문서파일 내의 개인정보탐색 및 암·복호화 시스템

김백엽, 이주희, 최경주
충북대학교 소프트웨어학과
e-mail: yeobssi@gmail.com

A System for Personal Information Detection Encrypt/Discryption

Beckyeob Kim, Juhee Lee, Kyungjoo Choi
Dept of Software, Chung-Buk University

요 약

개인정보보호법의 개정안이 1년간의 계도 기간을 거쳐 2014년 8월 7일에 시행되었다. 개인정보 보호법은 개인정보보호에 관한 법률을 포함하고 있으며 크게 관리적인 보호조치와 기술적 보호 조치로 구분된다. 본 시스템은 기술적인 보호조치에 해당하며, 개인정보가 포함된 문서를 탐지 및 암·복호화 한다. 개인정보를 보호하는 방법에 있어서 개인정보에 해당하는 패턴을 정의하고, 상기 패턴을 참조하여 컴퓨터에 저장된 파일에 개인정보가 포함되어 있는가를 검색하며, 정보보호를 위한 암·복호화 동작을 수행하는 정보보호단계를 제공함으로써, 개인정보가 포함되어 있는 파일의 외부 유출을 차단할 수 있을 것으로 기대된다.

1. 서 론

일반적으로 컴퓨터의 발달로 개인 또는 기업 내에서 이루어지는 작업들이 전산화되어지고 있으며, 이로 인하여 생성되는 데이터들은 개인용 컴퓨터에 저장되거나 별도의 이동식 저장매체에 저장되어 이동 또는 보관되어지고 있다. 이러한 저장매체는 데이터의 저장 및 공유가 가능한 컴퓨팅 환경을 확장한다. 이러한 환경이 개인정보유출을 확장시키며 개인정보 보호방법에 대한 시스템 개발의 필요성이 커지고 있다.

기존연구의 개인정보 암·복호화 시스템은 이와 같은 필요성에 따라 개발되어져 왔다. 그러나 기존연구의 암·복호화 시스템은 사용자가 개인 정보를 검색 할 경우 사용자의 의사 여부와는 관계없이 강제적인 암호화가 이루어졌다. 사용자가 필요로 하지 않는 문서도 암호화함으로써 불편하고 비효율적이었다. 강제적인 암호화 시에도 시스템은 미리 설정된 암호 키만 사용하여 암호 키가 유출되면 모든 문서가 유출 될 수 있는 위험이 있다. 암호화된 파일의 삭제는 복구가 불가능하도록 완전 삭제를 제공하고 있다.

문서파일 내의 개인정보탐색 및 암·복호화 시스템(Document Filter)은 기존연구의 개인정보 암·복호화 시스템의 문제점을 해결하기 위해 이루어진 것으로, 사용자가 개인정보가 포함된 문서 검색을 원할 경우, 컴퓨터 내의 문서를 검색하여 제공하며 검색된 문서의 암·복호화를 통해 개인정보유출을 막을 수 있는 방법을 제공한다.

본 논문에서는 컴퓨터 내의 개인문서가 포함된 문서를 관리하고 암·복호화 하는 시스템을 제안하고자 한다. 2장에서 제안 되는 개인 정보 보호 시스템에 대해 기술하고 실제 구현된 시스템의 결과 및 평가는 3장에 기술한다. 그리고 5장에서 결론을 맺는다.

2. 제안된 개인 정보 보호 시스템

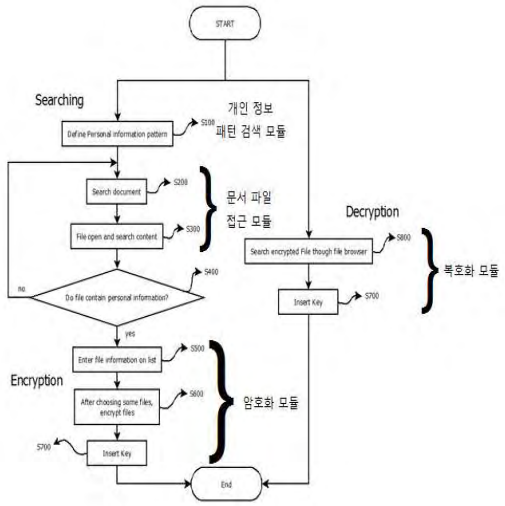
기존에 연구 된 개인정보 암·복호화 시스템은 각종 오피스웨어(txt, hwp, doc, docx, xls, xlsx, ppt, pptx)를 검색하여 사용자에게 개인정보(주민등록번호, 여권번호, 운전면허번호, 금융정보 등)가 담긴 문서를 보여주고 암호화 또한 가능하며 문서 내 개인정보를 *모양으로 표기하여 개인정보를 가린다.

제안된 개인정보보호 시스템(Document Filter)은 기존의 개인정보 시스템과 마찬가지로 오피스웨어(txt, hwp, doc, docx, xls, xlsx, ppt, pptx)를 검색 하여 개인정보를 탐지할 수 있고, 암호화 또한 가능한데, 암호화는 AES-128 방식을 사용하여 보안성을 높였다.

기존 시스템과의 가장 큰 차이점은 암호화에 대한 방법과 암호 키 설정, 문서 완전삭제에 대한 부분이다. 기존 시스템에서의 암호화는 문서 파일을 암호화 하지 않고, 문서 안의 개인정보만 암호화 하는데 한정되어 있다. 또한 암호화를 하려면 '암호 키'가 필요한데 기본적으로 이 키는 HDD의 시리얼 넘버를 사용함으로써 자신의 PC가 아닌 곳에서는 복호화를 할 수 없게 된다. 이에 비해 제안하는 시스템은 문서를 다른 PC로 옮기는 경우에 '사용자 정의

키'를 이용하여 사용자가 직접 정의한 8자리 키로 암호화 할 수 있게 함으로써 다른 PC에서도 복호화를 할 수 있게 된다. 또한 제안하는 시스템은 문서 삭제 기능도 추가하여 불필요한 개인정보 문서를 완전히 삭제하여 개인정보 보호에 대한 대비를 할 수 있다.

개인정보에 해당하는 패턴의 정의는 개인 정보 패턴 검색모듈에서 수행하며, 문서파일 접근 모듈에서 문서 파일을 탐색하여 파일의 내용을 확인하고, 개인정보패턴 모듈에서 개인정보를 검색 한다. 개인정보가 포함된 파일이라면, 파일명, 파일위치, 포함된 개인정보 타입, 보안 등급을 리스트에 나타내준다. 작업이 끝난 후 파일들을 직접 열어 볼 수 있으며, 암호화 작업을 선택적으로 진행한다. 암호화 작업이 꼭 필요하지 않은 파일은 진행하지 않을 수 있다. 암호화 알고리즘은 대칭형 블록 암호 알고리즘인 AES-128 방식으로 암호화를 수행하고, 암호화를 위해서는 키가 필요하다.



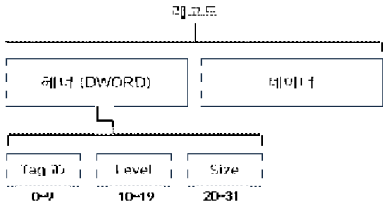
(그림1) 시스템 흐름도

2.1 개인정보 패턴 검색 모듈

개인정보 패턴 검색 모듈은 개인정보의 패턴을 정의하고, 그 패턴에 일치하는 정보를 검색하는 모듈이다. 정의되는 패턴은 사용자의 주민등록번호, 여권번호, 운전면허번호 등 개인정보를 식별하는 정보를 포함하며, 정규표현(regular expression)식과 키워드 복합 매칭으로 정의한다. 즉 예를 들어 주민등록번호를 정규 표현식으로 나타내면 [0-9]{6}(-|[])?(1|2|3|4)[0-9]{6}으로 표현할 수 있는데, 여기서 [0-9]{6}는 0~9 사이의 숫자가 6번 반복됨을 의미하고, (-|[])?는 하이픈(-) 또는 빈 여백이 0번 또는 1번 매치(Hyphen 또는 빈 여백이 있거나 없는)되는 것을 의미하며, (1|2|3|4)는 1 또는 2 또는 3 또는 4에 해당하는 숫자가 나오는 것을 의미한다.

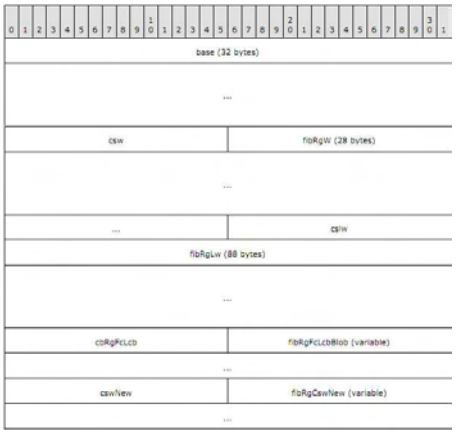
2.2 문서파일 접근 모듈

문서파일에서 개인정보를 검색하기 위해서는 파일에 접근을 할 수 있어야 한다. 문서파일에 접근하는 방식은 파일마다 구조가 다르기 때문에 파일별 접근 방식을 지정해 주어야 한다. 한글파일(hwp)의 경우에는 본문에서 사용하는 문단의 텍스트 데이터 레코드에 접근을 할 수 있어야 한다. Tag ID인 HWPTAG_PARA_TEXT를 사용하여 본문에 접근 할 수 있다.



(그림2) 한글 데이터 구조의 레코드

오피스파일(doc, ppt, xls)의 경우에는 문서의 데이터 스트림에서 Fib구조의 FibRgLw를 통하여 데이터에 접근 할 수 있다.



(그림3) FIB구조

2.3 암호화 모듈

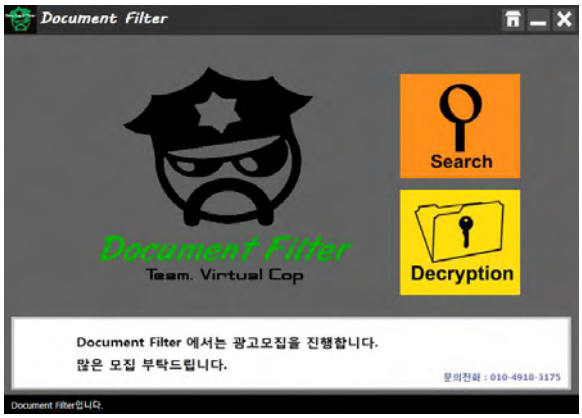
암호화/복호화 작업 시 사용하는 모듈로서 대칭형 블록 암호 알고리즘인 AES-128방식을 사용하여 보안에 대한 신뢰성을 높였다.

3. 결과 및 평가

본 논문에서는 C# 및 XML을 사용하여 설계한 시스템을 구현하였다. 시스템은 사용자의 편의를 위하여 직관적인 인터페이스와 다이얼로그 방식을 사용하였다.

3.1 시스템 실행 과정

다음은 설계한 프로그램을 진행 순서에 맞게 나타낸 것으로, (그림 4)는 Document Filter를 실행 시키면 나타나는 화면이다.



(그림4) Document Filter 메인화면

화면의 Search 버튼을 누르게 되면 (그림 5)와 같이 검색 할 드라이브와 검색할 개인정보 타입, 검사 대상 문서 포맷을 선택하고 검사를 실행하게 된다.

(그림 5) 검사 화면

암호화 버튼을 누르게 되면 기본키와 사용자 정의 키 중 선택하여 암호화를 하게 된다. 원본 파일은 암호화 완료 후 완전 삭제가 된다. (그림 6)은 이러한 흐름을 나타낸다.

(그림 6) 검사완료 및 암호화 과정

(그림 4)에서 Decryption 버튼을 누르게 되면 암호화 된 파일의 리스트가 뜨게 된다. 리스트 중 사용자가 복호화 하려는 파일을 선택하고 복호화 버튼을 누르게 되면 기본키 혹은 사용자 정의 키 중 선택하여 복호화 한다. (그림 7) 은 이러한 흐름을 나타낸다.

(그림 7) 복호화 흐름

3.2 시스템 실행 결과

5가지 포맷(txt, xls, doc, ppt, hwp)을 가진 1,286개의 파일로 시스템을 실험 해 보았다. 쓰레기 값은 랜덤으로 하였으며 결과는 <표 1>과 같다.

<표 1> 실험 결과

	주민번호	여권번호	카드번호
총 파일	453	347	215
검출파일	460	347	228
오탐률	1.52%	0%	5.71%

4. 결 론

기존연구의 개인정보 암·복호화 시스템은 사용자의 선택을 고려하지 않고 강제적으로 암호화를 하고 고정 키 값을 사용함으로써 보안에 취약하였다. 본 논문에서는 단순히 문서를 찾고 암호화에 그치던 시스템을 개선시켜 효율적인 개인정보보호 방법을 제안하였다. 개인정보가 포함된 문서 탐지 및 암·복호화 시스템(Document Filter)으로 불필요한 암호화를 줄일 수 있으며, 문서 검색 범위를 사용자가 선택 가능하게 함으로써 검색시간을 단축시킬 수 있는 효과가 있다. 이 시스템을 기반으로 사용자는 종래의 시스템보다 간편하고 효율적인 시스템을 바탕으로 개인정보유출을 차단할 수 있다.

참고문헌

[1] 박중환, 조남욱, 이기혁, 최일훈, “기업내부 개인 정보보호 시스템 개발”, 정보보호학회지,18권,6호, 2008

[2] 김동례, 심기창, 전문석, “개인정보보호법을 대비한 개인정보보호 시스템에 관한 연구”, 한국정보보호학회지, no6, p19, 2011

[3] ㈜한글과컴퓨터, “한글 문서 파일 구조”, 2010

[4] Microsoft Corporation, “[MS-DOC]: Word (.doc) Binary File Format”, 2014