Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.25, NO.6, Dec. 2015 ISSN 2288-2715(Online)
http://dx.doi.org/10.13089/JKIISC.2015.25.6.1399

빅데이터 로그를 이용한 실시간 예측분석시스템 설계 및 구현

이 상 준,† 이 동 훈‡
고려대학교 정보보호대학원

Real time predictive analytic system design and implementation using Bigdata-log

Sang-jun Lee,† Dong-hoon Lee‡
Graduate School of Information Security, Korea University

요 약

기업들은 다가오는 데이터 경쟁시대를 이해하고 이에 대비해야 한다며 가트너는 기업의 생존 패러다임에 많은 변화를 요구하고 있다. 또한 통계 알고리즘 기반의 예측분석을 통한 비즈니스 성공 사례들이 발표되면서, 과거 데이터 분석에 따른 사후 조치에서 예측 분석에 의한 선제적 대응으로의 전환은 앞서가고 있는 기업의 필수품이 되어 가고 있다. 이러한 경향은 보안 분석 및 로그 분석 분야에도 영향을 미치고 있으며, 실제로 빅데이터화되고 있는 대용량 로그에 대한 분석과 지능화, 장기화되고 있는 보안 분석에 빅데이터 분석 프레임워크를 활용하는 사례들이 속속 발표되고 있다. 그러나 빅데이터 로그 분석 시스템에 요구되는 모든 기능 및 기술들을 하둡 기반의 빅데이터 플랫폼에서 수용할 수 없는 문제점들이 있어서 독자적인 플랫폼 기반의 빅데이터 로그 분석 제품들이 여전히 시장에 공급되고 있다. 본 논문에서는 이러한 독자적인 빅데이터 로그 분석 시스템을 위한 실시간 및 비실시간 예측 분석 엔진을 탑재하여 사이버 공격에 선제적으로 대응할 수 있는 프레임워크를 제안하고자 한다.

ABSTRACT

Gartner is requiring companies to considerably change their survival paradigms insisting that companies need to understand and provide again the upcoming era of data competition. With the revealing of successful business cases through statistic algorithm-based predictive analytics, also, the conversion into preemptive countermeasure through predictive analysis from follow-up action through data analysis in the past is becoming a necessity of leading enterprises. This trend is influencing security analysis and log analysis and in reality, the cases regarding the application of the big data analysis framework to large-scale log analysis and intelligent and long-term security analysis are being reported file by file. But all the functions and techniques required for a big data log analysis system cannot be accommodated in a Hadoop-based big data platform, so independent platform-based big data log analysis products are still being provided to the market. This paper aims to suggest a framework, which is equipped with a real-time and non-real-time predictive analysis engine for these independent big data log analysis systems and can cope with cyber attack preemptively.

Keywords: Bigdata, Advanced bigdata analytics, Predictive analytics, Preemptive Countermeasure, Log Management

Received(07. 16. 2015), Modified(09. 24. 2015), Accepted(09. 24. 2015)

† 주저자, sjunee@unet.kr
‡ 교신저자, donghlee@korea.ac.kr(Corresponding author)

I. 서 론

빅데이터는 단순히 데이터의 양이나 크기에 의한 것이 아니고, 데이터에 대해 최대한의 전수(全數) 분석을 통해 기존에 알 수 없었던 다양하고 신뢰할 만한 유의미한 결과를 도출하기 위해 제반 한계 상황 극복을 목표로 하는 프레임워크라 할 수 있다.

데이터 전수 분석을 위해 극복해야 하는 한계 상황(또는 특성) 중에는 당연히 데이터의 양과 크기가 포함되며, 속도, 다양한 데이터 유형(정형, 반정형, 비정형성) 및 처리 유형(실시간, 배치, 스트림, 근실시간) 등이 있다.

2001년 메타그룹의 애널리스트 더그 레이니는 3가지 관점에서 빅데이터를 정의하였는데, 이것이 3V(Volume, Velocity, Variety)이고, 가장 일반적인 빅데이터의 요소로 통용되고 있으며, IBM은 Veracity(정확도)라는 요소를 추가해 4V를 정의하였고, 포레스터 리서치의 브라이언 홉킨스 등은 Variability(가변성)을 추가하여 4V를 정의하였다.

이외에도 빅데이터에 대한 이해와 가독성 향상을 위한 Visualization, 빅데이터 활용에 따라 발생되는 가치(Value)를 포함하기도 한다.

또한 통계 알고리즘 기반의 예측분석(Predictive Analytics)과 결합하여 다양한 분야에 활용될 것으로 기대되고 있으며, 뉴욕대학 스턴(Stern) 경영대학원의 배선트 다르(Vasant Dhar) 교수는 "구글과 아마존을 포함하여 인터넷 시대를 이끄는 기업들은 기계학습 기반의 예측 모델에 의존하는 비즈니스 모델을 가지고 있다"고 할 정도로 빅데이터 기반의 예측 분석(Predictive Analytics) 모델에 의한 비즈니스 성공 사례들이 늘어나면서, 과거 데이터 분석에 따른 사후 조치에서 예측 분석에 의한 선제적 대응으로의 전환은 앞서가고 있는 기업의 필수품이 되어 가고 있다.

이러한 경향은 보안 분석 및 로그 분석 분야에도 영향을 미치고 있으며, 실제로 빅데이터화되고 있는 대용량 로그(빅데이터 로그)에 대한 분석과 지능화, 장기화되고 있는 보안 분석에 빅데이터 분석 프레임워크를 활용하는 사례들이 속속 발표되고 있다.

한편 하둡 기반의 로그 예측분석 제품이나, 특정한 용도로 하둡 기반 로그 예측 시스템을 구축하는 사례들도 보고되고 있으나, 범용성은 많이 떨어지는 편이다.

또한 로그 분석 시스템에 요구되는 모든 기능 및 기술들을 하둡 기반의 빅데이터 플랫폼에서 수용할 수 없는 문제점들이 있어서 독자적인 플랫폼 기반의 빅데이터 로그 분석 제품들이 여전히 시장에 공급되고 있다.

따라서 독자적인 플랫폼 기반의 빅데이터 로그 분석 제품에도 예측 분석에 대한 요구가 늘고 있는 상황이다.

본 논문에서는 빅데이터 로그의 특성을 살펴보고, 독자적인 빅데이터 로그 분석 시스템과 하둡 기반 빅데이터 플랫폼의 차이점을 분석하여, 궁극적으로 빅데이터 로그 분석 시스템을 위한 예측 분석 프레임워크를 제안하고자 한다.

II. 관련 연구

예측분석(Predictive Analytics)은 미래 또는 알 수 없는 이벤트에 대한 예측을 하기 위해 현재 또는 과거 사실의 분석 방법들인 모델링, 머신 러닝, 데이터 마이닝 관련 다양한 통계적 기술을 포괄하는 것이다.[1]

빅데이터 예측 분석은 빅데이터 플랫폼과 더불어 발전하고 있으며, 실시간 및 비실시간 예측 분석이 가능은 하지만, 전문가에 의한 프로그래밍이 필요하다.

빅데이터 로그 분석 시스템은 실시간•비실시간 분석 및 로그 검색 기능을 제공하나, 예측 분석 기능을 제공하는 제품은 없다.

각각의 연구 또는 기술 내용을 정리하면 다음과 같다.

2.1 빅데이터 플랫폼

빅데이터 = 하둡이라 해도 과언이 아닐 정도로 하둡 에코시스템은 글로벌 IT 기업에서 개발 및 운영으로 검증된 후 오픈 소스로 공개되면서 빅데이터 플랫폼으로 널리 사용되고 있다.

빅데이터에 필요한 기술들은 수집, 저장, 분석, 관리 및 모니터링 등으로 구분하며, 배치 처리 아키텍처와 실시간 처리 아키텍처로 분류하기도 한다.

하둡은 비즈니스 효율적으로 빅데이터 분석 시스템을 구축할 수 있도록 다양한 서브 프로젝트 형태의 공개 소스가 제공되면서 하둡 에코시스템이 구성되었다.

빅데이터는 단일 솔루션으로 해결할 수 없으며,

데이터의 성격, 비즈니스 요구사항 등에 따라 다양한 오픈 솔루션이 조합되어야 한다. 뿐만 아니라 한번 구축하고 관리만 하는 시스템이 아니라 지속적으로 진화시켜 나가야 하는 시스템이다.

그러나 이렇게 많은 하둡 에코시스템들은 완전하지 않거나 혹은 현재도 개발이 진행되고 있다. 이러한 신생 기술로 빅데이터 시스템을 구축하고 관리하며 대규모 데이터의 고급 분석을 수행한다는 것은 대단한 전문성과 기술 및 훈련을 요구한다.

2.2 빅데이터 로그의 특성 및 플랫폼

Fig.1.과 같이 로그관리시스템과 ESM을 통한 로그에 대한 접근 방식은 IT 인프라 및 네트워크의 고속화에 따라 데이터양의 급격하게 증가하고, 해킹 및 공격 방법이 지능화, 장기화되면서 로그 분석 및 SIEM 시스템으로 발전하게 되며, 빅데이터의 영향을 받아 빅데이터 로그분석 시스템이 등장하게 된다.

빅데이터 로그의 특성은 로그량 뿐만 아니라 광범위한 수집 범위(로그 유형) 및 수집 대상 장비의 대량화도 함께 나타나며, 처리 성능, 분석의 유연성 및 수집하는 로그에 대한 자동 분류 등의 기능이 함께 고려되어야 한다.

뿐만 아니라 빅데이터 로그 분석 시스템은 전통적

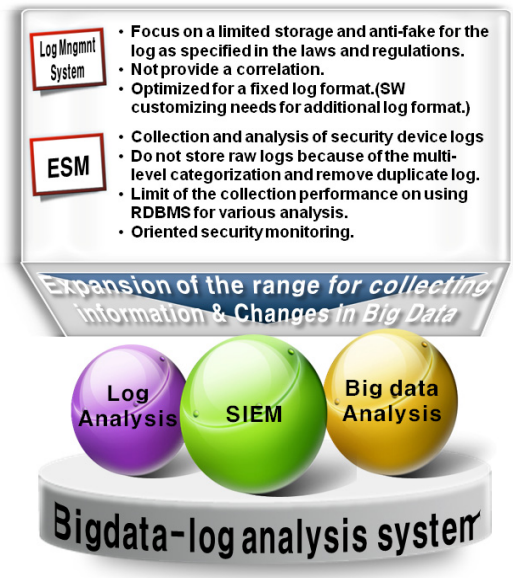

Fig. 1. The emergence background of big data analytics system

인 로그 관리 및 분석 시스템에 요구되는 Fig.2.와 같은 기능을 더불어 제공해야만 하며, Fig.3.와 같이 로그관리의 주요 목적과 함께 살펴 볼 때 주목할 것은 포렌식 분석, 법적 규제 대응 및 실시간 alert 부분이다.

포렌식 분석과 법적 규제 대응을 위해서는 원본 로그에 대한 저장 및 위변조 방지를 위한 기술을 포함해야 하며, 실시간 alert는 실시간 분석을 통해서만 지원 가능하다.

이러한 특징은 하둡 에코시스템을 활용한 빅데이터 로그 분석을 어렵게 하는 요인이 되며, 따라서 빅데이터 로그 분석시스템을 위한 독자적인 프레임워크도 함께 발전하고 있다.

필자가 관여하고 있는 빅데이터 로그분석 시스템은 파일 DB를 이용하여 수집·저장 성능을 확보하고 있으며, 병렬·분산 처리를 통해 실시간 분석 성능을 보장하고, 수집서버와 관리 서버를 분리하여 Scale out 방식의 확장이 가능한 구조를 채택하여 하둡 에코시스템과 유사하면서 단일 벤더의 일원화된 상용 아키텍처를 제공한다.

Fig.4.는 빅데이터 로그 분석 시스템의 논리적 구성도이며, 센터 매니저에 각종 보안 정책 및 모니터링 결과가 보관되며, 실제 로그의 수집 및 분석은 사이트 매니저에서 구현되는 구조이다.

사이트 매니저는 로그의 양에 따라 무한 병렬적으로(Scale Out) 증설이 가능한 구조이며, 실시간 분석은 각각의 사이트 매니저에서 실행된다.

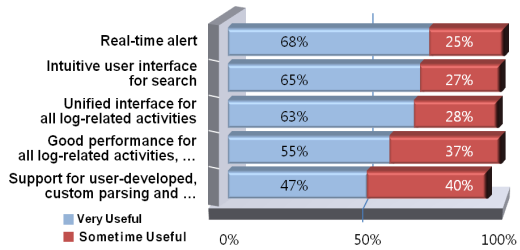

Fig. 2. Most Useful Feature(3)

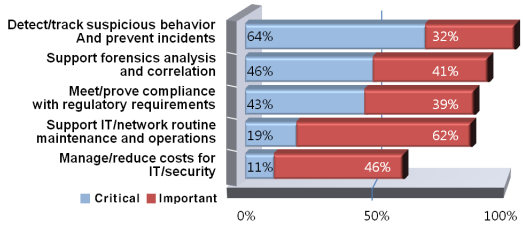

Fig. 3. Reasons for Collecting Logs(3)

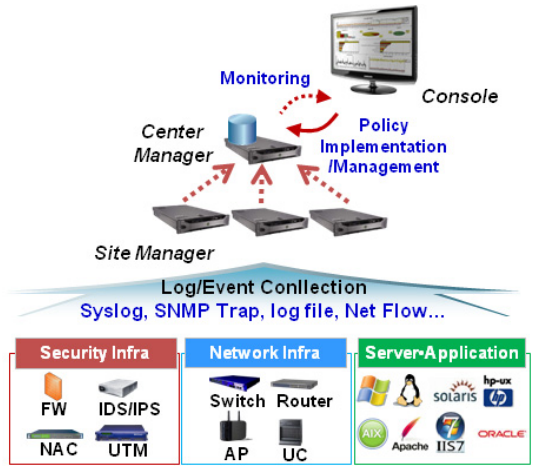

Fig. 4. Logical Architecture for Bigdata-log Analysis System

이러한 분산 분석 및 Scale Out 증설 방식은 하둡 에코시스템과 동일한 구조이며, NoSQL과 유사한 독자적인 파일 DB를 적용하여 검색 및 규제 준수가 가능하도록 구현하고 있다.

2.3 실시간 및 비실시간 로그 분석

2.3.1 실시간 로그 분석

실시간 분석은 분석에 필요한 모든 가용한 데이터를 활용하여 사용자가 분석을 수행하는 시점에 빠르고 적시에 지식을 제공해 줄 수 있는 분석 기법을 말한다.

하둡은 기본적으로 배치형태의 분산 처리 및 분석에 최적화되어 있는 프레임워크이나, 실시간 분석을 지원하는 storm과 같은 에코 시스템들을 통해 실시간 분석을 지원하고 있으며, 로그 분석 시스템에서도 병렬 분석(Fig. 6.) 및 Sliding Window 기법

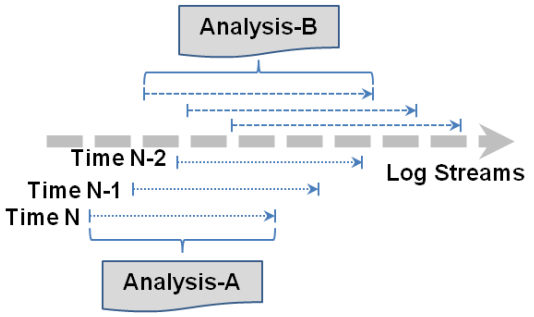

Fig. 5. Sliding Window Method

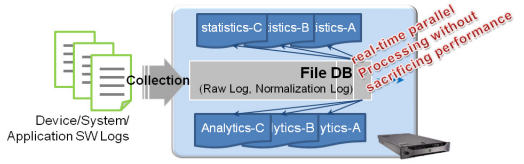

Fig. 6. conceptual model for Parallel Analysis

(Fig.5.)을 활용해서 실시간 분석을 지원한다.

실시간 분석은 현재 수집된 로그를 기준으로 각 분석 정책에 설정된 시간 만큼에 대한 과거 로그를 대상으로 분석하여 결과를 도출하는 기법을 적용하고 있으며, Fig.8.와 같이 도식화할 수 있다.

2.3.2 비실시간(시나리오) 분석

해킹 및 공격이 지능화되고 장기화되면서 세밀한 공격 시나리오에 대응하는 장기간에 걸친 분석이 반드시 필요한 기법이 비실시간 분석으로 단위 상관 분석 결과를 다음 상관분석의 입력으로 활용할 있도록 함으로써 단위 상관분석을 시간단위로 연결하는 시나리오를 완성할 수 있도록 한다.

Fig.7.은 외부 사용자의 개인정보 유출에 대한 네트워크 경로별 관련 시스템/장비를 시간 순서대로 분석한 것을 도식화한 것이다.

이렇게 도식화된 유통 경로 및 위협 분석 내용은

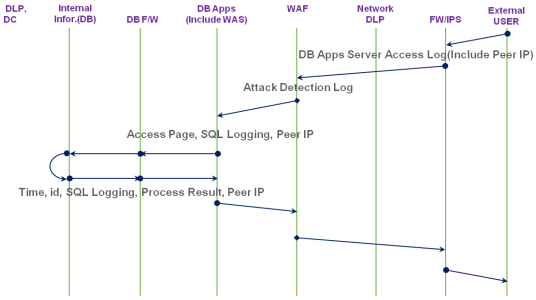

Fig. 7. Personal Information Leakage Path Analysis by external users

APT Threat Scenarios

Methods	Description
APT attack by the spread of malicious code	1. Upload malicious code through an external transit server hacking
	2. APT Attack with malicious code spread. - P2P, Portal Web Server, SNS, etc
	3. Obtain internal system information after infection
	4. Upload the data taken outside the transit server
	5. Data taken after the system accessible via the remote internal users
	6. Upload the data taken outside the transit server
	7. Destruction attack on the log and ...

Fig. 8. Detection scenario for APT

구체적인 탐지 시나리오로 구현되게 되며 Fig.8.은 APT 위협에 대한 탐지 정책(시나리오)이다.

2.4 예측 분석

2.4.1 빅데이터 예측 분석

KDnuggets에서 조사한 바에 따르면 2015년도 데이터 분석 도구로 가장 많이 사용된 도구는 'R'이다[5].

R은 뉴질랜드 오클랜드 대학 교수 로스 이하카와 로버트 젠틀맨의 주도하에 2000년에 버전 1.0으로 시작했으며, 이 두 교수의 이름 앞자가 R로 같다 보니, 이 패키지 이름을 R로 명명했다고 한다.[6]

R은 Warranty없이 사용 가능한 GNU Open Source이며, 통계 계산과 표현을 위한 환경 및 언어로서, 특징 및 장단점은 Table 1.과 같이 정리할 수 있다.

빅데이터 분석 관련 에코시스템은 기계학습·마이닝을 위한 Mahout, RHive 및 앞서 언급한 통계 언어인 R 등을 실시간·비실시간 분석 아키텍처에 프로그래밍을 통해 접목함으로써 구현될 수 있다.

이러한 점이 빅데이터 분석 시스템 구축에 있어 전문적인 기술력과 노하우를 요구하게 되며, 예측 분석을 위해서는 더욱 높은 수준의 기술력과 노하우가 필요하게 된다.

Table 1. Feature, Adv. & Disadv. of R[7]

Feature	Advantages	Disadvantages	Remark
In-Memory Architecture	Fast execution speed	Impossible to analyze big data	commercial R System
Open Source	Low cost Easy System Integration	Lack of training and technical support	
Language Structure	Easy to implement algorithms Detail Analysis	Required skills development program	S3, S4 Spec

2.4.2 로그 분석 시스템을 위한 통계 예측 엔진

R 또는 RHive를 실시간 로그 분석 시스템에 적용하기는 모듈 규모가 너무 크고, 별도의 프로그램 능력이 필요한 도구이기 때문에 경량화되고 쉽게 사용할 수 있도록 가장 많이 활용되는 알고리즘만을 선별하여 Fig.9.과 같은 구조의 통계 예측 엔진을 수학적 표준 라이브러리를 기반으로 개발하였다.

실제로 통계 예측 엔진은 로그 분석 시스템과 통합되었으며, 원본 로그를 이용한 직접적인 예측이 아닌 통계 DB에 저장되는 값을 기반으로 예측을 수행하는 모델이다.[9]

기존 로그분석 시스템에 부하 및 영향을 최소화하기 위해 로그분석 정책에 의한 결과물(예를 들면 지정 시간 동안의 특정 패킷 수)이 저장되어 있는 통계 DB를 예측엔진의 Input으로 설정하였으며, 이에 대한 일련의 처리과정을 거쳐 분석 결과를 로그분석 시스템의 콘솔에서 조회할 수 있도록 다시 통계 DB의 특정 영역에 주기적으로 저장하는 구조이다.

그러나 이러한 방법은 2차적인 통계 정보를 이용한 예측으로 배치분석만 가능하며, 정확성 측면 또는 의미있는 예측 측면에서 미흡함을 내재하고 있다.

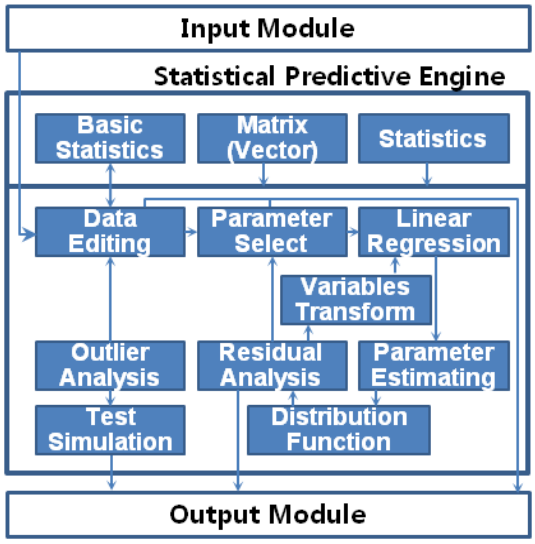

Fig. 9. Predictive Engine for Log Analytics System[8]

III. 실시간 및 비실시간 빅데이터 로그 예측 분석 시스템

3.1 고급 분석(Advanced Analytics)

Forrester의 James Kobielus는 고급분석 기술은 '비즈니스 상황을 예측하고 효율적인 의

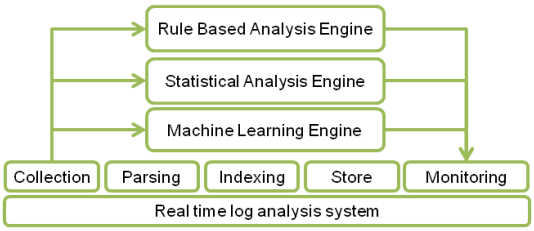

Fig. 10. Advanced Analysis Model for Proposed System

사결정을 지원하기 위해 구조화 및 비구조화된 복잡한 형태의 데이터에서 요인들 간의 상관관계와 의미있는 데이터의 패턴을 식별하고 예측하기 위한 모든 기법과 기술들'이라 정의하고 있으며, Neil Raden은 '기술분석(Descriptive Analytics), 예측분석(Predictive Analytics), 최적화(Optimization)'으로 고급 분석 기술을 분류하고 있다.

제안 시스템에서는 Fig.10.과 같이 고급 분석 기술을 지원하기 위한 아키텍처를 설계하였다.

룰 기반 분석 엔진이 기술 분석에 해당하고, 통계 분석 엔진은 예측분석이며, 머신 러닝과 모델 테스트 엔진(모델간의 평가)은 최적화에 해당된다.

기존에 개발한 로그분석시스템을 위한 통계 예측 엔진은 통계 DB를 기반으로 하였으나, 제한적인 분석 구조였기 때문에 원시로그를 그대로 활용할 수 있는 구조로 확장하였으며, 성능 이슈 해결을 위해 다양한 방안을 함께 검토해야 했다. 이런 의미에서 제한적 로그 예측분석 시스템이라 한다.

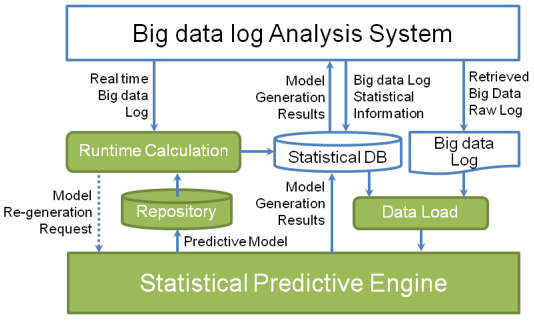

Fig. 11. Framework for Proposed System

3.2 실시간 예측 분석 모델

일반적인 실시간 분석은 CEP(Complex Event Processing)라고 하는 실시간 이벤트를 분석할 수 있는 기술이 개발되고 이를 기반으로 다양한 솔루션들이 출시되어 활용되고 있다.

ESP(Event Stream Processing)는 보다 많은 이벤트 스트림에 대해 Query나 수학적 알고리즘을 적용할 수 있도록 하는 기술이며, 빅데이터 분석 프레임워크들은 대부분 이 범주에 속한다고 볼 수 있다.

Apache S4나 트위터의 Storm, 아마존의 Hstreaming 및 IBM의 InfoSphere Stream 등이 있고, 트위터의 Storm이 요즈음 각광받고 있다.

실시간 분석은 위와 같은 분석 프레임워크에 R과 같은 통계 라이브러리를 적절히 통합하여 구현하게 되는 바, 쉽게 접근하기 어렵고, 더욱이 예측 분석을 실시간으로 구현하는 것은 이렇다 할 대안이 없는 실정이다.

이러한 현실을 감안하여, 제안시스템의 프레임워크를 Fig.11.과 같이 구현하였다.

우선 통계 DB를 경유한 예측 분석은 기존 선행 연구의 예측 분석 절차와 동일하며, 빅데이터 원본 로그를 검색 & 추출하여 예측 분석한 결과와 함께 통계 DB를 통해 빅데이터 로그분석 시스템에 전달된다.

"Runtime Calculation"은 실시간 예측 분석에서 가장 중요한 요소로 빅데이터 로그로 부터 생성되는 예측 모델을 이용하여 실시간으로 모든 로그에 대한 예측 추정값을 계산하고 결과를 반환하는 역할을 수행한다.

오른쪽의 "Data Load"는 빅데이터 로그 분석 시스템의 저장소로부터 분석 대상 로그를 추출하여 통계 예측 엔진에 전달하게 되며, 통계 예측 엔진은 해당 로그를 이용하여 예측 모델 생성 및 분석을 수행한다.

3.3 제안 시스템 설계 및 구현

3.3.1 제안 시스템 설계 구성 사항

3.3.1.1 전체 시스템 구성

우선 통계 DB를 경유한 예측 분석은 기존 선행

연구의 예측 분석 절차와 동일하며, 빅데이터 원본 로그를 검색 & 추출하여 예측 분석한 결과와 함께 통계 DB를 통해 빅데이터 로그분석 시스템에 전달된다.

통계 예측 엔진 자체가 실시간으로 동작할 필요는 없다. 통계 예측 엔진이 동작하는 경우는 예측 분석 모델을 생성 및 재 생성할 때와 "Data Load"를 통해 비실시간 분석을 하는 경우이다.

Fig.12.는 상세한 모듈 구성도이며, "Runtime Calculation" 모듈을 제3자에 제공하게 되면, 통계 예측 엔진은 빅데이터 로그 분석 시스템과는 독립적으로 운영이 가능한 구조를 갖게 된다.

"Data Load" 모듈은 다양한 형태의 입력을 지원할 수 있으나, 현재는 빅데이터 로그 분석 시스템의 원본 로그 저장소인 파일DB와 일반 RDBMS인 통계 DB 연동을 지원한다.

향후 빅데이터 로그 분석 시스템의 "Collect" 모듈에서 지원하는 다양한 형태의 수집 및 Data load 기법을 적용할 예정이다.

통계 예측 엔진은 배치 형태로 동작하게 되며, 빅데이터 로그 분석 시스템에서 저장한 원본 로그에서 필요한 데이터를 추출하여 예측 모델을 생성하게 되고, 생성된 모델을 'Runtime Calculation' 모듈을 통해 실시간 예측 추정치 계산을 하게 되며, 주기적인 비실시간 분석을 함께 수행하게 된다.

통계 및 예측 알고리즘은 가장 많이 사용되는 선

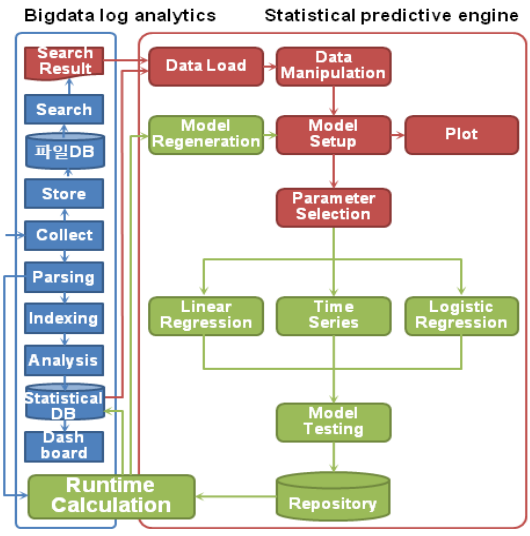

Fig. 12. Detailed module configuration of the proposed system

형 회귀 분석(Linear regression), 시계열 분석(Time Series) 및 로지스틱 회귀 분석(Logistic Regression)을 우선 개발하였다.

3.3.1.2 선형회귀분석(Linear Regression)

선형 회귀 분석은 2개 또는 그 이상의 변수 간에 인과관계가 있는지, 있다면 변수 별로 어느 정도의 비중으로 영향을 미치는지 분석하는 방법으로 다음과 같은 특성을 갖는다.

- 문제의 원인 분석에 사용
- 원인이 되는 요인 값을 예측 할 수 있다면 예측을 위한 용도로도 사용가능
- 실시간 분석 환경에서는 예측보다는 현재 수집 값에 대해 outlier 판정 및 이상치 검출

3.3.1.3 시계열분석(Time Series)

시계열 분석(Time Series)은 시간의 흐름에 따라 관측되는 자료를 바탕으로 일정 주기 별로 Data의 특성을 파악하고 미래의 값을 예측하는 분석 방법으로 다음과 같은 특성을 갖는다.

- 시간의 흐름에 따른 변동 요인
  - 추세 : 처음부터 계속 일정한 기울기로 움직이는 방향
  - 계절 : Data 한 주기 내에서 다시 일정 기간 별로 보여지는 패턴
  - 순환 : 이전 주기가 다음 주기에 영향을 미치는 경우
  - 잔차 : 위 세가지의 변동으로 예측되지 않는 잡음
- 원격 장비 장애 예측의 경우 장비 데이터의 특성상 특정한 주기를 가지고 있지 않음
  - 데이터의 변동에 따른 짧은 주기의 예측 위주로 사용 (이동평균법 등)

3.3.1.4 로지스틱 회귀분석(Logistic Regression)

로지스틱 회귀분석(Logistic Regression)은 추정하고자 하는 값이 True/False와 같이 이분형일 때 사용하는 분석방법이며 다음과 같은 특성을 갖는다.

- 회귀 분석이 연속된 데이터를 분석/예측 하는 방법

- 결과 값이 확률로 표시 될 수 있기 때문에 필요한 Data가 갖춰진다면 장애가 발생할 확률을 표현할 수 있음
- 원격 장비 장애 예측의 경우 장애 발생 위험성 및 장비의 한계 수명 등 판단에 사용

3.3.2 시스템 구현 및 성능 평가

3.3.2.1 구현 화면

Fig.12.와 같이 제안시스템의 통계 예측 엔진은 빅데이터 로그 분석 시스템에 Plug-in되는 구조를 갖기 때문에 독자적인 화면은 많지 않다.

최종 디자인 과정이 남아있어 좀 거칠기는 하지만 PoC가 완료된 결과물을 통해 동작 절차를 살펴보면 다음과 같다.

- 분석 대상 빅데이터 로그의 로드(load)를 위한 설정
- 빅데이터 로그의 로드(extract) 및 결과 확인
- 분포 분석
- 모델 생성을 위한 설정
- 생성된 모델을 Repository에 저장
- 저장된 모델 기반의 Runtime Calculation

먼저 Fig.13.은 빅데이터 로그 분석 시스템의 저장소로부터 필요한 데이터를 추출(Extract)하기 위한 설정화면으로 Category는 빅데이터 로그 분석 시스템에 등록되어 있는 내용 중에 선택하게 된다.

추출 결과는 Fig.14.와 같으며, 각 필드(또는 변수)간의 일반적인 관계도는 Fig.15.와 같이 표현된다.

Fig.16.은 예측 분석 모델 생성을 위한 설정 화면이며, 실행 결과로 생성된 모델은 Fig.17.과 같다.

Fig.17.은 운영자가 볼 필요는 없는 화면이며, 이 내용이 Repository에 등록된다.

등록된 Repository의 예측 모델이 'Runtime Calculation'에 적용되기 위해서는 기존의 빅데이터 로그 분석 시스템의 실시간 분석 정책으로 해당 내용이 포함되어 설정되어야 한다.

Fig.17.은 다중회귀분석을 통해 특정 장비의 장애 발생 가능성을 예측하기 위한 모델을 가정한 것이다.

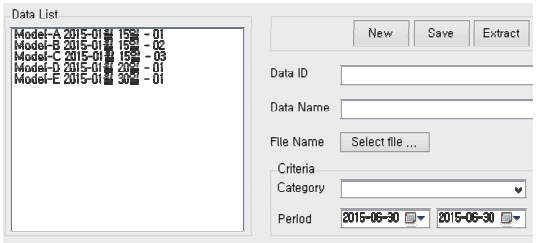

Fig. 13. Configuration menu for loading big data log

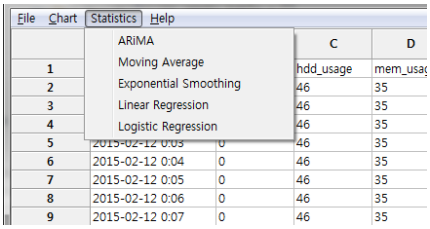

Fig. 14. Results for log extraction

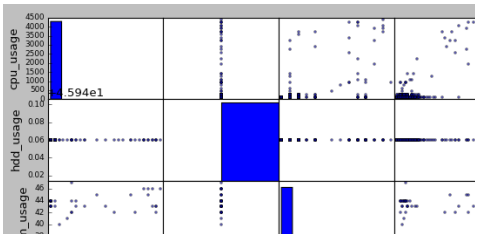

Fig. 15. Correlation graphs between variables

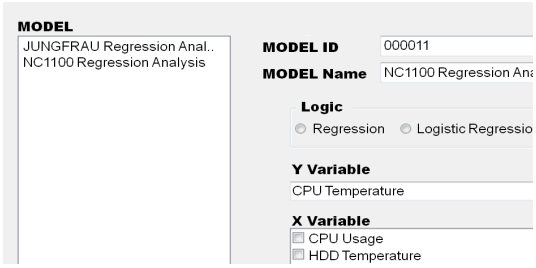

Fig. 16. Configuration menu for generating a prediction model

회귀분석에서는 예측 목적이 되는 변수를 종속변수(PoC에서는 CPU 온도)라 하고, 종속변수(PoC에서는 CPU 사용량, 메모리 사용량, HDD 온도)에 영향을 주는 변수를 독립변수라 한다.

종속변수와 독립변수의 상관관계를 모델이라 하며, 종속변수를 Y, 독립변수를 X_1, X_2, X_3, ...라 가정하면 $Y = a + bX_1 + cX_2 + dX_3$, ...와 같

OLS Regression Results

Dep. Variable:	cpu_temp	R-squared:	0.637
Model:	OLS	Adj. R-squared:	0.637
Method:	Least Squares	F-statistic:	2527.
Date:	Tue, 23 Jun 2015	Prob (F-statistic):	0.00
Time:	14:17:36	Log-Likelihood:	-5810.0
No. Observations:	4320	AIC:	1.163e+04
Df Residuals:	4316	BIC:	1.165e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
const	16.3388	1.703	9.592	0.000	12.999	19.678
cpu_usage	0.7938	0.012	67.313	0.000	0.771	0.817
mem_usage	0.1086	0.020	5.511	0.000	0.070	0.147
hdd_temp	0.6798	0.058	11.690	0.000	0.566	0.794

Omnibus:	3175.854	Durbin-Watson:	1.027
Prob(Omnibus):	0.000	Jarque-Bera (JB):	236190.768
Skew:	2.848	Prob(JB):	0.00
Kurtosis:	38.773	Cond. No.	5.35e+03

Fig. 17. As a result of generating a predictive model

은 방정식으로 표현된다.

따라서 Fig.16.의 PoC에서 해당 장비의 CPU 온도를 Y라 하면, Y = 16.3388 + 0.7938 * (CPU 사용량) + 0.1086 * (메모리 사용량) + 0.6798 * (HDD 온도)와 같은 방정식이 도출되며, 실시간으로 수집되는 각각의 로그 값과 연산을 통해 결과값(예측값)이 생성되고 경보를 발생하게 된다.

3.3.2.2 예측 결과

PoC의 예측 대상을 CPU 온도로 설정한 것은 좀 더 유의미한 예측(예를 들면 장비 장애)을 위해서는 해당 예측을 동반하는 많은 량의 로그(장비 장애 상태의 독립 변수값)가 필요하나 현실적으로 어렵기 때문에 함께 수집이 가능한 CPU 온도를 종속변수로 나머지 값들을 독립변수로 가정한 것이다.

제안시스템은 수집하는 로그 중에서 독립변수를 설정하고, 종속변수를 예측하는데 있어 가장 적합한 예측•분석 알고리즘을 선별(선형분석, 회귀분석, 로지스틱 회귀분석, 의사결정트리 등)하며, 자동으로 선별된 모델을 개선하는 프레임워크를 제공하는 시스템이 목표이므로 쉽게 종속변수를 비교할 수 있는 측정 가능한 값을 선택한 것이다.

Fig.18.은 PoC 환경에서 수집한 독립 변수들을 이용해서 CPU 온도를 예측하고, 실제 측정된 CPU 온도와 비교한 그래프이며, 상당히 정확하게 예측되는 사실을 확인할 수 있다.

동일한 방법으로 CPU 사용량이나 메모리 사용량 또는 HDD 온도를 종속변수로 예측할 수 있다.

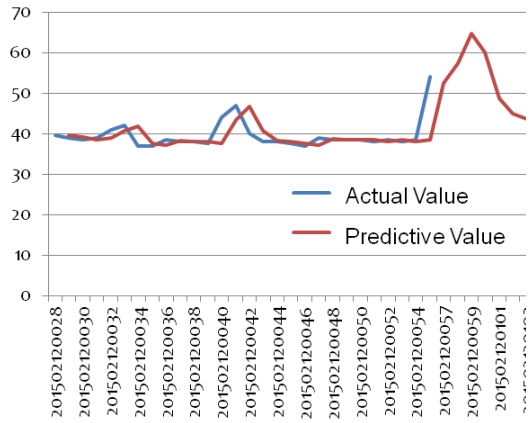

Fig. 18. Predictions for CPU temperature

한편 위 결과에 포함되지 않았던 "팬 속도"를 독립변수로 추가해서 유사한 형태의 그래프이지만 좀 더 높은 정확성을 갖는다는 것을 확인하였다. 이는 종속변수 예측에 있어 관련있는 독립변수의 발굴이 얼마나 중요한지를 단적으로 보여주는 것이다.

독립변수를 UI를 통해 손쉽게 추가•삭제할 수 있는 본 제안시스템을 이용하여 보다 많은 예측할 수 없었던 독립변수들을 찾을 수 있을 것이라 기대한다.

3.3.2.3 성능 평가

본 시스템은 빅데이터 로그 분석 시스템을 확장하여 실시간 및 비실시간 예측 분석이 가능하도록 하는데 목적이 있으며, Table 2.는 빅데이터 고급 분석을 위한 하둡 에코 시스템과 비교한 것이다.

Table 2. Comparison between the proposed system and the hadoop ecosystem

ITEM	Proposed System	Hadoop ecosystem
System complexity	Simple	Complex
Log analysis function (Compliance, Crypto, etc)	provide	Programming is required
Implementation difficulty	Normal	High
Statistical experts	needless	Need
Implementation cost	Low	High
Operating costs	Low	High
Maintenance Provider	Delivery Company	unclear
Predictive analysis	provide	Programming is required
Real time Predictive analysis	provide	Programming is required

3.4 적용 사례

제안시스템의 가장 큰 목적은 사이버 공격에 대한 예측을 통한 선제적 대응이다.

그러나 기술적인 측면에서 보면 사이버 공격에 대한 예측이나 IT 또는 IoT 디바이스에 대한 장애 예측 및 금융 사기 예방 시스템의 사기 행위 예측 등은 모두 동일한 예측 프레임워크로 대응이 가능하다.

다만, 사이버 공격 발생 당시의 로그와 일반적인 로그, 장애 발생 로그와 비발생 로그 및 사기 발생 로그와 비발생 로그를 구분하여 확보할 수 있어야 하며, 이를 통한 예측 모델링 과정이 반드시 필요하며, 이러한 과정을 통해 예측 목적에 맞는 로그 수집 내용 및 범위에 대한 접근도 아울러 가능해 진다.

3.4.1 IoT 디바이스 장애 예측 분석

현재 적용 완료되어 가동되고 있는 원격 장비 유지관리를 포함하는 장애 예측 관련 적용 사례를 통해 제안 시스템의 가능성을 확인해 볼 수 있다.

해당 사례는 궁극적으로 IoT 장애 예측으로 확대·통합될 것으로 기대하며, 보안 장비·네트워크 장비에 대한 장애 예측, 공장 자동화 또는 생산 설비에 대한 장애 예측 등을 포함할 수 있을 것이다.

원격 장비로부터 관련 로그를 수집하여, 실시간으로 각 장비별로 장애를 예측하고, 장애 확률이 높은 장비에 대해서는 모니터링 요원에게 alert을 하고, 헬프데스크를 통한 선제적 조치 또는 현장 엔지니어의 사전 방문 수리를 한다.

이러한 선제적 장애 예방 프로세스를 통해 장애 발생 후 콜 접수를 통한 지원 대비 장비 down time 감소 및 유지 관리 비용 절감 등의 정량적 효과가 발생하고 있다.

Fig.19.는 장애 예측 관련 화면으로서 장애 발생 확률 TOP N에 대한 원인 및 조치 방안에 대한 대시보드이다.

Fig.19.에서 장애 예측 가능성이 높다고 해서 장애가 발생하지는 않는다. 이는 예방적 활동에 의해서이기도 하고, 충분한 운영 시간을 갖지 못한 결과이기도 하다. 그리고 실제 장애 건수가 극소수에 불과하기 때문에 장애 예측 결과와 실제 장애를 비교하여 성능을 측정하기에는 무리가 있다.

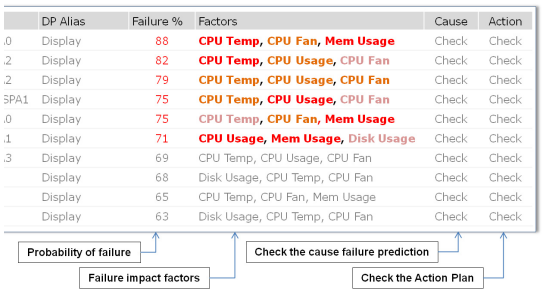

Fig. 19. Dashboard for Failure Prediction

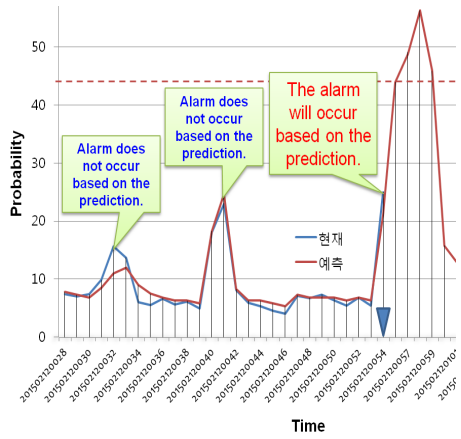

Fig. 20. Performance for Failure Prediction

앞으로 장기간의 운영 경험과 로그를 이용하여 좀 더 정교한 모델로 발전시켜야 할 것이며, 장애에 영향을 미치는 독립변수를 꾸준히 발굴하는 노력이 필요한 시스템이다.

본 논문에서는 각각의 독립변수에 대한 시계열 분석에 의한 예측치를 이용한 장애 예측값과 실제 수집되는 독립변수에 의해 계산되는 예측값의 비교를 통해 프레임워크의 안정성 및 신뢰성에 대한 검증을 대신하며, Fig.20.에서와 같이 신뢰도 높은 결과를 얻을 수 있었다.

3.4.2 프로파일링 기법

프로파일은 행위 기반의 평균값을 다양하게 모아 놓은 저장소 정도로 정의할 수 있다.

예를 들어 금융 사기 방지 시스템인 FDS의 경우 각 계정 또는 사용자별로 일회 평균 이체 금액, 일일 평균 이체 횟수, 주요 이체 지역, 시간 등등을 데이터 마트에 저장해 놓고, 이체 금액, 금일 누적 이체

횟수와의 비교를 통해 사기 여부를 판단하게 된다.

한편 기업 내부의 개인정보 유출 모니터링을 위해서도 임직원별로 다양한 프로파일이 필요하다.

DRM 문서 복호 평균 건수, 외부 메일에 파일 첨부 평균 건수, PC 사용을 위한 인증 실패 또는 성공 건수 등등 다양한 평균이 프로파일되고 비교되면서 이상행위를 판단하는 지표로 활용되고 있다.

그러나 그 동안의 프로파일은 단순 평균이거나 트렌드 기법이라고 하는 표준편차를 활용하면서 아웃라이어를 제거하는 정도이지 앞서 살펴 본 바와 같은 종속변수에 대한 독립변수와의 상관관계 등을 고려되고 있지 않다.

예를 들어 평일 평균 10건, 주말에는 1건의 DRM 문서를 복호하는 임직원이 있다면, 전체 평균은 7.42(건/일)가 될 것이다.

오늘이 일요일이고 해당 임직원이 7건의 DRM 문서를 복호했다면 기존의 방법으로는 이상행위에 대한 알람이 발생하지 않는다.

그러나 제안시스템에서 종속변수를 DRM 복호 일 평균 건수로 하고, 독립변수를 요일, 공휴일 여부, 출장 여부, PC On/Off 여부 등을 포함하여 분석한다면 훨씬 정확한 해당 일에 대한 평균이 예측될 것이라는 것은 자명하다. 또한 모델에 의한 독립변수의 계산만으로 종속변수 값이 예측되기 때문에 별도의 데이터마트가 필요하지도 않다.

이외에도 방화벽 로그를 이용한 공격 예측 모델링을 진행하고 있으며 좀 더 폭넓은 모델링을 위해 현업 시스템에 시범 적용 중이다.

IV. 결론 및 향후 과제

빅데이터 및 빅데이터 로그에 대한 예측 분석의 중요성은 점점 더 높아져 갈 것으로 보인다.

'빅데이터의 다음 단계는 예측분석이다'라는 에릭 시걸의 저서명을 언급하지 않더라도 아주 많은 성공 사례들이 보고되고 있고, 구축되어 가고 있다.

본 논문에서는 로그 관리/분석을 위한 기본 기능을 지원하면서 빅데이터 로그 분석과 함께 예측 분석이 가능한 프레임워크를 제시하였다.

이는 현업 보안 담당자나 로그 관리자가 빅데이터나 빅데이터 로그 분석 시스템에 전문적이 소양이 없어도 쉽게 예측 분석에 다가설 수 있도록 하는 계기가 될 것이며, 다음과 같은 효과가 있을 것으로 기대한다.

- 수집된 로그의 데이터를 분석하여 미래에 발생할 상황(장애, 공격)등을 예측하고 더 나아가 예방에 대한 정보를 제공
  - 데이터의 추세 분석을 통해 미래의 장애요인 예측
  - 상황의 원인이 되는 요인을 파악, 사전에 위험 요소를 제거함으로써 최적의 안정된 상태를 유지
- 보관된 데이터를 분석하여 기존 상황에 대한 원인을 분석하고 영향을 준 항목에 대한 정보를 제시
  - 기존 장애 발생 전후의 데이터를 분석하여 상황에 대한 원인을 설명
  - 원인 분석을 통해 도출된 위험 요소는 사전 관리 대상 또는 중점 관리 지표로 관리
- 기존 데이터의 정보를 이용하여 새로운 데이터를 구분, 분류하고 사용자의 의사결정을 지원
  - 특정 환경, 요인에 따른 대상의 분류
  - 사용자의 의사결정을 도와주기 위한 보조 데이터를 제시

또한 IoT 환경에서 거대화되고 있는 IoT 로그에 대한 분석 및 예측 프레임워크로서 손색없는 역할을 할 것으로 기대하며, 이를 위해 좀 더 정교하고 고도화된 프레임워크로 진화하기 위해 다음과 같은 통계 및 분류 알고리즘을 추가적으로 개발할 예정이다.

- 추가 예정 통계 알고리즘
  - 생존 분석 : 생명체의 사망까지 시간이나 장비의 고장까지 시간을 측정한 데이터에 대한 분석 방법
  - 원격 장비 유지 관리를 위한 장애 예측 시스템에 적합할 것으로 기대
- 추가 예정 분류 알고리즘
  - 군집 분석
  - 의사 결정 나무 (Decision Tree)
  - AHP (Analytic Hierarchy Process)

끝으로 모델 검증과 함께 고급분석 최적화(Optimization)의 한 축이 되는 머신러닝을 다음과 같이 개발하여 자가 발전 모델을 완성하는 것을 향후 과제로 남겨 놓는다.

- 베이지안(Baysian) 분석
  - 표준 분포를 따른다고 가정하고 입력된 데이터에 따라 우도를 수정하여 적절한 모델을 찾는 방법

- 장애 예측 응용 : 사전 데이터를 분석하지 않고 지속적으로 입력되는 데이터만으로 장애 모델을 추출

● 뉴럴 네트워크 (Neural Network)
- 동물의 뇌와 같은 신경망에서 영감을 받아 만들어진 통계적인 분석 방법
- 장애 예측 응용 : 사전 데이터를 분석하지 않고 지속적으로 입력되는 데이터만으로 장애 모델을 추출

References

[1] WIKIPEDIA, "Predictive analytics," Retrieved May. 27. 2015 from https://en.wikipedia.org/wiki/Predictive_analytics

[2] Quinton Anderson, "Storm Real-Time Processing Cookbook," acorn publishing, 2014

[3] Jerry Shenk, "SANS Seventh Annual Log Management Survey Report," pp. 5-6, April. 2011.

[4] Sang-Jun Lee et al. (Unetsystem), "Unusual action decision system," Patent 10-2013-0134805, 2013.

[5] Gregory Piatetsky, "R leads Rapid Miner, Python catches up, Big Data tools grow, Spark ignites," Retrieved May. 17. 2015 from http://www.kdnuggets.com/2015/05/poll-r-rapidminer-python-big-data-spark.html

[6] Sung-Min Hong, "Open Language, R Language! Beyond that age," Cheil Communications(Magazine of Cheil Worldwide Inc.), 2004.

[7] Choong-Hyun Yoo, "Technology Trends in Big Data Analytics and Introduction to R," NexR, 2012.

[8] Kwang-Man KO, Beom-Chul Kwon, Sung-Chul Kim, Sang-Jun Lee, "Development of Statistical Prediction Engine for Integrated Log Analysis Systems," Journal of The 2013 Fall Conference of the KIPS, Vol. 20, No. 2, 2013.

[9] Sang-Jun Lee et al.(Unetsystem), "Integrated log analysis system," Patent 10-1484290, 2015.

〈저자소개〉

이 상 준 (Sang Jun Lee) 정회원
1992년 2월: 동국대학교 전자계산학과 졸업
2012년 2월: 고려대학교 정보경영공학전문대학원 석사(수료)
1993년 7월~1994년12월: 삼성전자 컴퓨터시스템사업부 SW 개발실
1995년 1월~2000년 3월: 삼성SDS 공공개발팀
2000년 4월~2003년 2월: 시큐아이닷컴 PKI 개발팀장
2003년 3월~현재: 유넷시스템 무선보안연구소 소장
〈관심분야〉 빅데이터 로그 분석, 무선랜 보안, 네트워크 보안, PKI

이 동 훈 (Dong Hoon Lee) 종신회원
1983년 8월: 고려대학교 경제학과 학사 졸업
1987년 12월: Oklahoma University 전산학과 석사 졸업
1992년 5월: Oklahoma University 전산학과 박사 졸업
1993년 3월~1997년 2월: 고려대학교 전산학과 조교수
1997년 3월~2001년 2월: 고려대학교 전산학과 부교수
2001년 3월~현재: 고려대학교 정보보호대학원 교수
2015년 ~ 현재: 고려대학교 정보보호대학원 원장
〈관심분야〉 암호프로토콜, 암호이론, USN 이론, 키 교환, 익명성 연구, PET 기술