Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.24, NO.5, Oct. 2014 ISSN 2288-2715(Online)
http://dx.doi.org/10.13089/JKIISC.2014.24.5.871

이미지 및 코드분석을 활용한 보안관제 지향적 웹사이트 위·변조 탐지 시스템*

김 규 일,[1†] 최 상 수,[1] 박 학 수,[1] 고 상 준,[1,2] 송 중 석[1,2‡]
[1]한국과학기술정보연구원, [2]과학기술연합대학원대학교

Website Falsification Detection System Based on Image and Code Analysis for Enhanced Security Monitoring and Response*

Kyu-il Kim,[1†] Sang-soo Choi,[1] Hark-soo Park,[1] Sang-jun Ko,[1,2] Jung-suk Song[1,2‡]
[1]Korea Institute of Science and Technology Information,
[2]Korea University of Science & Technology

요 약

최근 경제적 이윤을 목적으로 한 해킹조직들이 국가 주요 웹사이트 및 포털사이트, 금융 관련 웹사이트 등을 해킹하여 국가적 혼란을 야기 시키거나 해킹한 웹사이트에 악성코드를 설치함으로서 해당 웹사이트를 접속하는 행위만으로도 악성코드에 감염되는 이른바 'Drive by Download' 공격이 빈번하게 발생하고 있는 실정이다. 이러한 웹사이트를 공격목표로 하는 사이버 위협에 대한 대응방안으로 웹사이트 위·변조 탐지 시스템이 주목을 받고 있으며, 국내에서는 국가사이버안전센터(NCSC)를 중심으로 분야별 사이버 보안을 담당하는 부문 보안관제센터에서 해당 시스템을 구축·운영하고 있다. 그러나 기존 위·변조 탐지기술의 대부분은 위·변조 탐지 시간이 오래 걸리고 오탐율 또한 높기 때문에, 신속성 및 정확성이 중요한 보안관제 분야에서는 직접적 활용이 어렵다는 문제점을 안고 있다. 따라서 본 논문은 웹사이트 위·변조 탐지시스템의 오탐률을 최소화하고 실시간 보안관제에 활용하기 위해 이미지 및 코드 분석기반의 웹사이트 위·변조 탐지 시스템을 제안한다. 제안 시스템은 웹크롤러에 의해 비교검증의 대상이 되는 정보만을 수집하고 정규화를 통해 위·변조 판별에 영향을 미치는 이미지 및 코드를 추출하여 유사도를 분석하고 이를 시각화함으로서 보안관제요원의 직관적인 탐지 및 웹사이트 위·변조에 대한 신속성 및 정확성을 향상하는데 목적을 둔다.

ABSTRACT

New types of attacks that mainly compromise the public, portal and financial websites for the purpose of economic profit or national confusion are being emerged and evolved. In addition, in case of 'drive by download' attack, if a host just visits the compromised websites, then the host is infected by a malware. Website falsification detection system is one of the most powerful solutions to cope with such cyber threats that try to attack the websites. Many domestic CERTs including NCSC (National Cyber Security Center) that carry out security monitoring and response service deploy it into the target organizations. However, the existing techniques for the website falsification detection system have practical problems in that their time complexity is high and the detection accuracy is not high. In this paper, we propose website falsification detection system based on image and code analysis for improving the performance of the security monitoring and response service in CERTs. The proposed system focuses on improvement of the accuracy as well as the rapidity in detecting falsification of the target websites.

Keywords: Security Monitoring and Control, Website Falsification Detection System, Image and Code Analysis

접수일(2014년 8월 19일), 수정일(2014년 9월 19일), 게재확정일(2014년 9월 29일)

* 본 연구는 2014년도 미래창조과학부의 수탁사업 「과학기술사이버안전센터 구축 및 운영사업」의 지원을 받아 수행된 연구임 (G-14-GM-IR02)

† 주저자, kisados@kisti.re.kr
‡ 교신저자, song@kisti.re.kr(Corresponding author)

I. 서 론

최근 급변하는 사회와 더불어 인터넷의 성능이 획기적으로 향상되면서 네트워크를 통한 정보자원의 공유나 협력 연구가 활성화되고 있다. 그러나 인터넷 발전과 더불어 최근 경제적 이윤을 목적으로 한 해킹조직들이 국가 주요 웹사이트 및 사용자가 빈번하게 이용하는 웹사이트를 해킹하여 국가적 혼란을 야기시키거나 해킹한 웹사이트에 악성코드를 설치함으로서 웹사이트를 접속하는 행위만으로도 악성코드에 감염되는 이른바 'Drive by Download' 형태의 공격이 발생하고 있는 실정이다.

이러한 사이버 위협에 대응하기 위해 국내에서는 국가사이버안전센터(NCSC)를 중심으로 분야별 사이버 보안을 담당하는 부문 보안관제센터를 구축·운영하고 있으며 전주기적 정보보호 활동을 위한 관제, 분석, 대응지원의 체계를 갖추고 있다. 과학기술사이버안전센터(S&T-SEC) [6]은 부문보안관제 센터 중 하나로 과학기술 공공·연구 기관에 대한 보안관제 서비스를 제공하고 있다.

그러나 보안관제센터의 대상기관 수가 증가함에 따라 실시간으로 웹사이트 위·변조를 탐지하는데 한계가 있으며, 연구학계 및 기업 등에서 발표한 웹사이트 위·변조 탐지시스템을 보안관제 서비스에 적용하기에는 다음과 같은 문제점을 가지고 있다.

첫째, 기존 웹사이트 위·변조 시스템[1][2][5]은 웹 문서의 태그, 속성 및 콘텐츠 중에서 하나라도 일치하지 않을 경우 이를 위·변조로 탐지하기 때문에 시스템 담당자가 시각적 효과를 주기위해 글자 색깔, 폰트크기 등을 변경하였을 때에도 정상적인 행위임에도 불구하고 위와 동일하게 위·변조로 탐지하는 사례가 빈번하게 발생하고 있다.

둘째, 기존의 연구는 주로 시스템 담당자가 웹 서버의 전체 디렉토리를 검색하여 정상 파일과 현재의 파일 리스트, 크기 등을 토대로 변경유무를 검사하는 로컬(내부) 방식을 사용하여 왔다. 해당 방법은 웹 관련 파일 뿐만 아니라 웹 서비스에 필요한 설정파일 등의 작은 단위(Fine-grained)까지 검사대상으로 하기 때문에 탐지시간이 증가하게 된다. 보안관제는 특정한 곳의 기관이 아닌 수십~수백 개의 대상기관에게 신속·정확한 관제 서비스 제공을 목적으로 하기에 해당 방식을 적용하기 어렵다.

셋째, 현재 국내·외 시장에 출시된 제품들은 웹사이트 위·변조 시스템에 특화되기 보다는 각종 정보보안 시스템(침입탐지, 침입차단, 침입방지)을 통합한 종합관리시스템으로 개발되기 때문에 시스템 규모가 크며 상당히 높은 도입비용이 발생한다. 또한 해당 시스템 내의 많은 수의 모듈을 연동 및 동기화 하는데 시간소요가 많다.

넷째, 기존의 연구는 텍스트 기반으로 웹사이트 위·변조의 탐지과정 및 결과를 보여주기 때문에 보안관제 요원이 직관적으로 탐지하는데 한계가 있다. 대부분의 보안관제 요원은 해킹 사고를 미연에 방지 및 대응하기 위해 침해위협관리시스템(TMS), 통합위협관리시스템(UTM) 등 다양한 보안시스템 운영·관리하기에 관제를 위한 해당 시스템의 시각화가 절실히 요구된다.

따라서 본 연구는 웹사이트 위·변조 여부를 신속·정확하게 판별하기 위해 이미지 및 코드 분석기반의 웹사이트 위·변조 시스템을 제안한다. 제안 시스템은 웹크롤러에 의해 비교검증의 대상이 되는 정보만을 수집하고 정규화를 통해 위·변조 판별에 영향을 미치는 이미지 및 코드를 추출하여 유사도를 분석하고 이를 시각화함으로서 보안관제요원의 직관적인 탐지 및 웹사이트 위·변조에 대한 신속성 및 정확성 향상을 목적으로 한다.

본 논문의 구성은 다음과 같다. 2장에서는 웹사이트 위·변조 탐지를 위해 현재 진행되고 있는 기법들을 소개하고 3장은 제안 시스템 원리와 구조에 대해 기술한다. 4장에서는 웹사이트 위·변조 시스템의 설계·구현을 제안하며 5장은 제안 시스템의 산출물에 대한 우수성을 제시하고 6장에서는 본 논문의 최종 결론을 맺는다.

II. 관련 연구

2.1 경로탐색기반 웹사이트 비교에 관한 연구

특정 웹 페이지 상에서 HTML 태그 및 속성 등의 변경유무를 탐지하기위한 경로(path)탐색 기반 [1][4]의 방법들이 제안되었다. 이들 기법은 우선 검사 대상을 선정한 다음 각 해당 정보자원의 위치에 대한 URL 목록을 작성하여 HTML 파일을 생성하게 된다. 생성된 파일은 경로 관리자(Path Manger)를 통해 URL 목록에 포함된 모든 웹 페이지의 변경여부를 검사하게 된다. Fig.1.은 선별된 웹페이지에 대한 변경 결과를 보여준다. 탐지방식은 정상 및 현재 웹페이지의 비교를 통해 진행되며 4가

지 시그니처(단락 검사, HEAD 검사, 링크 검사 및 키워드 검사)를 바탕으로 유사도를 판별한다. 해당 기법은 웹페이지 변화 탐지율이 높은 편이나 필터링 및 축약과정 없이 전체 HTML의 웹페이지를 대상으로 정상 웹페이지와 비교하기 때문에 변경된 웹페이지를 탐지하는 시간이 매우 오래 걸리는 단점을 지닌다.

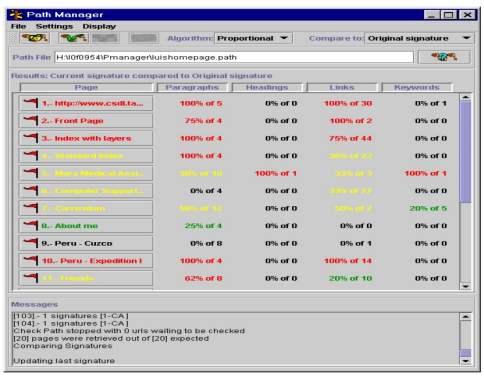

Fig. 1. View of Change Metrics

2.2 노드기반 웹사이트 비교에 관한 연구

웹 페이지 정보의 추가, 삭제, 및 업데이트 될 시이를 탐지하기 위한 노드 기반의 비교 알고리즘[2][3][8]들이 제안되었다. 해당 기법들은 Fig.2.와 같이 DOM(Document Object Model)을 이용하여 HTML 문서의 논리적 구조를 트리 형태로 추출하고 이를 다시 XML 형태로 변환한다. 변환된 XML 노드 트리를 토대로 정상 웹 페이지와 현재 웹 페이지에 대한 변경여부를 판별하게 된다. Fig.2는 웹페이지 내용 변경을 탐지하기 위한 아키텍처를 보여준다. 먼저, Input Unit을 통해 탐지대상을 지정하고 Crawler[7][12]는 모니터링 주기 시간을 기준으로 정상 및 현재 웹페이지를 수집한다. 수집된 파일은 매니저를 통해 XML 형식으로 변환되며 비교 분석기를 통해 해당 웹 페이지에 대한 대조확인을 수행한다. 해당기법은 웹사이트 구조변경을 탐지하기 위해 XML변환 및 관련 모듈 설계를 고려하였으나 HTML로 작성된 웹문서를 XML 문서로 변환하는 시간이 소요되며 경로탐색기반과 마찬가지로 불필요한 정보 수집 및 무의미한 비교측정 방지를 위한 필터링 및 축약과정이 없기 때문에 탐지시간이 증가하는 단점을 가진다.

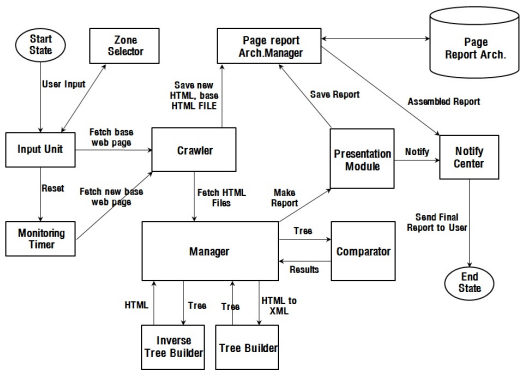

Fig. 2. The Architecture for Content Changes

2.3 사용자 관심기반 웹사이트 비교에 관한 연구

사용자가 자주 이용하는 웹 페이지의 정보변경을 탐지하기 위해 사용자 관심 기반[5][9][10][11]의 방법들이 제안되었다. 이들 기법은 사용자가 관심을 갖는 특정 웹페이지 또는 내용을 지정하여 일정 시간 주기로 해당 변화를 탐지하는 것을 특징으로 하며 HTML뿐만 아니라 XML 문서까지 탐지가 가능하다는 장점을 가지고 있다. Fig.3.은 해당 방법의 대한 아키텍처를 나타나며 7가지 모듈로 구성된다. 우선, 감시모듈을 통해 사용자가 요청한 특정 웹페이지를 수집하여 검증모듈로 보내면 수집된 문맥 구조 및 의미를 검증하고 지식기반 모듈을 통해 관계형 DB로 저장된다. 변화탐지 모듈은 저장된 DB를 토대로 해당 데이터의 특징을 추출하여 시간주기로 정상페이지와 현재 페이지를 비교하여 변화여부를 판별한다. 사용자 관심기반 기법은 탐지 정확도는 높은 편이나 데이터 마이닝, 기계학습 등 응용 알고리즘의 적용으

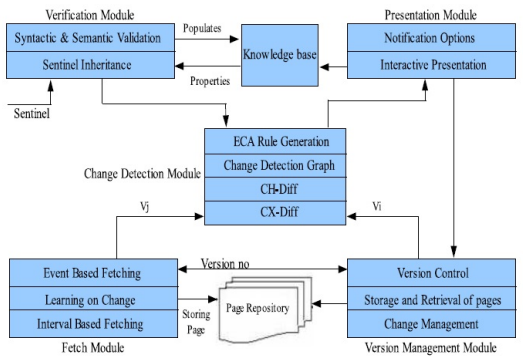

Fig. 3. Web Change Detection Architecture based on User

로 인해 시간복잡도가 높다는 단점을 가진다.

III. 보안관제를 위한 위·변조 탐지시스템

3.1 웹사이트 위·변조 시스템 기본원리

웹사이트 위·변조 시스템의 기본원리는 Fig.4.와 같이 크게 수집·탐지시스템 및 모니터링 뷰로 구성된다.

① 수집·탐지 시스템은 인터넷에 연결된 국가 주요 웹사이트 및 포털사이트를 대상으로 위·변조 여부를 수집 및 탐지한다. 데이터의 형태는 이미지 및 코드이며 시간주기 설정에 따라 해당 정보가 저장된다.

② 저장된 정보는 모니터링 뷰에 의해 보안관제요원이 직관적으로 이해하기 쉽도록 정상화면과 현재화면을 비교화면을 표출한다.

③~④ 관제요원은 모니터링 뷰를 통해 웹사이트 위·변조를 신속·정확하게 판별할 수 있으며 만일 특정 대상기관의 웹사이트가 비정상적인 경우, 관련기관에 대한 즉각적인 대응조치를 수행하게 된다.

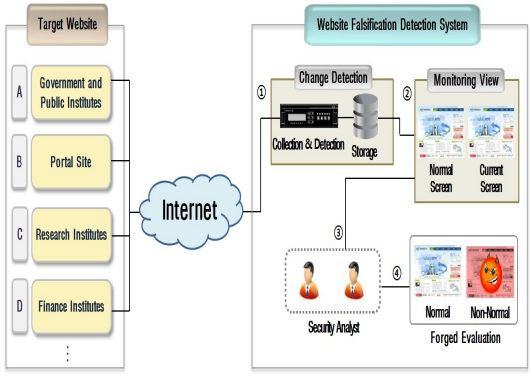

Fig. 4. Basic Framework of Website Falsification Detection System

3.2 웹사이트 위·변조 시스템 구조

본 절에서는 제안시스템의 구조 및 각 모듈별 기능에 대해 자세히 다루며 Fig.5.는 제안된 5가지(웹크롤러, 정상 웹사이트 빌딩, 현재 웹사이트 조사, 유사도 비교 및 결과 관리)모듈 및 아키텍처를 보여준다.

3.2.1 웹크롤러

웹크롤러는 지능형 소프트웨어 에이전트이며 수집하고자 할 대상기관의 URL 리스트를 기반으로 해당 기관의 웹사이트를 방문 및 관련 데이터를 수집한다. 웹크롤러는 수집 후, 일정주기에 따라 URL 리스트를 갱신하여 재귀적으로 다시 방문하는 과정을 거친다. 웹크롤러에 의해 수집된 데이터는 정상 웹사이트 빌딩 모듈 및 현재 웹사이트 조사모듈로 전달한다.

3.2.2 정상 웹사이트 빌딩모듈

정상 웹사이트 빌딩모듈은 각 대상기관의 웹사이트가 악성코드 감염 및 위·변조가 발생되지 않은 상태에서 데이터를 추출하는 모듈이다. 해당 모듈은 크게 3가지(수집, 선별 및 추출) 단계로 구분된다. 우선, 수집단계는 웹크롤러를 통해 이미지 및 소스코드를 수집하게 된다. 해당 정보수집 시 일시적인 접속 장애 및 데이터 전송오류 등으로 인해 수집을 완료하지 못할 경우 2~3회에 걸쳐 해당 데이터를 수집한다.

선별단계는 수집된 이미지 및 소스코드가 실제로 해당 대상기관의 데이터인지 여부를 확인하기 위해 기관명, URL 및 특정 키워드를 조사하게 된다. 특정 키워드는 각 대상기관의 식별이 가능한 고유특징을 뜻한다. 만약, 위의 3가지 사항에 부합되지 않을 경우 해당 데이터는 삭제되며 반대로 부합된 데이터는 파일개수 및 파일용량 등을 확인하는 과정을 통해 복수 개로 수집된 대상기관의 데이터 중 최종 원본 데이터를 선정하게 된다.

추출단계는 최종원본 데이터를 토대로 중복 및 무의미한 소스코드를 제거하는 과정이다. 제안시스템은 HTML로 구성된 웹페이지에 대한 위·변조를 보다 신속하게 탐지하기 위해 위·변조 판별에 영향을 미치는 특정코드 및 이미지만을 추출한다. 추출한 특정코드는 HTML의 구성요소 중 Table 1.과 같이 보안에 취약한 태그, CSS[1](Cascading Style Sheet) 및 자바스크립트이다. 위의 3가지 요소는 해커로부터 스크립트 구문을 삽입하여 사용자의 개인정보를 탈취하는 XSS(Cross Site Script), CSRF(Cross Site Request Forgery) 및 신·변종 웹페이지 위·변조 공격을 시도할 때 빈번하게 악용되는 항목이다.

추출한 소스코드는 유사도 비교모듈에서 정상 소스코드와 유사도를 측정하게 되며 정확도 향상 및 오탐을 줄이기 위해 픽셀 기반의 이미지 유사도 측정과 함

1) CSS(Cascading Style Sheets) : 웹 문서의 전반적인 스타일을 미리 저장해 둔 스타일시트로서 문서 전체의 디자인 변경 및 일관성 유지가 가능함

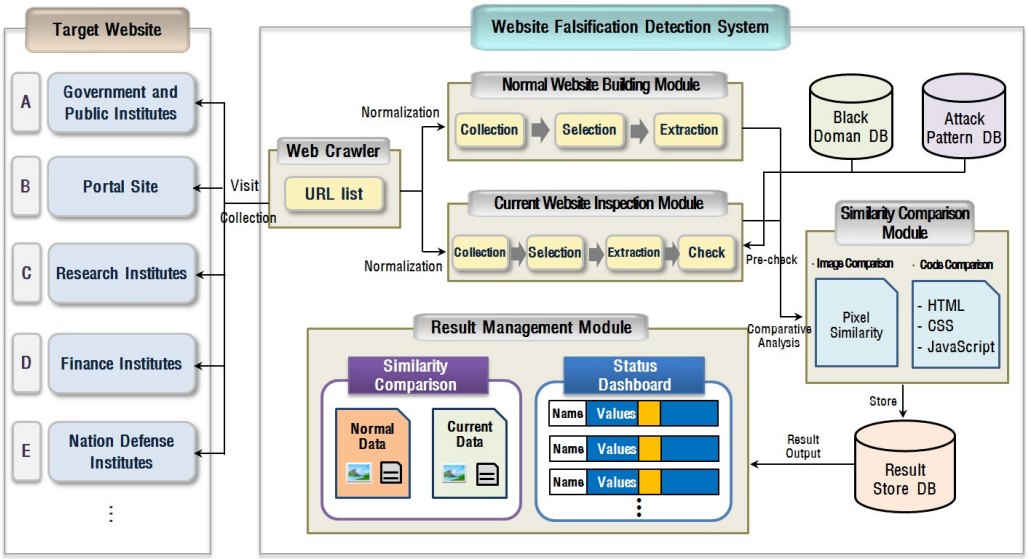

Fig. 5. Architecture of Proposed System

께 수행된다.

Table 1. Item Extraction having Security Weakness

Item	Tag (function)	Contents
HTML	〈script〉	Abuse by XSS and CSRF attacks as the tag enabling script execution in web
	〈iframe〉	Abuse by malicious URL insertion as partition tag in browser
	〈a href〉	Abuse by malicious URL as the tag creating hyperlink
	〈img〉	Abuse by forged images as the tag for the image insertion
	〈li〉	Abuse by forged contents as the tag displaying the catalogue in web
CSS	〈link href〉	Abuse by malicious URL insertion as the tag referring CSS for HTML and XML design
	〈style〉	Abuse by malicious script and URL as the tag inserting design item in HTML
JAVA Script	〈script src〉	Abuse by forged js files as the tag for inserting Javascript codes
	document.getElementById	Abuse by malicious script as the function for calling specific values in XML documents
	function	Abuse by malicious function call as the function using the declaration and call

3.2.3 현재 웹사이트 조사모듈

현재 웹사이트 조사모듈은 웹크롤러를 통해 일정시간 주기로 각 대상기관의 웹사이트 관련 데이터를 추출하는 모듈이다. 해당 모듈은 수집, 선별, 추출 및 검사의 4단계로 구성되며 추출단계까지 정상 웹사이트 빌딩모듈과 동일한 과정으로 진행된다. 검사 단계는 정상 및 현재 웹사이트의 유사도를 비교하기 전에 추출단계에서 축약된 요소를 토대로 블랙 도메인 및 특수문자를 사용한 스크립트를 우선적으로 검사하여 위·변조 여부를 발견하는 과정이다. 추출 요소 중 블랙도메인 리스트 및 사고대응을 수행하여 도출된 공격 유형패턴과 비교하여 하나라도 일치하는 경우 해당 웹사이트는 위·변조가 되었다고 판단하여 유사도 비교 없이 빠른 대응지원이 이루어지게 된다.

Table 2.은 주요 블랙도메인에 대한 예시를 보여주며 보안을 위해 도메인 중간을 *로 표기하였다.

Table 2. List of Main Black Domain

Black Domain	Contents
http://copy.yo**gkoala.com	Transfer of system information by infection
http://ddi.astr**ger.com	
http://inds.m**o.com	
http://korea1.m**o.com	
http://wy001.it**ns.com	
http://scholes.o*s.org	Information transfer to compromised system
http://www.15**t.com	
http://zxzx.k**bf.com	

3.2.4 유사도 비교모듈

유사도 비교 모듈은 정상 및 현재 웹사이트 조사 모듈을 통해 도출된 각 요소간의 유사도를 측정하는 모듈이다. 해당 모듈은 이미지 비교, 소스코드 비교 및 통합 비교로 구분된다. 이미지 비교는 정상 및 현재 웹사이트의 이미지를 비교하여 유사도를 측정한다. 이미지 유사도 측정방식은 이미지 1개의 픽셀에 대해 RGB(Red, Green, Blue)값이 모두 동일할 때 같은 픽셀로 판정하는 단일픽셀 방식과 이웃한 픽셀들과 유사도를 비교하는 인접픽셀 방식이 있다. 인접픽셀 방식은 단일픽셀 방식보다 정확도가 높은 반면에 많은 계산량을 필요로 하기 때문에 신속성을 요구하는 환경에서는 적용이 어렵다는 문제점을 지닌다.

따라서 본 논문에서는 단일픽셀 방식을 적용한 유사도 측정을 제안하며 정확도 향상을 위해 유사도를 비교하기 전에 정상 및 현재 이미지의 크기를 측정한 후 이를 토대로 이미지의 비교 범위를 설정한다. 이미지 유사도 비교의 식 (1)은 다음과 같다.

$$\text{IS(Image Similarity)} = \frac{\sum \text{현재동일픽셀수}}{\sum \text{정상픽셀수}} \times 100 \quad (1)$$

소스코드 비교는 이미지 비교와 마찬가지로 정상 및 현재 웹사이트의 소스코드를 비교하여 유사도를 측정한다. 추출코드 유사도(Source Similarity)는 식 (2)와 같이 추출된 소스코드 비율과 정상 및 현재 소스코드 파일 비율의 산술적인 평균으로 계산한다.

$$\text{SS(Source Similarity)} = \frac{NEC + ECS}{2} \quad (2)$$

NEC(Number of Extracted Code)는 추출된 소스코드 수를 나타내며 식 (3)과 같이 정상 및 현재 웹사이트의 추출코드 수와 비교한 결과를 도출한다.

$$NEC = \frac{\sum \text{현재동일 추출코드수}}{\sum \text{정상추출코드수}} \times 100 \quad (3)$$

식 (4)는 추출된 정상 및 소스코드 크기 비율(Extracted Code Size)을 나타내며 해당 식을 대입하기 전에 정상 및 현재 코드크기를 백분율로 환산하여 대입한다. 만약 해당 추출코드의 크기가 서로 동일할 경우 결과 값은 1에 가까우며 반대로 코드 크기가 서로 상이할 경우 0에 가까운 값을 도출하기 때문에 NEC와 ECS간 합의 평균을 적용함으로서 보다 세밀한 유사도 측정이 가능하게 하였다.

$$ECS = \frac{1}{a}x \ (x \le a), \ \frac{a}{x} \ (x > a) \quad (4)$$

· a : 정상 소스코드 크기
· x : 현재 소스코드 크기

통합비교는 이미지 및 소스코드 비교에 대한 결과를 바탕으로 위·변조 여부를 판별한다. Table 3.은 위·변조의 판별기준을 나타내며 이미지 및 소스코드의 유사도가 임계치 미만이면 1로 표기하고 그 반대이면 0으로 표기한다. 임계치는 각 대상기관에 대해 사전에 경험적으로 설정한 평균값이다. 웹사이트는 사이버 해킹과 상관없이 대상기관의 관리자에 의해 하루에도 빈번하게 업데이트가 발생하기 때문에 대상기관의 상황에 적합한 임계치를 설정하였다.

우리는 설정한 임계치를 기반으로 만일 이미지 및 소스코드의 합계가 2인 경우 해당 웹사이트는 위·변조가 되었다고 판단하며 합계가 1인 경우에는 위·변조의 가능성 있다고 예측한다. 반대로 합계가 0인 경우 위·변조가 발생하지 않는다고 판별한다.

Table 3. Decision of Web Change Detection

	Image	Code	Type
Similarity Results	1	1	Danger
	1	0	Care
	0	1	
	0	0	Normal

3.2.5 결과 관리모듈

결과 관리모듈은 통합비교의 결과를 토대로 시각화를 제공하는 모듈이다. 본 모듈은 통합비교에서 도출된 현재 및 정상 웹사이트의 결과 값을 관제요원이 이해하기 쉽게 유사도 비교 뷰와 관제 현황판으로 표출되며 또한, 표출 시간설정으로 자유롭게 모니터링이 가능하도록 하였다.

IV. 웹사이트 위·변조 시스템 설계 및 구현

4.1 웹사이트 위·변조 시스템 설계

제안 시스템은 Fig.6.과 같이 6가지(관제대상, 수집, 생성, 검사, 비교 및 관리) 프로시저로 설계 하였으며 절차는 다음과 같다.

① 수집단계는 관제대상기관의 웹사이트 정보를 검색 및 수집하기 위해 웹크롤러를 이용하여 관련 전체 이미지 및 소스코드를 받아온다.

②~③ 생성단계는 정상 및 현재 웹사이트에 대한 데이터를 추출하고 생성하는 과정으로 위·변조의 위험도가 높은 이미지 및 소스코드를 선별하였으며 현재 데이터 생성의 경우, 데이터 생성 전에 검사과정을 두어 치명적인 공격에 대해 사전검사를 수행할 수 있도록 하였다

④~⑥ 검사단계는 ③번에서 잠시 언급하였듯이 추출한 현재 데이터를 토대로 블랙도메인 및 공격패턴을 검사하여 위·변조 여부를 신속히 판별하는 과정과 생성된 정상 및 현재 데이터를 검증DB에 저장하는 역할을 수행한다.

⑦~⑧ 비교단계는 저장된 검증DB로부터 정상 및 현재 웹사이트의 유사도를 비교하고 해당결과를 토대로 위·변조 판별을 수행한다.

⑨ 관리단계는 유사도 비교결과를 출력하는 과정으로 도출된 텍스트기반의 결과를 시각화하여 보안관제요원이 웹사이트 위·변조에 대해 직관적으로 모니터링이 가능하게 설계하였다.

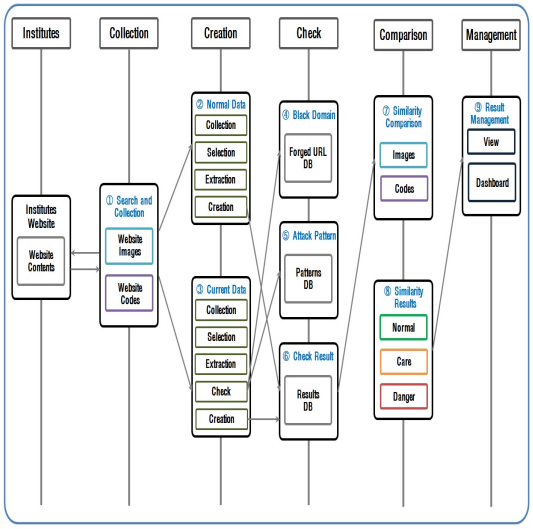

Fig. 6. Procedures of Proposed System

4.2 웹사이트 위·변조 시스템 주요함수

제안 시스템은 정상 웹사이트 데이터를 생성하는 「정상데이터 빌딩클래스」, 현재 웹사이트 데이터를 생성하는 「현재데이터 조사클래스」, 생성된 데이터의

Table 4. Class for Normal Data Building

Class for Normal Data Building

/* Class execution creating normal data */
void main( )
{
//Path check storing normal data
Folder_Check(path);
//Receiving Institute name and URL information
Get_Inst_Value(inst_value,path);
//Receiving Forged URL information
Get_Info.Get_TXT_URL(URL, path);
//Images and codes collection through institute website's connection
Get_Source_Image(inst_value,tmp_source,path,cnt);
//Origin check of source codes
Source_Check(tmp_source,inst_value,cnt);
//The most compatible code selection among the collected codes
Select_Source(tmp_source,source,selected_num,cnt);
//The rest deletion of collected data except selected data
Arrange_Source(source,selected_num,path,cnt);
//Specific tag extraction among the selected data
Extract_Source(cnt,path);
}

유사도를 비교하는 「유사도 비교클래스」 및 유사도 결과를 출력해주는 「결과 관리클래스」로 구성된다. Table 4.는 정상데이터 생성 클래스의 주요함수를 보여준다. 해당 모듈은 대상기관의 정보(기관명, URL)를 텍스트로 정보를 받아 각 기관의 웹사이트에 접속하여 대상기관 웹사이트의 관련 정보(이미지, 소스코드)를 받아온다. 수집한 웹사이트 정보는 선별 및 추출 과정을 거쳐 정상 데이터를 생성한다.

Table 5.는 현재 웹사이트 데이터 조사모듈의 주요함수를 보여준다. 해당모듈은 정상데이터 빌딩모듈과 유사하나 데이터를 생성하기 전에 블랙도메인 및 공격패턴 검사를 통해 위·변조 여부를 판별한다.

Table 5. Class for Current Data Inspection

Class for Current Data Inspection

/* Class execution creating current data */
void main( )
{
//Receiving current institute information
Get_Current_Info(inst,path,cnt);
 	
//Checking forged URL insertion of source codes
URL_Check(URL,inst[inst_num][1],path,inst_num,total_result);
 	
//Checking attack patterns of source codes
Character_Check(inst_num,path);
 	
//Current data extraction
Extract_Source(path,cnt);
}

Table 6.은 생성된 정상 및 현재 웹사이트 데이터를 비교하는 유사도 비교 클래스의 주요함수를 보여준다. 대상기관의 생성 이미지와 소스코드 유사도를 계산하여 반환 값을 통해 위·변조 여부를 판단한다. 이미지 유사도가 임계치 미만일 경우 total_result값 1을 증가하고 마찬가지로 소스코드 유사도가 임계치 미만일 경우에도 total_result값 1을 증가시킨다. 만약 total_result 값이 2가 되는 경우 이미지 및 소스코드가 모두 정상 데이터와 유사하지 않기 때문에 해당 웹사이트는 위·변조 되었다고 판별한다.

Table 6. Class for Similarity Comparison

Class for Similarity Comparison

/* Class execution comparing the similarity of normal and current data */
void main( )
{
for(int inst_num=0;inst_num<cnt;inst_num++)
{
//Image comparison
result[inst_num][0] = Image_Compare(inst[inst_num][1],path,inst_num);
 	
//Code comparison (weight * code similarity)
result[inst_num][1] = Text_Weight(path,inst_num) *
 	
Text_Compare(inst[inst_num][1],path,inst_num);
 	
//when image similarity is fewer than threshold
if(result[inst_num][0] < image_th)
total_result++;
 	
//when codes similarity is fewer than threshold
if(result[inst_num][1] < text_th)
total_result++;
 	
if(total_result > 1){
//Images and codes store
Save_Log(path,date,inst_num);
}
}

Table 7.은 유사도 결과 관리클래스의 주요함수를 나타낸다. 해당 클래스를 통해 유사도 결과값을 토대로 시각화 하였으며 정상 및 현재 웹사이트를 한 화면에 출력하여 직관적으로 웹사이트 상태를 확인할 수 있게 하였다.

Table 7. Class for Similarity Result Management

Class for Result Management

/* Class execution visualizing similarity results */
void main( )
{
for(int inst_num=0;inst_num<cnt;inst_num++){
 	
//Creatiojn of Image frame

```
Frame f = new ImageFrame(url,inst_name,img_result,txt_
result,path,inst_num);
//Starting point of print images
f.setLocation(start_x, start_y);

//Size adjust of print images
f.setSize(height ,width);

//Title set of print images

f.setTitle("Analysis results of images");
f.show();

//Time set of print images
sleep((int)sleep_time);

//Memory return of print images
f.dispose();
((ImageFrame) f).flush();

}
}
```

4.3 웹사이트 위·변조 시스템 구현

웹사이트 위·변조 시스템의 화면은 유사도 비교결과 화면과 대상기관 현황판으로 구분된다. 먼저 비교결과 화면은 Fig.7.과 같이 화면 상단중앙에는 대상기관 이름을 보여주며 바로 아랫줄에는 이미지 및 소스코드 유사도와 함께 해당결과를 정상, 주의 및 위험의 3단계로 종합하여 나타낸다. 참고로 우리는 보안을 위해 대상기관의 이름을 부분적으로 블라인드 처리하였다.

또한, 비교결과 이외에도 화면 중앙에 대상기관 웹사이트의 정상화면과 현재화면의 이미지를 출력함으로써 보안관제요원이 웹사이트 상황을 직관적으로 파악할 수 있도록 구현하였다. Fig.8.은 정상화면과 현재화면의 유사도가 일치하지 않은 경우를 나타낸 것이며 확인결과 관리자에 의한 웹사이트 점검으로 확인 되었다.

웹사이트 위·변조 현황판은 Fig.9.와 같이 각 기관에 대한 최근 12시간의 유사도 비교결과 값을 보여준다. 해당 결과 값은 위에서 언급하였듯이 정상, 주의 및 위험의 3단계로 구분하고 각 단계별 색상(녹색, 노란색, 빨강)을 달리 지정하여 표기하였다. 또한 위험단계가 1회 이상 발생한 경우 대상기관 이름의 바탕색도 동일 색상으로 표현하여 가독성을 높였다. 무엇보다 위·변조 현황판은 전체 대상기관의

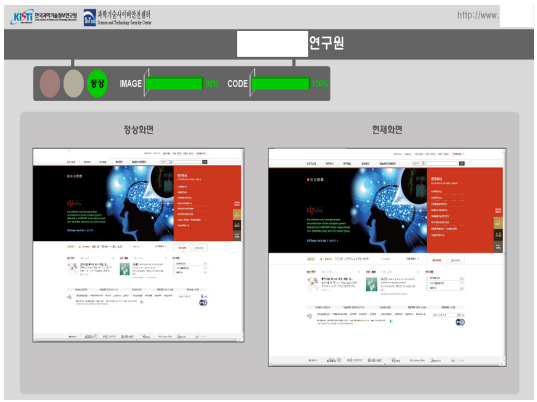

Fig. 7. Result Screen of Similarity Comparison-1

Fig. 8. Result Screen of Similarity Comparison-2

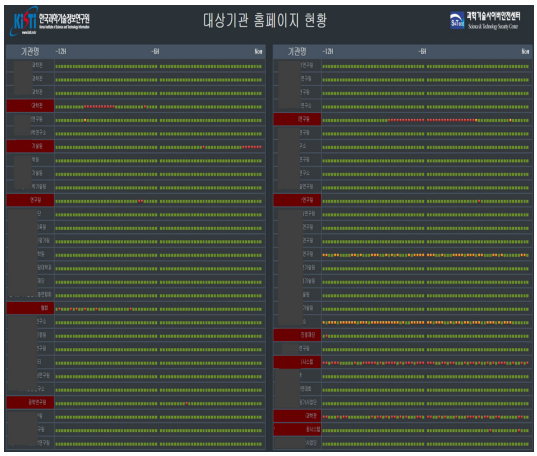

Fig. 9. Dashboard of Related Institutes

이력을 직관적으로 파악할 수 있기 때문에 웹사이트 현황 및 상태변화에 대한 모니터링 및 분석이 용이할 뿐만 아니라 불안전한 웹사이트에 대해 집중적인 보안관제가 가능하다는 장점을 가진다.

V. 웹사이트 위·변조 시스템 결과분석

5.1 웹사이트 위·변조 시스템 정확성 및 탐지시간

본 연구는 정보보호기관 웹사이트를 대상으로 웹사이트 위·변조에 대한 정확도를 측정하였다. 측정방법은 14년 7월 1일부터 14년 7월 15일까지 연구·공공기관의 58개 웹사이트에 대한 유사도를 10분주기(총 125,280회)로 비교하였다. Fig.10.은 도출된 결과(3단계: 정상, 주의 및 위험)의 정확도를 나타내며 정상 및 위험단계인 경우 100%의 정확도를 얻었으며 주의단계는 약 92%의 정확도를 보였다.

먼저, 위험단계 분석결과에 관해 국내·외 해커그룹에 의한 웹사이트 위·변조 공격 보다는 네트워크 장애 및 웹사이트 점검으로 인한 접속불가 및 기관 업데이트로 인한 웹사이트 변경이 거의 대부분을 차지하였다. 또한, 주의단계에서는 이미지 및 소스코드 중 한 개 특성의 유사도가 임계치 미만인 경우 탐지되며 아직 위·변조의 행위가 발생하지 않는다고 판단하는 경우이다. 주의단계 대부분이 소스코드의 유사도는 거의 일치하지만 플래쉬 적용이 되어 있는 웹사이트의 이미지가 바뀔 때 탐지되었다.

그러나 281건(총3510건)은 위험단계로 탐지되었어야 했으나 주의 단계로 탐지된 사례이다. 이는 네트워크 및 시스템 환경 등으로 인해 대상기관의 수집되는 데이터양이 적어 임계치 계산이 어려운 경우 나타나며 이 현상은 정확성 향상을 위해 앞으로 개선해

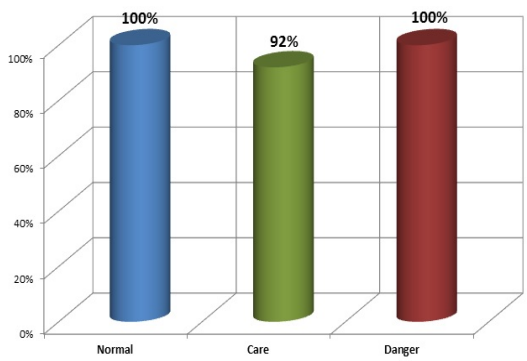

Fig. 10. Accuracy of Proposed System

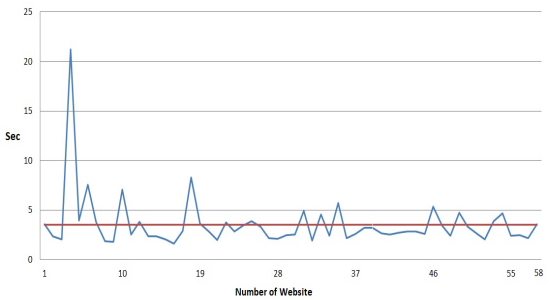

Fig. 11. Detection Time of Proposed System

야 할 점이라 하겠다.

Fig.11.은 웹사이트 위·변조에 대한 탐지 시간을 그래프로 나타낸 것이다. 탐지시간은 데이터 수집에서부터, 선별, 추출 및 유사도 측정까지 소요되는 시간을 의미한다. 전체 58개 웹사이트에 대한 평균 탐지 소요시간은 3.52초로 신속한 위·변조 탐지가 가능하며 웹사이트 접속이 원활하지 않은 특정 기관을 제외하고 대부분의 웹사이트는 5초 안에 위·변조 탐지가 완료되는 것을 확인할 수 있다.

5.2 웹사이트 위·변조 시스템 비교 분석

Table 8.은 기존 연구들과 제안시스템을 비교분석한 결과를 보여준다. 먼저, 제안시스템은 보안 관제를 위한 특화된 시스템으로 외부에서 웹사이트 위·변조를 판별하여 보안관제요원이 정보보호 기관을 24시간 모니터링을 실시함으로서 운용·관리 측면에서 효율성이 높다.

둘째로, 제안시스템은 외부에서 수집 가능한 HTML 태그, 속성, CSS, 및 자바스크립트 등 다양한 유형의 위·변조를 탐지할 수 있으며 특히, 정상 웹사이트와 유사도를 비교하기 전에 추출된 데이터를 토대로 블랙도메인 및 해킹코드를 사전에 검사하므로 신속한 탐지 및 대응이 가능하다.

셋째로, 기존 연구들은 각 개별 특성(XML 구조 변환, 응용 알고리즘 적용 및 모듈 동기화) 등으로 인해 시간 복잡도가 높은 반면에 제안 시스템은 데이터를 정규화(수집, 선별, 추출, 검사 및 생성) 하여 유사도를 비교하는데 걸리는 시간이 평균 3.5초 밖에 소요되지 않기 때문에 상대적으로 시간 복잡도가 낮다.

Table 8. Comparison with the Existing Approaches

	Path-Detection	Node-Detection	User-interested	Proposed System
Object	System administrator	User/System administrator	User	Security Monitoring and Control
Detection	Internal	External	Internal/External	External
Type	HTML	HTML	HTML/XML	HTML (CSS, Javascript)
URL List	○	○	○	○
Hacking Code Detection	×	×	×	○
Pre-Check	×	×	×	○
Filtering/ Contraction	×	×	×	○
Time Complexity	High	High	High	Low

VI. 결 론

본 연구는 웹사이트 위·변조 여부를 신속·정확하게 판별하기 위해 이미지 및 코드 분석기반의 웹사이트 위·변조 시스템을 제안하였다. 제안시스템은 웹크롤러에 의해 분석에 필요한 정보만을 수집하여 정규화 및 유사도 비교를 통해 웹사이트 위·변조를 탐지하는 특화된 단일 시스템으로 구축 및 적용이 쉽고 별도의 보안장비와의 연동 및 구입을 요구하지 않기 때문에 도입비용이 상대적으로 낮은 강점을 가진다.

또한 실제 해킹시도와 관련한 웹사이트 위·변조를 시각화하여 신속·정확하게 분석함으로서 보안관제 업무의 효율성 향상에 직접적인 기여가 가능할 뿐만 아니라 해킹공격에 대한 예방 업무수행도 가능하였다. 개발된 웹사이트 위·변조 시스템은 현재 과학기술사이버안전센터(S&T-SEC)를 통해 구축·운용 중이며 해당 기술의 안정화·고도화를 위해 실시간 보안관제 및 침해대응 서비스에 적극 활용할 예정이다.

References

[1] Francisco-Revilla, L. F. Shipman, Furuta R., Karadkar U., and Arora A., "Managing Change on the Web," Proc. of the 1st ACM/IEEE-CS joint conference on Digital libraries(JCDL '01), pp. 67-76, Jun. 2001.

[2] Khandagale H. P. and Halkarnikar, "A Novel Approach for Web Page Change Detection System," In International Journal of Computer Theory and Engineering, vol. 2, no. 3, 1793-8201, pp. 364-368, Jun. 2010.

[3] Varshey N. K. and Sharma D. K, "A Novel Architecture and Algorithm for Web Page Change Detection," Proc. of the Advance Computing Conference, pp. 782-787, Feb. 2013.

[4] Furuta, R., Shipman, F., Marshall, C., Brenner, D., and Hsieh, H. "Hypertext Paths and the World-Wide Web: Experiences with Walden's Paths," Proc. of the 8th ACM conference on Hypertext(HYPERTEXT '97), pp. 167-176, Apr. 1997.

[5] Chakravarthy S., Hara S.C.H., "Automating Change Detection and Notification of Web Pages," Proc. of the 17th International Workshop on Database and Expert Systems Applications(DEXA '06), pp. 465-469, Sep. 2006.

[6] Science&Technology Security Center, http://www.sntsec.or.kr/

[7] Olston C., and Najork M., "Web Crawling," Journal of Foundatgions and Trends in Information Retrieval, vol. 4, no. 3, pp. 175-246, Mar. 2010.
[8] Wang Y., DeWitt D. J., and Cai J., "X-Diff: an effective change detection algorithm for XML documents," Proc. of 19th International conference on Data Engineering, pp. 519-530, Mar. 2003.
[9] Tomczak J. M., and Zieba M., "On-line bayesian context change detection in web service systems," Proc. of the international workshop on Hot topics in cloud services(HotTopiCS '13), pp 3-10, Apr. 2013.
[10] Khoury I., EI-Mawas, R.M, EI-Rawas O and Mounayar E.F., "An Efficient Web Page Change Detection System Based on and Optimized Hungarian Algorithm," Journal of IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 5, pp. 599-613, Mar. 2007.
[11] Prusiewicz A., and Zieba M., "The Proposal of Service Oriented Data Mining System for Solving Real-Life Classification and Regression Problems," Journal of Advances in Information and Communication Technology, vol. 349, pp. 83-90, Feb. 2011.
[12] Mali S. and Meshram B. B., "Implementation of multiuser personal web crawler," Proc. of the 6th International conference on Software Engineering(CONSEG '12), pp. 1-12, Sep. 2012.

〈저 자 소 개〉

김 규 일(Kyu-il Kim) 정회원
2005년 2월: 성균관대학교 컴퓨터공학과 석사
2010년 2월: 성균관대학교 컴퓨터공학과 박사
2010년 6월~현재: 한국과학기술정보연구원 과학기술정보보호실 선임연구원
〈관심분야〉 보안관제, 침해사고대응, 악성코드 분석

최 상 수 (Sang-soo Choi) 정회원
2001년 2월: 한남대학교 컴퓨터공학과 졸업
2003년 2월: 한남대학교 컴퓨터공학과 석사
2006년 2월: 한남대학교 컴퓨터공학 박사
2006년 2월~현재: 한국과학기술정보연구원 과학기술정보보호실 선임연구원
〈관심분야〉 정보보호, 보안관제, 침해사고대응

박 학 수 (Hark-soo Park) 정회원
1989년 2월: 한남대학교 전자계산학과 졸업
1991년 2월: 한남대학교 컴퓨터공학과 석사
2003년 2월: 한남대학교 컴퓨터공학 박사
1991년 3월~현재: 한국과학기술정보연구원 과학기술정보보호실 책임연구원
〈관심분야〉 정보보호, 보안관제, 침해사고대응

고 상 준 (Sang-jun Ko) 학생회원
2013년 2월: 한국항공대학교 정보통신공학 졸업
2013년 3월~현재: 과학기술연합대학원대학교 그리드 및 슈퍼컴퓨팅 석사과정
〈관심분야〉 정보보호, 네트워크 보안, 악성코드 분석

송 중 석 (Jung-suk Song) 정회원
2003년 2월: 한국항공대학교 통신정보공학 졸업
2005년 2월: 한국항공대학교 정보공학 석사
2009년 3월: 교토대학교(일본) 지능정보학 박사
2009년 4월~2010년 9월: 일본정보통신연구원 정보통신 보안 연구소 전문연구원
20010년 10월~2011년 9월: 일본정보통신연구원 네트워크 보안 연구소 선임연구원
2011년 10월~현재: 한국과학기술정보연구원 과학기술정보보호실 선임연구원
2012년 9월~현재: 과학기술연합대학원대학교 그리드 및 슈퍼컴퓨팅 조교수
〈관심분야〉 보안관제, 침해사고대응, 악성코드 분석, 네트워크 보안