KIPS Tr. Comp. and Comm. Sys.
Vol.6, No.2 pp.87~94 pISSN: 2287-5891



https://doi.org/10.3745/KTCCS.2017.6.2.87

# A Research of Anomaly Detection Method in MS Office Document

Sung Hye Cho† · Sang Jin Lee††

## ABSTRACT

Microsoft Office is an office suite of applications developed by Microsoft. Recently users with malicious intent customize Office files as a container of the Malware because MS Office is most commonly used word processing program. To attack target system, many of malicious office files using a variety of skills and techniques like macro function, hiding shell code inside unused area, etc. And, people usually use two techniques to detect these kinds of malware. These are Signature-based detection and Sandbox. However, there is some limits to what it can afford because of the increasing complexity of malwares. Therefore, this paper propose methods to detect malicious MS office files in Computer forensics' way. We checked Macros and potential problem area with structural analysis of the MS Office file for this purpose.

Keywords : MS Office, Malware, Anomaly Detection, doc, ppt, xls, Compound File Binary Format, OLE, Forensic

# MS 오피스 문서 파일 내 비정상 요소 탐지 기법 연구

조 성 혜† · 이 상 진††

## 요 약

최근 각종 공문서와 증빙 서류를 비롯하여 대부분의 문서가 디지털 데이터의 형태로 사용되고 있다. 특히 MS 오피스는 전 세계적으로 공공기관, 기업, 학교, 가정 등 다양한 곳에서 가장 많이 사용하고 있는 문서 편집 소프트웨어로써 악의적인 목적을 가진 사용자들이 해당 문서 프로그램의 범용성을 이용하여 MS 오피스 문서 파일을 악성 행위를 위한 매개체로 사용하고 있으며, 최근에는 단순한 사용자뿐만 아니라 국내외 정부 기관과 주요기업을 비롯하여 기반시설에서도 MS 오피스 문서 파일 형태의 악성코드가 유입되고 있다. MS 오피스 문서에 악성 코드를 삽입하는 방법은 단순히 미할당 영역에 은닉하는 방법을 사용할 뿐만 아니라 매크로 기능을 이용하는 등 다양한 방법을 통해 점점 정교한 형태로 진화되고 있다. 이러한 악성 코드들을 탐지하기 위해서 시그니처를 이용하거나 샌드박스를 이용한 탐지방법이 존재하지만, 유동적이고 복잡해지는 악성 코드들을 탐지하기에는 한계가 있다. 따라서 본 논문에서는 디지털 포렌식 관점에서 MS 오피스 문서 분석에 필요한 주요 메타데이터와 파일 포맷 구조 분석을 통해 매크로 영역과 그 외 악성 코드가 삽입될 가능성이 존재하는 영역들을 확인함으로써 MS 오피스 문서 파일 내 비정상 요소를 탐지하는 기법을 제안한다.

키워드 : MS 오피스, Malware, 비정상 탐지, doc, ppt, xls, 복합 파일 이진 구조, OLE, 포렌식

## 1. 서 론

디지털 기기의 발달로 인해 다양한 문서 작업이 디지털 데이터의 형태로 저장 및 활용되고 있다. 또한 태블릿 PC, 노트북, 스마트폰 등과 같은 휴대가 편의한 제품이 보편화 됨에 따라 많은 문서 작업이 디지털 기기로 이루어지고 있으며, 학교의 강의 자료를 비롯하여 각종 공문서와 증빙서류도 디지털 데이터의 형태로 사용되고 있다.

다양한 문서가 디지털 데이터의 형태로 사용되고, 웹 브라우저와 이메일 등을 통해 활발히 공유됨에 따라 악의적인 목적을 가진 사용자들은 문서 파일을 통해 악성코드를 배포하기 시작하였다. 특히 MS(Microsoft) 오피스는 국내 및 국외에서 70% 이상 점유하고 있기 때문에 MS 오피스 문서에 삽입된 악성 코드의 위험성이 증가하고 있다[1, 2].

MS 오피스 문서파일은 국내외 정부 기관과 주요 기업에서 사용 빈도가 매우 높기 때문에 고위 간부를 대상으로 사회공학적 기법에 사용할 MS 오피스 문서파일 형태의 악성코드가 급증하고 있다. 이러한 악성 코드의 피해 유형은 단순한 정보 유출 외에 DDoS 공격 등을 통한 시스템 파괴와 같이 다양해지고 있으며[3], 2015년 12월에 발생한 우크라이나 정전사태처럼 악성코드로 인한 피해 범위도 커지고 있다. 따라서 사이버 안보 관점에서도 악성 MS 오피스 문서

† 준 회 원 : 고려대학교 정보보호대학원 정보보호학과 석사과정
†† 종신회원 : 고려대학교 정보보호대학원 교수
Manuscript Received : November 1, 2016
Accepted : November 23, 2016
* Corresponding Author : Sang Jin Lee(sangjin@korea.ac.kr)

파일을 탐지하고 대응하기 위한 MS 오피스 문서 내 비정상 요소 탐지 기술을 확보할 필요성이 존재한다.

악성 혹은 비정상 MS 오피스 문서 파일을 사전에 탐지하고 피해를 최소화하고자 하는 노력은 계속 진행되었지만, 기존의 시그니처(signature)를 통한 탐지기법은 행위 패턴이 변경되거나 알려지지 않은 신종 행위들에 대해서는 대응하기 힘들다. 최근에는 가상 환경에서 비정상 행위를 유발시켜 악성 코드를 분석하는 샌드박스(sandbox) 기반의 분석 도구들이 개발되었지만, 이를 통해 문서형 악성 코드를 분석하려면 가상환경에서 해당 문서 파일의 뷰어(Viewer)가 반드시 필요하다는 단점이 있다. 또한 분석 시스템을 우회하는 악성 코드들도 존재하기 때문에[1], 악성 코드 뿐만 아니라 다양한 관점에서 비정상 요소들에 대해 탐지할 수 있는 방법에 대한 연구가 필요하다.

이에 본 논문에서는 '.doc', '.ppt', '.xls' 파일과 같은 MS 오피스 버전 97-2003 문서 파일을 대상으로 파일 내부 구조를 파악하여 악성 코드가 삽입될 수 있는 구간을 정의하고, 해당 구간 내의 비정상 요소를 식별하여 문서 파일 내에 비정상적인 행위를 위한 코드를 탐지할 수 있는 방법을 제안한다.

## 2. 관련 연구

MS 오피스를 많은 사용자들이 사용함에 따라 악의적인 목적을 가진 사용자들이 MS 오피스에 악성코드와 은닉데이터와 같이 비정상적인 행위를 위한 코드를 삽입하는데 활용하고 있다. 이에 따라 비정상 MS 오피스 문서 파일을 탐지하고, 실행을 차단하기 위한 여러 가지 탐지 기법에 대한 연구가 진행되었다.

### 2.1 디지털 포렌식 관점에서의 은닉 기법 연구

Byers 등[4]과 Castiglione 등[5]은 MS 오피스의 복합 파일 이진 구조 형식을 기반으로 데이터를 은닉하는 방법을 설명하고, 디지털 포렌식 관점에서 의미있는 데이터를 찾는 방법을 연구하였다.

Park 등[6, 7]과 Yoo 등[8]은 문서 파일 내 데이터 은닉의 가능성을 언급하며, 문서 파일 특징과 파일구조를 분석하고 이에 대한 오피스 문서에 대한 디지털 포렌식 조사 절차 및 분석 방법을 제시하였으며, 필터 도구를 소개하였다.

이와 같이 MS 오피스는 복합 파일 이진 구조 형식으로 미사용 영역 등에 대해서 데이터 은닉에 대한 가능성이 존재한다. 하지만 지능화되는 문서 삽입형 악성코드를 전부 탐지하기에는 부족한 실정이며, 문서 삽입형 악성코드의 특징을 파악하고, 유형화하여 은닉된 비정상 데이터를 탐지하기 위한 추가적인 방법이 필요하다.

### 2.2 문서 삽입형 악성코드 탐지 기법 연구

한국인터넷진흥원의 연구보고서[9]에서는 악성코드 탐지 기법 현황에 대하여 설명하였고, RTF 문서 등 문서파일을 주요 보안 위협 요소로 언급하였다.

Lee 등[10]은 문서 삽입형 악성코드의 악성 행위 정보를 파일 및 레지스트리, 네트워크, 프로세스의 관점에서 분석하여 해당 행위 정보를 기반으로 문서 삽입형 악성코드의 탐지 방식을 제안하였다.

Park 등[11]은 악성코드 바이너리 실행파일에 포함된 API에 대한 호출 빈도수 및 문자열의 유사도를 비교하여 악성 여부를 판별하였다.

HwpScan2[12]는 한컴 오피스의 취약점을 분석하는 도구로 내부 구조 및 알려진 취약점에 대해서 보여주지만, 오피스 프로그램 중 중 가장 많이 사용되는 MS 오피스 문서에 대해서는 분석을 제공하지 않는다.

이러한 연구와 도구들은 Idika 등[3]이 기존에 제시된 시그니처 기반과 샌드박스 기반의 탐지기법에 대한 한계점이 존재하고 새로운 탐지기법의 필요성을 제시한 바와 같이 다양하게 진화하고 있는 악성코드를 정확하게 탐지하기에는 부족하다. 따라서 MS 오피스 문서의 구조와 특징을 디지털 포렌식 관점에서 분석하는 연구와 함께 문서 파일의 악성코드의 특징을 유형화한 연구가 필요하다.

따라서 본 논문에서는 MS 오피스 파일 포맷의 독자적인 구조를 분석하고, 복합 파일 이진 구조의 미할당 영역에 있는 은닉된 비정상 데이터를 판단한다. 그리고 분석된 악성코드의 유형을 파악함으로써 MS 오피스에 삽입된 악성 코드를 문서 포맷 관점에서 탐지하는데 활용한다.

## 3. MS 오피스 파일 내부 구조

MS 오피스는 버전에 따라 복합 파일 이진 구조(Compound File Binary Format)와 OOXML(Open Office XML) 구조로 나뉜다. 복합 파일 이진 구조는 2003 버전까지 사용하던 파일 구조이며, OOXML은 2007 버전 이상에서 사용하는 파일 형식이다.

본 논문에서는 .doc, .ppt, .xls 확장자를 가지는 MS 오피스 97-2003 버전의 문서 파일을 대상으로 문서 파일 내 악비정상 요소 탐지 기법 연구를 진행하였으며, 해당 절에서는 다양한 탐지 기법을 제시하기 위해 복합 파일 이진 구조 형식에 대한 소개와 .doc, .ppt, .xls 파일 별로 내부 구조를 설명하고, 필수 스트림과 그 사용 용도를 소개한다.

### 3.1 복합 파일 이진 구조

복합 파일 이진 형식은 여러 파일과 디렉터리를 하나의 파일에 저장하는 마이크로소프트의 자체적인 파일 형식으로 텍스트, 오디오, 동영상 등 서로 다른 데이터 형식을 포함한다[13]. 이러한 파일 형식은 여러 응용프로그램에서 생성된 데이터를 한 문서 파일에서 편집할 수 있는 환경을 제공한다. 예를 들면, MS 워드 파일에 MS 파워포인트나 MS 엑셀 문서를 삽입하면, 삽입된 MS 파워포인트나 엑셀을 따로 편집하지 않고, 해당 워드 파일만을 편집함으로써 문서를 관리할 수 있는 것을 뜻한다. 이러한 특성을 OLE(Object Linking

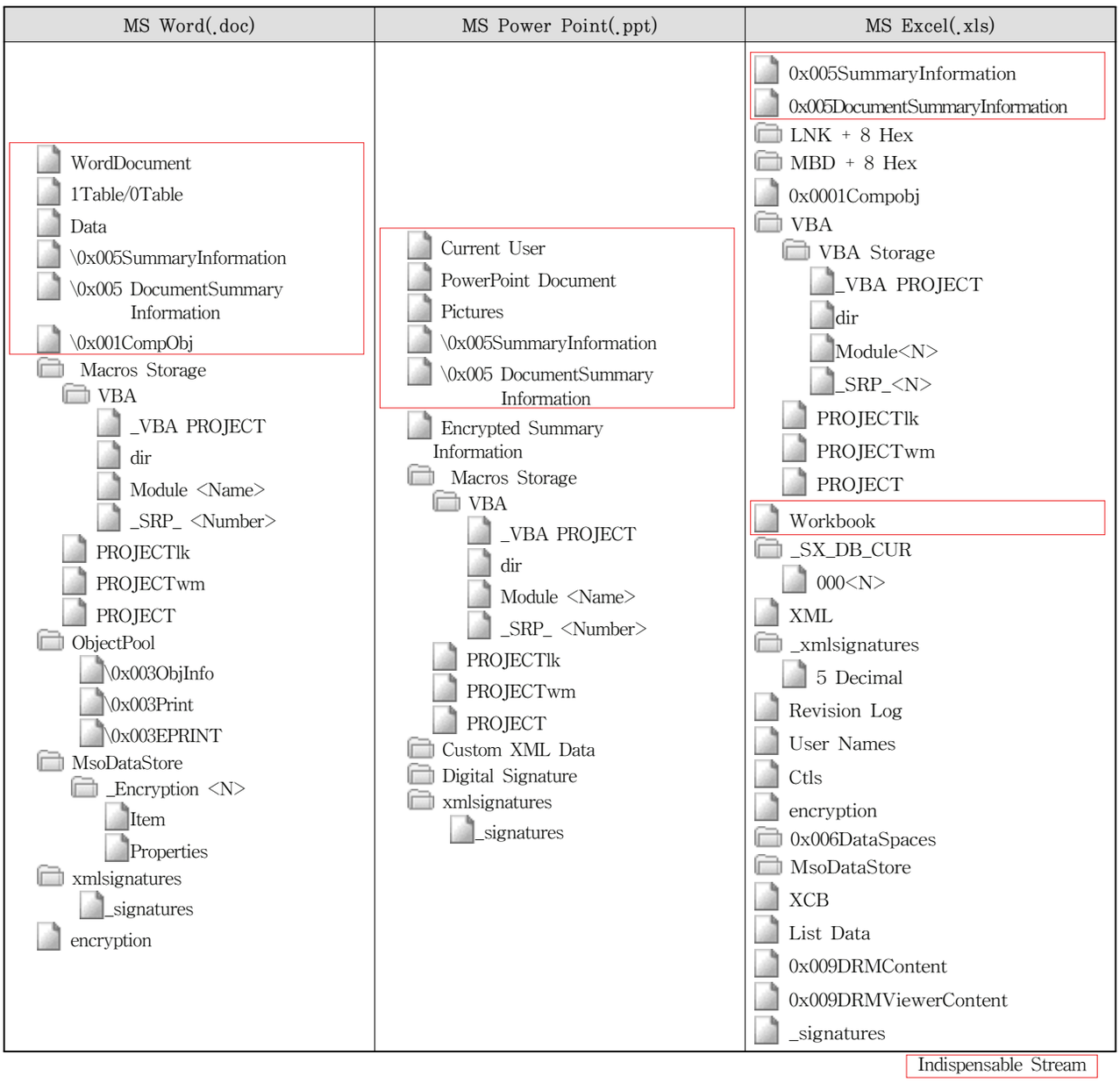


Fig. 1. MS Office Document File Structure

Embedding)라고 부르기 때문에 복합 파일 이진 포맷의 문서를 OLE 복합 문서라고도 한다.

복합 파일 이진 구조는 FAT 파일 시스템과도 유사하며, 스토리지(storage)와 스트림(stream)의 계층 구조로 구성된다. 스토리지는 파일 시스템에서의 디렉터리에 해당하며, 스트림은 파일 시스템에서의 파일과 유사하다. 즉, 스토리지는 하위에 또 다른 스토리지 또는 스트림을 가질 수 있다. 스토리지와 스트림을 관리하기 위해서 메타데이터가 존재하며, 메타데이터는 사용자가 입력한 데이터를 설명하거나 관리하기 위한 데이터로 응용 프로그램에서 자동으로 생성된다.

### 3.2 MS 오피스 문서 파일(.doc, .ppt, .xls) 구조

MS 오피스 파일 내부는 Fig. 1과 같은 구조로 이루어지며, 파일 헤더 정보, 문서 정보, 본문, 삽입 객체, 이미지 데이터, 문서 속성, 매크로 정보, 전자 서명 정보 등이 저장되며, 각 파일 별로 필수 스트림은 상이하다.

Word Document 스트림은 MS 워드의 메인 스트림으로, Fig. 2와 같이 MS 워드 파일 헤더정보와 본문 텍스트 정보를 저장하고 있다[14].

헤더 영역은 FIB(File Information block)라는 MS에서 자체적으로 정의한 데이터 구조로 이루어져 있으며, 문서가


Fig. 2. WordDocument Stream Format

Table 1. Potentially Malicious Code Insertion Possibility Area in MS Office Document (Unit : #, Allow Repetition)

| Category | Details | MS Word | MS Excel | MS PowerPoint |
|---|---|---|---|---|
| Macros | A Macros Code Conducting Some Malicious Acts | 266 | 183 | 17 |
| Data Record Area | Abnormal Structure of BIFF Record | - | 89 | - |
| OLE Object Area | Malicious Embedded OLE Object | 151 | 52 | 213 |
| Unused Area | Extra Data Hided in Unused Area | 98 | 75 | 62 |
| Stream & Storage | Undefined Stream and Storage Area | 32 | 144 | 79 |
| Etc. | Possibility of Additional Data is Out of Range | 3 | 2 | 2 |

작성된 MS 워드의 버전 정보, Table 스트림과 텍스트 영역의 데이터를 가리키는 오프셋 정보 등이 들어 있다. 텍스트 영역에는 블록 단위로 본문의 텍스트가 저장된다. 이 외에 그림 파일은 Data 스트림에, 표와 각종 서식 정보는 Table 스트림에 저장된다.

MS 파워포인트의 레코드는 트리(tree) 형식으로 구성되며, 레코드의 종류에는 아톰(atom)과 컨테이너(container)로 구분된다. 아톰은 레코드의 최소 단위로 레코드 헤더와 데이터로 나타내며, 레코드 헤더는 레코드의 타입과 길이를 저장하고 있으며, 데이터 부분에는 실질적인 데이터가 저장된다[15].

MS 엑셀 파일은 Fig. 3과 같이 Type과 Size, Data로 구성된 BIFF 레코드 구조로 이루어져 있으며, 주로 본문의 데이터를 저장하는 워크, 차트, 매크로 시트 등의 문서가 BIFF 레코드 구조로 저장된다. 필수 스트림 중 Workbook 스트림은 본문에 저장되는 시트에 따라 5개의 서브스트림(Substream)으로 구성될 수 있으며, 서브스트림은 BIFF 레코드 구조들의 모음 형태로 저장되고, 서브스트림이 모여 워크북 스트림을 이룬다[16].

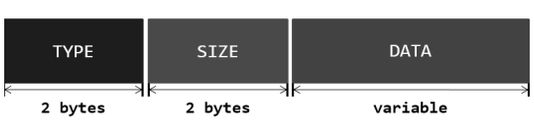

Fig. 3. BIFF(Binary Interchange File Format)

## 4. MS 오피스 내부 비정상 요소 탐지 방법

MS 오피스 문서에는 다양한 비정상 데이터들이 포함될 수 있다. 이러한 비정상 데이터들은 데이터 은닉이 가능한 미사용 영역 및 매크로 기능의 취약점을 이용하여 해당 문서 내에 삽입되고, 문서가 실행될 때 악성코드가 실행되어 사용자의 PC를 감염시키는 등 악성코드로 활용될 수 있다.

MS 오피스 문서 내 비정상 요소를 탐지하기 위해 본 논문에서는 Virus Total과 Malswares.com 서비스에서 Anti Virus에 5개 이상 탐지된 악성 MS 오피스 파일을 각각 포맷별(.doc, .xls, .ppt)로 300개씩 수집하였으며, 정상 파일과 비교를 통해 문서 파일 내에 악성 코드가 삽입 가능한 영역을 Table 1과 같이 정리하였다. 이와 같은 결과를 바탕으로 각 영역에서 비정상 데이터를 탐지할 수 있는 방법 6가지를 제안하였으며, 자세한 내용은 세부 절에서 소개한다.

### 4.1 매크로 내 비정상 요소 탐지

MS 오피스에서는 특정 문구나 행위에 대한 반복적인 기능을 빠르고 효율적으로 사용할 수 있게 매크로 기능을 제공하며, VBA(Visual Basic for Application)를 통해 사용자가 직접 원하는 기능을 소스 코드로 작성하여 사용할 수 있다[17]. 따라서 VBA를 이용하여 복잡한 수식 계산과 대문자-소문자 변환, 특정 값들에 대한 형태 변화와 같은 번거로운 작업을 소스코드로 구현하여 편리하게 사용할 수 있다. 하지만 VBA를 통해 일반적인 파일에 대한 생성과 삭제, 실행 등이 가능하며, 레지스트리와 같은 시스템 파일에 대한 접근도 가능하기 때문에 악의적인 목적을 가진 사용자가 VBA 코드를 활용하여 악성코드를 제작하고 있다.

MS 오피스 파일에 매크로 기능을 추가하면 Fig. 4와 같이 VBA 프로젝트 형태로 스토리지의 하위 스토리지로 생성된다. 해당 스토리지는 MS 워드의 경우 Macros, MS 엑셀과 파워포인트의 경우 VBA Project로 존재하며, VBA 코드는 VBA 스토리지에 생성된 Module 스트림 내에 Run-Length 압축[1]으로 생성된다. 따라서 Run-Length 압축을 해제한 후, VBA 코드를 확인함으로써 삽입된 매크로가 비정상 요소인지 파악할 수 있다.

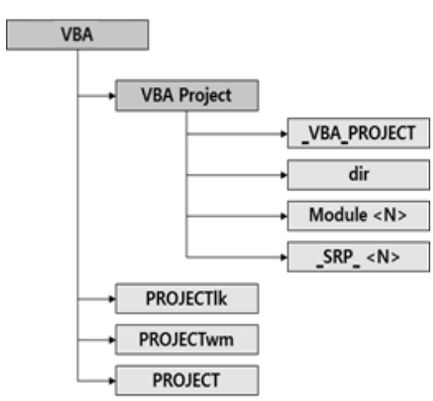

Fig. 4. MS-OVBA Structure

일반적으로 매크로를 이용한 비정상 MS 오피스 문서는 해당 코드를 자동으로 실행하기 위해 자동실행 함수를 활용하여 매크로를 작성한다. VBA 코드에서는 특정 함수 이름

1) 비손실 압축 방법으로, 데이터에서 같은 값이 연속해서 나타나는 것을 그 개수와 반복되는 값만으로 표현하는 방법

Table 2. Autorun Macro Function in MS Office Files

| Category | Function | Details |
|---|---|---|
| Common | AutoOpen | When you Open Document |
| MS Word | Document_Open | |
| | DocumentOpen | |
| MS Excel | Workbook_Open | |
| Common | Auto_Close | When you Close Document |
| | AutoCLose | |
| | AutoExec | |
| | AutoExit | |
| MS Word | Document_Close | |
| | DocumentBeforeClose | |
| Common | DocumentChange | When you Edit document |
| | AutoNew | When you Make New Document |
| | Document_New | |
| | NewDocument | |

을 사용하면 문서 실행 시 함수의 기능을 자동으로 수행하도록 구성할 수 있으며, 해당 함수 이름은 Table 2와 같다.

정상적인 행위의 매크로와 비정상 행위를 하는 매크로를 구분하기 위해서는 함수가 수행하는 기능을 확인해야 한다. 함수 내부에는 비정상 행위에 대한 기능이 구현되어 있다. 예를 들어 외부 url에 접속하여 파일을 다운로드하거나, 윈도우 시스템 명령어를 통해 실행파일을 레지스트리에 등록하는 등의 기능이 존재할 수 있다. Table 3은 악성 코드가 삽입된 오피스 문서 파일 중 자동 실행 매크로가 포함된 문서를 분석한 결과를 바탕으로 비정상 행위 동작에 사용되는 주요 VBA 코드들을 나타낸다. 이를 포함하고 있는 코드는 비정상 행위를 수행한다고 의심할 수 있으며, 특히 Environ, CreateObject, Shell과 같은 VBA 메소드는 대다수의 악성 매크로 파일에서 사용되었다. 근래에는 MS 오피스 문서 파일이 드롭퍼(Droopper)로 사용됨에 따라 URLDownloadToFileA와 같은 메소드도 많이 사용되고 있다.

### 4.2 비정상적인 레코드 탐지

상당수 악성코드들은 MS 엑셀에서 존재하는 Workbook 스트림 내부의 BIFF 구조에서 생기는 취약점을 이용하여 자신이 원하는 바이너리 데이터를 삽입하였다. BIFF는 3.2. 절에서 살펴본 바와 같이 Type, Size, Data의 구조를 가지고 있으며, Type을 나타내는 2byte를 통해 해당 레코드의 기능을 파악할 수 있다. 따라서 명시된 Type 외 다른 값을 가진 레코드의 존재 여부를 파악함으로써 비정상적인 레코드를 탐지할 수 있다.

### 4.3 외부 객체 참조 영역 내 비정상 요소 탐지

MS 오피스는 해당 문서 파일 내부에 다른 문서 파일이나 PDF, 이미지 파일, Adobe Flash 등 32가지의 파일을 삽입할 수 있으며, 삽입된 파일을 객체라고 칭하고, 이러한 기능을 '외부 OLE 객체 삽입'이라고 부른다. 악의적인 목적을

Table 3. Major VBA Method used in Abnormal Behavior

| Keywords | Details |
|---|---|
| Environ | Return the Value of an Operating System Environment Variable |
| Open | Enable File I/O |
| Write, Put, Output, Print # | Write Data on Files (With Open VBA Method) |
| Binary | Read / Write Binary Files (With Open VBA Method) |
| FileCopy CopyFile | Copy Files |
| Kill | Remove File frin Disk |
| CreateTextFile ADODB.Stream WriteText, SaveToFile | Create Files and Insert Specific Strings to Stream |
| Shell, vbNormal vbNormalFoucs vbHide vbMinimizedFocusvbMaximized FocusvbNormalNoFocusvbMinimizedNoFocus WScript.Shell Run | Run Executable Files or System Command |
| PowerShell | Execute Windows System Command |
| Application.Visible, ShowWindow, SW_HIDE | Hide Application |
| Mkdir | Create Directory |
| CreateObject | Create an OLE Object |
| Shell.Application | Execute Application (With CreateObject VBA Method) |
| Lib | Execute Code from DLL |
| CreateThread VitualAlloc | Insert Code to Other Process |
| URLDownloadToFileA Msxml2.XMLHTTP Microsoft.XMLHTTP MSXML2.ServerXMLHTTP User-Agent | Download Files from Internet |
| New-Object System.Net.WebClient DownloadFile | Download Files from Internet with Window Power Shell Command |
| RegOpenKeyExA, RegOpenKeyEx, RegCloseKey RegQueryValueExA RegQueryValueExRegRead | Read / Write Registry Key |
| r'SYSTEM\ControlSet001\Services\Disk\Enum', VIRTUAL, VMWARE, BOX | Search for Bypass Virtual Machine Behavior Detection |
| GetVolumeInformationA, GetVolumeInformation, 1824245000, r'HKEY_LOCAL,MACHINE\SOFTWARE\Microsoft\Windows NT\CurrentVersion\ProductId', 76487-337-8429955-22614, andy, sample, r'C:\exec\exec.exe' popupkiller | Search for Bypass Sandbox based Behavior Detection (Check Process/DLL/System User) |
| SbieDll.dll, SandboxieControlWndClass | |
| r'C:\file.exe', currentuser, | |

가진 사용자는 이러한 해당 파일에서 다른 OLE 객체를 참조할 때 발생하는 취약점을 이용하여 악성 코드 실행과 같은 비정상 행위를 할 수 있다. 해당 취약점을 통해서 악의적인 목적을 가진 사용자는 사용자와 동일한 권한을 획득할 수 있으며, 해당 취약점을 활용하여 쉘코드, ROP(Return-Oriented Programming) 체인, 힙스프레이 등의 공격기법에 활용될 수 있다.

외부 OLE 객체는 참조하고 있는 OLE 객체의 고유 스트림 정보를 함께 저장한다. 일반적으로 파일은 MS 워드의 경우 ObjectPool 스토리지의 하위 스트림 형태로 삽입되며, 엑셀의 경우에는 MBD+8HEX 스토리지의 하위 스트림 형태로 삽입된다. 따라서 상위 스토리지의 정보와 스트림의 내용이 일치하는지 여부를 판단한 후에 해당 스트림 영역을 추출함으로써 삽입된 파일을 획득하고, 획득한 파일에 대해서는 시그니처 기반 등과 같이 일반적인 탐지 기법을 통해 삽입된 파일에 대해 세부 분석을 진행할 수 있다.

### 4.4 미사용 영역 데이터 탐지

앞서 3장에서 살펴본 바와 같이, MS 오피스 파일은 복합 파일 이진 구조로 FAT이나 NTFS와 같은 파일시스템과 비슷한 구조를 가진다. 따라서 내부적으로 슬랙(slack)이나 비할당(unallocated) 영역과 같은 미사용 영역이 존재할 수 있으며, 해당 영역에 쉘코드(API Hashing, Stack frame)와 악의적인 실행 파일과 같은 바이너리 데이터가 저장 가능하다.

따라서 이와 같은 특성을 기반으로 하여 문서 파일 내부에서 실제로 활용되지 않는 미사용 영역을 분류하고, 미사용 영역 내에 데이터의 존재 여부를 판단함으로써 비정상 데이터의 존재 여부를 확인할 수 있다. 더불어 바이너리 데이터는 미할당 영역뿐만 아니라 일반적으로 사용되는 스트림 내부에 사용되지 않는 미사용 영역에도 삽입될 수 있으므로 각 스트림 별 미사용 영역에 대한 데이터 존재 여부를 같이 판단해야 한다.

### 4.5 스트림과 스토리지 구조 검증

3.2.절에서 살펴본 바와 같이 MS 오피스는 데이터의 종류와 목적에 따라 지정된 이름의 스트림과 스토리지가 생성되며, 필수 스트림과 스토리지가 존재한다.

비정상 MS 오피스의 경우 악성 코드와 같은 비정상 행위에 대한 내용을 삽입하기 위해 악의적인 목적을 가진 사용자가 임의로 스토리지나 스트림을 생성하거나 변조할 수 있다. 따라서 해당 MS 오피스 문서 파일 내에 알 수 없는 스트림이나 스토리지가 존재하는지 여부를 판단함으로써 해당 파일의 비정상 요소 존재 여부를 파악한다.

또한, 버퍼오버플로우나 응용프로그램이 사용하는 라이브러리 취약점 등을 사용한 문서 삽입형 악성코드의 경우 해당 파일의 필수 스토리지와 스트림이 존재하지 않을 수 있다. 따라서 파일 내에 필수 스트림과 스토리지의 존재 여부를 살펴봄으로써 손상된 파일이나 악성 코드와 같은 비정상 요소를 파악할 수 있다.

추가적으로 문서 파일 내부의 각 스트림이 정상적인 부모 스토리지 하위에 위치하고 있는지에 대한 검증도 수행함으로써 비정상 요소를 확인할 수 있다.

### 4.6 기타 검증

MS 오피스 문서 파일은 일반적으로 512 byte와 같이 특정 블록 크기 단위로 저장된다. 따라서 해당 파일이 블록 단위로 저장되어 있지 않다면, 비정상적으로 생성되었거나 변조된 것으로 의심할 수 있다. 또한 복합 파일 이진 구조의 관점에서 파일의 유효 범위 이후에 추가적으로 데이터가 존재한다면 파일의 뒷부분에 데이터가 숨겨져 있을 가능성이 존재한다.

### 4.7 비정상 요소 탐지 방법

MS 오피스 문서 내 비정상 요소를 탐지하기 위해서 악성 코드 샘플을 통해 악성코드 삽입 가능 영역을 확인하였고, 해당 영역 별 비정상 요소 탐지 방법을 소개하였다. 이를 바탕으로 Fig. 5와 같이 전체적인 탐지 방법을 정립하였으며, 해당 방법을 적용하여 문서삽입형 비정상 요소 탐지 도구를 개발하였다. 해당 도구는 다음 장에서 소개한다.

## 5. 구현 및 성능 평가

본 논문에서 제시한 MS 오피스 내 비정상 요소 탐지 기법을 바탕으로 Fig. 6과 같이 탐지 도구를 개발하였다.

해당 도구는 3개의 영역으로 구성되어 있으며, ①번 영역은 해당 MS 오피스 문서의 스트림과 스토리지를 확인할 수 있으며, 각 스트림에 대한 내용을 ②번 영역에서 확인할 수 있다. ③번 영역에서는 문서 내에 비정상 행위를 위한 코드가 삽입 가능한 부분에 대한 비정상 여부를 표시한다.

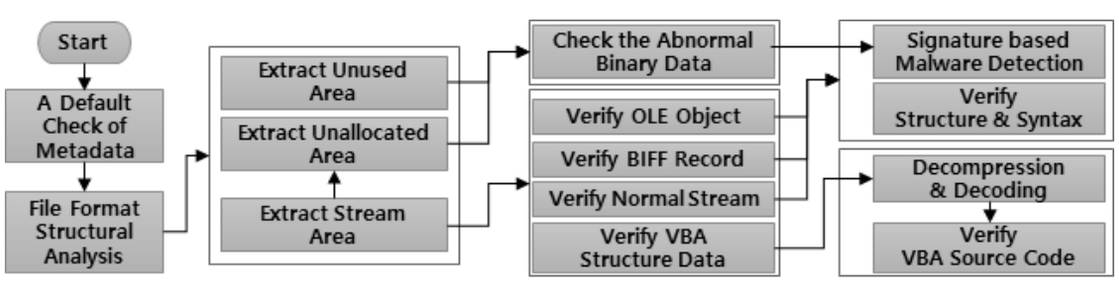


Fig. 5. Detection Process of Anomaly MS Office Document

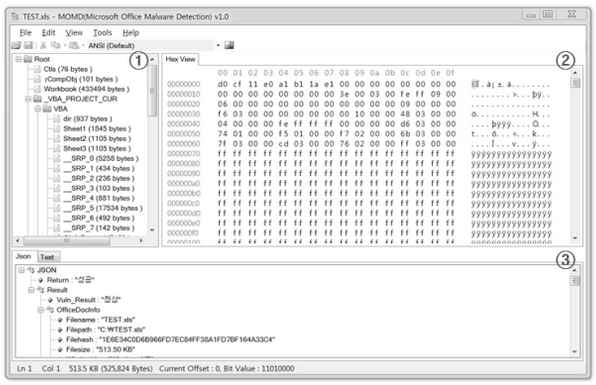

Fig. 6. Abnormal MS Office Files Detection Tool

해당 도구는 3개의 영역으로 구성되어 있으며, ①번 영역은 해당 MS 오피스 문서의 스트림과 스토리지를 확인할 수 있으며, 각 스트림에 대한 내용을 ②번 영역에서 확인할 수 있다. ③번 영역에서는 문서 내에 비정상 행위를 위한 코드가 삽입 가능한 부분에 대한 비정상 여부를 표시한다.

수집한 MS 오피스 포맷별(.doc, .xls, .ppt) 300개의 악성파일과 정상파일 300개의 샘플데이터를 통해 본 논문에서 제시한 탐지 기법을 적용한 탐지 도구와 Anti Virus 제품에 대한 성능 비교를 하였으며, 성능 비교 결과는 Table 4와 같다.

Table 4. Comparison Between the AV(Anti-Virus) Tool and the Suggested Method (#/300)

| | MS Word | | MS Excel | | MS Powerpoint | |
|---|---|---|---|---|---|---|
| | Abnor-mal | Nor-mal | Abnor-mal | Nor-mal | Abnor-mal | Nor-mal |
| Suggested Method | 300 | 3 | 300 | 2 | 300 | 0 |
| ALYac | 137 | 0 | 233 | 1 | 64 | 0 |
| AhnLab-V3 | 169 | 2 | 240 | 0 | 47 | 0 |
| Avast | 298 | 3 | 291 | 2 | 98 | 3 |
| McAfee | 244 | 2 | 263 | 2 | 101 | 4 |
| Kaspersky | 219 | 3 | 285 | 3 | 144 | 5 |

## 6. 결 론

문서 삽입형 악성코드는 국내외 정부 기관과 주요 기업을 비롯하여 기반시설까지 위협하고 있으며, 이에 따라 많은 연구기관과 기업에서 문서 삽입형 악성코드와 같은 비정상 요소를 탐지하기 위해 시그니처 탐지를 비롯하여 다양한 방법들을 제안하고 있다. 하지만 그러한 노력에도 불구하고 이들을 우회하는 악성코드로 인한 침해사고는 계속해서 발생하고 있다. 따라서 본 논문에서는 가장 많이 사용하고 취약점이 많이 공개된 MS 오피스 97-2003 문서 파일을 대상으로 디지털 포렌식 관점에서 문서 파일의 내부 구조를 분석하고, 문서 파일 내 존재하는 다양한 정보들을 통해 비정상 요소를 식별하여 문서 내의 비정상 코드를 탐지하는 기법을 제안한다.

본 논문에서는 MS 오피스 문서 파일의 각 스트림 별로 존재하는 고유한 데이터와 내부 저장 구조를 상세하게 파악하였고, 기존의 알려진 MS 오피스 문서 삽입형 악성코드 900개를 바탕으로 악성 행위의 특징을 유형화하였다. 제시된 방법으로 탐지도구를 개발하였으며, 주요 백신 제품들과의 비교를 통하여 탐지 기법의 우수성을 평가하였다. 이를 바탕으로 악성코드일 수 있는 MS 오피스 문서를 탐지할 수 있음을 입증하였다.

MS 오피스 문서파일은 가장 널리 사용되는 기본 문서 포맷이기 때문에 본 연구의 결과물을 통해 다양한 기관에서 지속적으로 활용될 수 있을 것으로 판단된다. 그리고 VBA 매크로 기능은 오피스뿐만 아니라 MS Visio, AutoCAD 등 다양한 제품군내에서 사용된다. 이들 포맷에서 매크로기능은 오피스의 매크로와 같이 VBA를 사용하기 때문에 본 연구의 VBA 탐지 방법은 오피스뿐만 아니라 VBA를 사용하는 다른 파일 포맷에서도 활용할 수 있을 것이다.

향후에는 MS 오피스 버전 2007 이후부터 사용하는 OOXML 포맷의 문서 내부를 분석하고, 해당 문서 내 비정상 요소를 탐지하는 연구도 진행하여 연구를 확장할 계획이다.

## References


[1] Graham Chantry, New developments in Microsoft Office malware [Internet], https://nakedsecurity.sophos.com/2015/03/06/from-the-labs-new-developments-in-microsoft-office-malware/.

[2] Foetron, MS Office is Still The Productivity Suite Leader [Internet], http://www.foetron.com/microsoft-office-is-still-the-productivity-suite-leader/.

[3] N. Idika and A. P. Mathur, “A Survey of Malware Detection Techniques,” *Purdue University*, 2007.

[4] Simon Byers, “Information leakage caused by hidden data in published documents,” *IEEE Security Privacy*, Vol.2, No.2, pp.23-27, Apr., 2004.

[5] A. Castiglione, De Santis, and C. Soriente, “Taking advantages of a disadvantage: Digital forensics and steganography using document metadata,” *The Journal of Systems and Software*, Vol.80, Iss.5, pp.750-764, 2007.

[6] J. H. Park, Bora Park, S. J. Lee, S. H. Hong, and J. H. Park, “Extraction of Residual Information in the Microsoft PowerPoint file from the Viewpoint of Digital Forensics considering PerCom Environment,” in *Pervasive Computing and Communications, 2008. PerCom 2008. Sixth Annual IEEE International Conference on*. IEEE, p.584-589, 2008.

[7] J. H. Park and S. J. Lee, “Forensic Investigation of MS Office Files,” Graduate School of Information Security, Korea University, Feb., 2009.

[8] B. Y. Yoo and S. J. Lee, “Documents Filter Tool Development for Forensic Investigation,” Graduate School of Information Security, Korea University, Feb., 2011.

[9] KISA, “A Study on Analyzing the Current Malware Detection Technologies and Planning for the Development Model of Detection & Response System,” Research Report, Feb., 2016.
[10] C. Y. Lee, H. G. Kang, T. J. Lee, H. C. Jeong, and Y. J. Won, “A Behavior based Analysis & Detection for Docuent Malicious Code,” The Korea Society of Management Information Systems, pp.532-537, 2012.
[11] J. W. Park, S. T. Moon, G. W. Son, I. K. Kim, K. S. Han, E. G. Im, and I. G. Kim, “An Automatic Malware Classfication System using String List and APIs,” *Journal of Security Engineering*, Vol.8, No.5, pp.611-626, 2011.
[12] Nurilab, HwpScan2 [Internet], http://www.nurilab.net/hwpscan2.
[13] Microsoft Corporation, “Compound Binary File Format Structure Specification,” Microsoft Corporation, 2010.
[14] Microsoft Corporation, “Word Binary File Format(.doc) Structure Specification,” Microsoft Corporation, 2013.
[15] Microsoft Corporation, “PowerPoint Binary File Format(.ppt) Structure Specification,” Microsoft Corporation, 2013.
[16] Microsoft Corporation, “Excel Binary File Format(.xls) Structure Specification,” Microsoft Corporation, 2013.
[17] Scott Driza, Learn Word 2000 VBA Document Automation, Wordware Publishing Inc., 2000.

조 성 혜
e-mail : sunghye0809@gmail.ac.kr
2015년 동덕여자대학교 컴퓨터학과(학사)
2015년~현 재 고려대학교 정보보호대학원 정보보호학과 석사과정
관심분야 : 디지털 포렌식, 정보보호

이 상 진
e-mail : sangjin@korea.ac.kr
1987년 고려대학교 수학과(학사)
1989년 고려대학교 수학과(석사)
1994년 고려대학교 수학과(박사)
1989년~1999년 ETRI 선임연구원
1999년~현 재 고려대학교 정보보호대학원 교수
2008년~현 재 고려대학교 디지털포렌식연구센터 센터장
관심분야 : Digital Forensic, Steganography, Hash Function