情 報 保 護 學 會 誌
第19卷 第6號, 2009. 12

개인정보 노출대응 체계

최 진 영*, 하 태 균*, 이 강 신*, 원 유 재*

요 약

IT 기술이 발전함에 따라 인터넷을 통한 개인정보의 수집·이용이 용이해지고 있다. 하지만 이에 대한 역기능으로 웹사이트를 통하여 개인정보가 노출되어 공개되어 있거나, 악의적인 사용자에 의해 개인정보가 유출되어 악용되는 사건이 지속적으로 발생하고 있다. 이렇게 노출 및 유출이 된 개인정보는 오·남용이 되어 국민에게 정신적·금전적 피해를 준다. 이를 해결하기 위해 한국인터넷진흥원은 2006년부터 인터넷 상의 개인정보 노출에 대해 지속적인 삭제조치 등의 노력을 하고 있으나 여전히 개인정보가 노출되고 있다. 따라서, 한국인터넷진흥원은 개인정보 노출을 최소화하기 위해 예방·대응·사후관리를 종합적으로 할 수 있도록 “개인정보 노출대응 체계(Privacy Incident Response SysTem)”를 2009년 11월에 구축하여, 개인정보를 신속히 검색하여 삭제 및 대응할 수 있는 체계를 마련하였다. 본 논문에서는 개인정보 노출대응 체계에 대하여 소개하고자 한다.

Ⅰ. 서 론

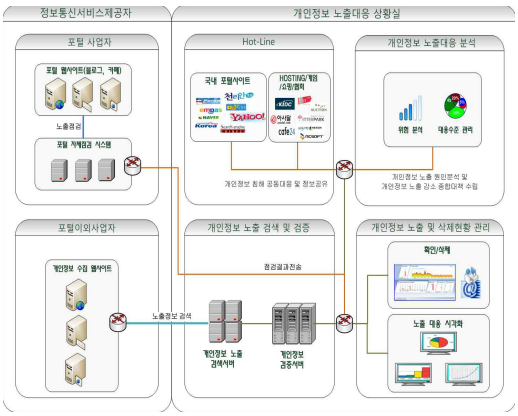
컴퓨터 기술의 꾸준한 발전으로 인터넷도 계속적으로 발전하여, 정보의 유통에 혁명적인 변화를 가지고 왔다. 정보통신 서비스를 제공하는 사업자는 인터넷을 통하여 개인정보를 손쉽게 수집하고 이를 이용해 상업적으로 사용하고 있다. 하지만 이러한 정보유통의 부정적인 효과로 웹사이트를 통하여 국민의 개인정보가 노출되어 공개되어 있거나, 악의적인 사용자에 의해 개인정보가 유출되어 이를 통해 오·남용하여 2차 피해를 발생시키고 있다. 한국인터넷진흥원은 2006년부터 인터넷 상 노출된 개인정보를 삭제하기 위하여 구글의 검색엔진을 이용해 개인정보를 검색하고, 검색된 결과를 지속적으로 삭제 조치하는 노력을 하고 있으나, 여전히 개인정보가 노출되고 있다. 또한, 구글의 검색엔진을 통해 찾는 개인정보는 원사이트를 직접 검색하여 찾는 방법에 비해 검색량이 적으므로 개인정보를 수집하여 보관 및 이용하는 웹사이트를 직접 검색할 필요가 있게 되었다.

한국인터넷진흥원은 웹사이트에 노출된 개인정보를 신속하게 검색하여 검증하고 삭제 조치 등 대응을 할 수 있는 개인정보 노출대응 체계를 구축하였다. 이를 통하여 기존의 3자 제공(구글) 검색엔진의 의존도를 탈피해 독자적인 개인정보 노출 검색엔진으로 대응을 할 수 있게 되었다. 또한 검색대상을 양적으로 확대했으며, 개인정보보호 업무의 효율성을 향상시켜 인터넷상에 노출된 한국인의 개인정보를 크게 감소시킬 수 있는 체계를 갖추었다. 본 논문의 II 장에서는 개인정보 노출대응 체계의 개요를 설명하며, III 장은 개인정보 노출대응 체계 세부 기능 및 알고리즘에 대해 소개하며, IV 장에서는 개인정보 노출대응 체계의 성능에 대하여 비교하고, V 장에서는 결론을 설명한다.

Ⅱ. 개인정보 노출대응 체계 개요

개인정보 노출대응 체계는 웹 페이지를 통해 노출되는 한국인의 개인정보를 신속히 검색하고 이를 삭제조치, 노출 상황전파 등 종합적인 대응을 하는 체계를 말한다. 개인정보 노출대응 체계는 PIRST(Privacy Incident Response SysTem)로 웹상에 노출된 개인정보를 가장 먼저(FIRST) 검색하여 대응하는 체계이다. 이 체계의 개념도는 다음과 같다.

* 한국인터넷진흥원 개인정보보호기획팀



[그림 1] 개인정보 노출대응 체계 개념도

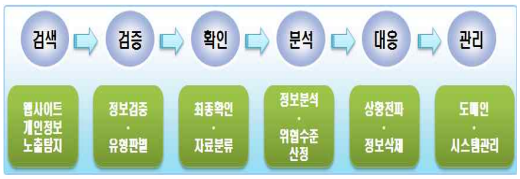
PIRST의 역할은 웹사이트에 노출된 개인정보를 검색하여 검증, 개인정보 노출 삭제 현황 관리 및 삭제 지원, 개인정보 노출 원인 분석 및 노출 감소 종합대책 연구, 국내 개인정보 삭제·정보공유를 위한 핫라인 운영, 국외에 노출된 개인정보 삭제 협력을 수행한다.

[그림 2] 개인정보 노출대응 체계 특징

PIRST는 신속성, 정밀성, 확장성, 자동화의 특징을 갖는다. 첫째, 검색 대상을 최초 검색한 이후 다음에 다시 검색을 할 때 신규 웹페이지만 방문하여 검색하므로 중복되지 않도록 검색하여 신속성을 높였다. 둘째, 다양한 웹페이지 언어(JavaScript, 플래시 등)를 분석하여 개인정보 노출여부를 판단할 수 있도록 파이어폭스 웹브라우저에서 사용된 자바스크립트 해석엔진을 튜닝하여 해석 능력을 높여 정밀성을 향상 시켰다. 셋째, 개인정보라 판단이 되는 패턴을 추가하여 검색 범위를 확장할 수 있다. 넷째, 개인정보가 노출된 페이지를 확인하고 삭제 지원하기 위해 업무 처리를 자동화하여 편리함을 제공할 수 있다.

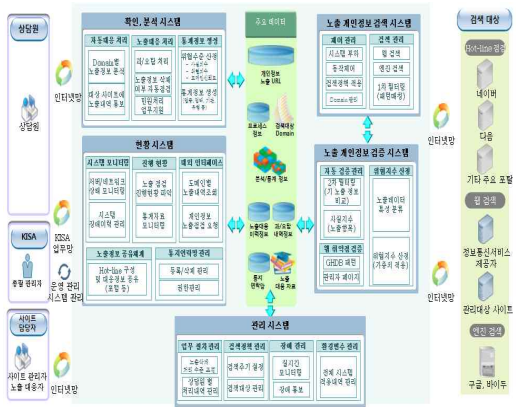
Ⅲ. 개인정보 노출대응 체계 세부 기능

PIRST는 검색, 검증, 확인, 분석(현황), 대응, 관리의 총 6개 논리적 시스템과 물리적 시스템의 관제와 제어를 담당하는 제어시스템으로 구성이 된다.



[그림 3] 개인정보 노출대응 체계 논리적 시스템

각각의 논리적 시스템은 확장성을 고려한 응용 프로그램 인터페이스(API)를 통해 동작하여 유연성·한정성을 최대화하여 설계하였다.



[그림 4] 개인정보 노출대응 체계 기능도

각각의 논리적인 시스템은 다음과 같은 기능을 수행한다.

1. 검색 시스템

검색 시스템은 검색대상 웹사이트의 페이지들을 검색하여 개인정보가 노출이 되었는지 필터링을 통하여 검출하며, 구글·바이두와 같은 기존의 검색엔진을 이용하여 개

인정보가 노출된 페이지를 검색할 수 있도록 설계되었다.

최근 인터넷 홈페이지 기술의 발달로 게시판을 구성할 때, JavaScript 기반의 페이지 이동함수 및 DHTML 등 동적인 웹페이지를 많이 사용하고 있다. 개인정보 노출은 게시판을 통하여 노출되는 사례가 많으므로 동적인 페이지를 효과적으로 분석하고 URL을 추출하여 처리할 수 있도록 검색엔진을 개발해야 한다.

또한 검색 속도 향상을 위해 초기 방문 이후 재방문 시 신규 페이지만을 방문하는 기법으로 UPM(URL Parameter Matrix)과 HTTP 헤더정보를 이용하여 페이지의 변경 여부와 신규로 생성된 페이지만을 검색할 수 있도록 알고리즘을 적용하였다. 각각의 알고리즘은 다음과 같다.

1.1 UPM 기법

검색엔진의 결과 파일 중 동적인 웹페이지들은 Full URL 리스트파일(domain.urllist)을 이용하여 URL 파라미터 메트릭스를 생성한다. Full URL은 '?'로 구분되는 페이지 스트링과 파라미터 스트링으로 나눌 수 있다. UPM은 페이지 문자열을 Key로 하여 여러 개의 파라미터 문자열 리스트를 Value로 가질 수 있다. 파라미터는 파라미터명과 파라미터 타입, 파라미터 값을 가지며 "name:type:value"으로 표현된다. type 구분으로 스트링형 파라미터는 S, 정수형 파라미터는 I로 표현된다. 정수형 파라미터는 동일한 파라미터 명을 가지더라도 파라미터 값이 다르면 다른 파라미터가 되며, 정수형 파라미터는 파라미터 명이 같으면 동일한 파라미터로 보며 파라미터 값들 중 최대 값을 파라미터 값으로 설정한다. 이 값을 이용해 해당 웹사이트에 다시 방문할 경우 이 값보다 높은 값이 있는지를 찾아 신규 페이지를 검색하여 검색 효율을 높인다.

1.2 HTTP 헤더 정보 분석 기법

검색엔진의 결과 파일 중 정적인 웹페이지들은 HTTP 헤더 정보를 얻고 반환코드가 200인지 확인하여 페이지를 저장한 후 페이지 크기와 시그니처를 생성한다. 파라미터를 삭제하지 않은 페이지의 시그니처, 페이지 크기와 비교하여 동일한 페이지인지 확인한다. 반환코드가 200이 아니거나 동일한 페이지가 아닐 경우 신규 페이지라 인식하여 검색하게 된다.

2. 검증 시스템

검색 시스템에서 검색된 다량의 웹자원에서 정확하게 개인정보 노출 사실을 추출하기 위해서 검증 시스템은 검색한 페이지를 분석하여 개인정보 유효성을 적용하고 위협지수, 사실지수를 산정한다. 이 두 지수는 후에 확인 시스템에서 신뢰도 지수와 결합하여 위협수준을 측정하는데 사용이 되며, 위협수준은 PIRST의 검색 대상 목록의 검색 주기를 동적으로 설정하는데 사용이 된다. 위협지수, 사실지수 산정 방식은 다음과 같다.

2.1 위협지수(value) 산정

검증 시스템에서 각 페이지별 노출정보 속성(노출 유형, 건수, 노출 유형 조합)을 이용해 수치를 산정한다.

$$value_k = \sum_{k=1}^{n} (S_k \cdot C_k \cdot P_k + W_k) \qquad (1)$$

위협지수를 산정하기 위해 사용되는 파라미터는 다음과 같다.

[표 1] 위협지수 피라미터

파라미터	설명
n	개인식별정보의 개수(9개)
S	사실지수
C	개수
P	개인식별정보의 가중치
W	웹취약점

2.2 사실지수(S) 산정

검증 시스템에서 개인식별정보(PII-Personal Identification Information) 노출로 판단된 것에 대해 사실인지를 판단하여 지수를 설정한다.

3. 확인 시스템

검증시스템에서 필터링한 노출 개인정보/취약점을

육안 확인하고, 노출 개인정보/취약점 위협수준 산정 인터페이스를 통해 노출/취약점의 사실여부 및 심각성 여부를 판단하며, 노출개인정보/취약점 대응수준 인터페이스를 통해 자동처리, 상향분배 및 이력을 추적할 수 있다. 또한, 담당자가 삭제 요청을 할 수 있도록 SMS, e-mail 전송 인터페이스를 제공하며, 담당자의 삭제 조치에 따라 위협수준에 사용이 될 신뢰도 지수를 산정되게 된다.

3.1 신뢰도 지수(T) 산정

삭제 담당자가 처리한 내역과 PIRST 시스템의 대응 이력을 이용해 웹사이트별로 신뢰도를 생성한다. 예를 들어 노출된 사이트를 삭제에 대한 대응을 신속히 해주거나, 노출 건수가 작을 경우 신뢰도가 높아진다.

$$T_k = \sum_{k=1}^{n} \frac{R_k}{Total} W_k \qquad (2)$$

신뢰도 지수를 산정하기 위해 사용되는 파라미터는 다음과 같다.

[표 2] 신뢰도 피라미터

파라미터	설명
R	대응 건수
Total, n	노출 건수
W	대응 기간 가중치

이렇게 생성된 신뢰도 지수는 후에 분석 시스템에서 사이트별 위협수준을 사정할 때 사용이 된다.

4. 분석(현황) 시스템

분석 시스템은 검증 시스템과 확인 시스템에서 산출된 위협지수, 사실지수, 신뢰도 지수를 통해 위협수준을 산정하고, 노출 대응 현황 분석을 위한 통계 자료를 생성한다. [그림 5]는 개인정보 노출대응 상황실에서 관리하는 통계화면들로 개인정보 노출에 대해 신속하게 대응을 하게 된다.

[그림 5] 개인정보 노출대응 체계 통계 화면

이러한 통계 화면은 주민등록번호 노출 추이, 주요 개인정보(주민등록번호 제외) 노출 현황, 실시간 개인정보 노출 규모, 위험군별 대상 웹사이트 분포, 주민등록번호 노출 추이(구글검색), 주민등록번호 노출 추이(바이두검색), 국가별 주민등록번호 노출 추이(구글,바이두검색), 전세계 주민등록번호 노출 분포, 국내 노출 주민등록번호 삭제 처리현황, 국내 노출 주요 개인정보(주민등록번호 제외) 삭제 처리현황, 중국노출 주민등록번호 삭제 처리현황, 국외 노출 주민등록번호 삭제 처리현황, 웹사이트 검색 처리현황, 위험군별 검색 현황, 당일 검색 완료율, 주민등록번호 노출 삭제 처리현황, 검색 완료 웹사이트 현황의 그래프를 표현하여 상황실 근무자가 실시간으로 개인정보 노출추이를 모니터링하고, 노출 원인 분석을 통한 개인정보 감소 종합대책을 연구할 수 있는 자료를 제공해 준다.

분석 시스템에서 생성된 각종 지표들을 확인하여 관리 시스템에서 검색, 검증, 대응에 대한 정책 반영이 수행되어 지능화된 시스템 구축이 가능하다. 또한, 배치 batch job을 주로 수행하는 분석 시스템을 Green IT 개념을 도입하여 batch job을 수행하지 않는 유휴 시간에 검증 시스템 역할을 수행할 수 있도록 해당 모듈을 탑재하는 등 최적화된 자원을 활용할 수 있는 시스템으로 구축하였다.

4.1 위협수준(val) 산정

위험 수준을 산정하기 위한 수식과 사용되는 파라미터는 다음과 같다.

$$val = \sum_{k=1}^{n} (value_k \bullet T_k) \qquad (3)$$

[표 3] 신뢰도 피라미터

파라미터	설명
n	해당 웹사이트의 노출 페이지 건수
T	해당 웹사이트의 신뢰도
value	위협지수

이렇게 산정된 위협수준은 후에 검색 대상 목록을 각각의 서버에 분배할 때, 이 값을 기준으로 5개의 검색그룹(고위험군, 위험군, 다소위험군, 보통군, 안정군)으로 나누어 동적으로 검색 정책을 설정하는 척도가 된다.

예를 들어, A라는 웹사이트에서 대규모의 개인정보가 노출이 되어 PIRST 시스템에 검색되었다면, PIRST 시스템은 이 페이지에 대한 위협지수, 사실지수와 해당 웹사이트의 신뢰도 지수를 이용해 위협수준을 새롭게 산정하여, 이 웹사이트가 자주 검색될 수 있도록 검색대상의 그룹을 재조정하여 시스템의 유연성을 제공한다.

5. 대응 시스템

확인/분석시스템을 통해 검증되어 확정된 노출개인정보/취약점에 대한 위협수준이 설정되고, 대상 도메인 관리자에게 이메일, SMS, 전화 또는 공문의 방법으로 해당 사이트에 시정조치 요청을 의뢰하며, 요청 된 시정조치에 대한 이력을 추적 관리하도록 기능을 제공한다. 또한 이슈 사항, 정보 등을 공유하는 인터페이스를 제공한다. 대응 시스템에서 동작하는 노출삭제프로세스와 검색대상 사이트 관리자에게 제공하는 웹 인터페이스 등으로 구성된 대응 시스템은 실제 사이트 담당자에게 각 사이트의 개인정보 위협 내용을 알려주는 인터페이스를 갖는다.

6. 관리 시스템

관리 시스템은 노출삭제 프로세스 관리, 검색정책 관리, 환경변수 관리 등 PIRST 시스템의 전체적인 정책과 시스템 자원을 관리할 수 있는 인터페이스를 제공한다. 세부적으로 노출 개인정보 대응 프로세스 수행결과, 검색정책, 노출/취약점 시그니처 및 환경변수 관리, 점검대상 도메인 관리, 노출점검시스템 서버/네트워크 상태관리, 개인정보 노출대응 상황실 연동 데이터항목 설정 등 노출대응 체계 시스템의 통합 관리를 제공할 수 있도록 시스템을 구축하였다.

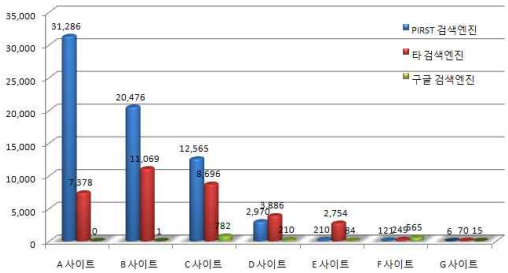
특히 분석 시스템에서 산정된 위협수준을 바탕으로 5개의 검색그룹으로 검색 대상 웹사이트를 분류하여 서로 다른 주기로 웹사이트에 방문하여 노출된 개인정보를 찾도록 정책을 설정하는 기능으로 시스템을 유연하게 운영할 수 있다.

[표 4] 검색그룹별 주기

검색그룹	검색 주기
고위험군	1일
위험군	7일
다소위험군	14일
보통군	30일
안정군	60일

Ⅳ. 성능 평가

PIRST의 개발 과정에서 검색 시스템의 검색 엔진 성능 평가를 하였다. PIRST 검색 엔진과 구글의 검색 엔진, 타 검색엔진을 검색량을 비교하기 위해 7개의 웹사이트 선정하여 테스트하였다.



[그림 6] 검색량 비교 결과

다음과 같이 PIRST 검색엔진의 성능을 개선하여 타 검색엔진과 성능을 비교해 본 결과 대부분 더 많은 웹 자원을 검색하는 것으로 측정되었다.

PIRST의 시스템은 전반적으로 다음과 같은 사항을 개선하였다.

[표 5] 개인정보 노출대응 체계 기대효과

구분	기존 방식	개선방식	기대효과
성능	모든 페이지 재검색	초기 검색 이후 신규 페이지만 검색	동일 규모로 검색 대상 사이트 확대
	수동에 의한 검색	시스템에 의한 검색	인력절감
			점검시간단축
범위	주민등록번호 등	PII 9개 항목 (주민등록번호, 운전면허번호, 여권 번호 등)	다양한 개인정보 노출 검색
	검색 항목 확대시 코드 수정	패턴 입력에 의한 항목확대 적용	점검 유연성 확대
처리	E-mail/삭제확인 등 사람에 의한 수동 업무 처리	시스템에 의한 자동 업무 처리	신속한 노출 대응처리
			업무 편의성 제공
	전체적 현황파악 불가	시스템에 의한 자동 업무 처리	신속한 판단 및 대응
	수작업에 의한 통계	편리한 사용자 인터페이스	

V. 결 론

PIRST는 기존의 구글 점검 등 3자 제공 검색엔진의 의존을 하지 않고 독립적 시스템으로 구축되어 개인정보 노출의 신속한 검색을 통해 검증하고 상황에 맞게 대응해 개인정보 오·남용의 2차 피해를 사전에 예방할 수 있다. 그리고 개인정보보호 업무의 효율성을 고려하여 자동화된 업무 절차와 직관적인 사용자 인터페이스 구현하였다. PIRST 체계 구축으로 국민의 개인정보보호를 효과적으로 할 것으로 기대된다.

향후 새로운 언어 및 기술 개발로 다양한 웹페이지가 등장할 것이며, 검색엔진의 지능화에 따라 잠재적인 개인정보의 노출 위협이 있으므로 새로운 검색 기술에 대해 지속적으로 연구할 필요가 있으며, 계속적인 운영을 통하여 개선 사항을 도출할 것이다.

<著者紹介>

최 진 영 (Jin-young Choi)
정회원

2007년 2월 : 성균관대학교 컴퓨터공학과 졸업
2009년 2월 : 성균관대학교 전자전기컴퓨터공학과 석사
<관심분야> 개인정보보호, 센서 네트워크

이 강 신 (Gang Shin Lee)
정회원

1989년 8월 : 한양대학교 수학과 이학석사
2005년 8월 : 고려대학교 정보보호대학원 공학박사
1990년 7월 ~ 1992년 6월 : 데이콤 종합연구소 연구원
1992년 7월 ~ 2000년 8월 : 한국전산원 정보화표준부장
2000년 9월 ~ 현재 : 한국인터넷진흥원 개인정보보호기획팀장
<관심분야> 개인정보보호, 네트워크보안, 정보보호아키텍처

하 태 균 (Tae-gyun Ha)
정회원

2001년 2월 : 인하대학교 수학과 졸업
2005년 8월 : 불법스팸대응센터 민원처리시스템 구축
2005년 9월 : 실시간스팸차단리스트(KISARBL) 구축
2006년 10월 : 도메인신뢰도평가시스템 구축
<관심분야> 정보보호, DRM, 홈네트워크

원 유 재 (Yoo-jae Won)
정회원

1987년 : 충남대학교 계산통계학과 이학석사
1998년 : 충남대학교 전산학과 이학박사
1987년 ~ 2001년 : 한국전자통신연구원 팀장
2001년 ~ 2004년 : 안철수연구소, 안랩유비웨어 CTO
2004년 ~ 현재 : 한국인터넷진흥원 단장
<관심분야> VoIP 보안, IPv6 보안, 멀티캐스트 보안, 무선 인터넷 보안, PKI 등