligent Intrusion Prevention model by using Fuzzy Cognitive Maps on Denial of Service Attack,” *INTERNATIONAL JOURNAL of FUZZY LOGIC and INTELLIGENT*, Vol.13, No.2, pp 148-153, 2003
[15] Kwee-Bo Sim, Jae-Won Yang, Young-Soo Kim, Se-Yul Lee, “Intrusion Detection Learning Algorithm using Adaptive Anomaly Detector,” *INTERNATIONAL JOURNAL of FUZZY LOGIC and INTELLIGENT*, Vol.14, No.4, pp 451-456, 2004
[16] Kwee-Bo Sim, Jae-Won Yang, Dong-Wook Lee, Dong-Il Seo, and Yang-Seo Choi, “Adaptive Intrusion Detection Algorithm based on Learning Algorithm,” *INTERNATIONAL JOURNAL of FUZZY LOGIC and INTELLIGENT*, Vol.14, No.1, pp 75-81, 2004

저 자 소 개



최보민(Bomin Choi)
2012년 2월 : 경원대학교 컴퓨터미디어학과 졸업(공학사)
2012년~현재 : 가천대학교 일반대학원 전자계산학과 석사과정

관심분야 : Security, Algorithm, Big Data
Phone : +82-31-750-5818
E-mail : cbm0728@gmail.com

공 종 환 (Jong-Hwan Kong)
2012년 2월 : 경원대학교 컴퓨터 소프트웨어과 졸업(공학사)
2012년~현재 : 가천대학교 일반대학원 전자계산학과 석사과정

관심분야 : Network Security, Information Security
Phone : +82-31-750-5818
E-mail : ball3314@naver.com

한명묵(Myung-Mook Han)
1980년 : 연세대학교 공과대학 졸업 (공학사)
1987년 : 뉴욕공과대학교 컴퓨터공학과 석사 졸업 (공학석사)
1997년 : 오사카시립대학교 정보공학부 졸업(공학박사)
1998년～현재 : 가천대학교 IT대학 컴퓨터공학과 교수
1998년～현재 : 한국지능시스템학회 이사

관심분야 : Security, Data Mining, Algorithm
Phone : +82-031-750-5522
E-mail : mmhan@gachon.ac.kr