논문 2011-6-32

인터넷 웹에서의 개인정보보호 시스템 구현

Personal Information Protection System for Web Service

황호영*, 김남윤**

Hoyoung Hwang, Namyun Kim

요 약 인터넷에서의 개인정보 보호의 중요성이 갈수록 증가하고 있다. 개인정보의 노출은 대고객 서비스의 접점인 홈페이지 상의 게시판 및 첨부파일을 통해 주로 노출되고 있으며, 운영 중인 게시판과 게시물 첨부파일의 숫자가 워낙 방대하여 관리자의 수작업으로는 기 존재하거나 최신 입력되는 개인정보를 진단하고 차단하는 데는 한계가 있다. 따라서 홈페이지를 운영하는 기관에서는 효율적인 개인정보보호를 위하여 자동화된 개인정보 스캐닝 및 필터링 솔루션이 필요하다. 본 연구에서는 인터넷과 웹 등에서 개인정보의 노출을 효과적으로 차단 및 치환하기 위한 개인정보보호 시스템을 개발한다.

Abstract Recently, Internet web services including bulletin board of companies, public intuitions, or governments became the main ways of personal information leakage. However, it is hard to succeed to detect personal information and protect it from leakage manually by system administrators, because the number of web boards and attached files is so huge. Therefore, automatic scanning and filtering methods are needed to protect personal information like resident registration numbers from unexpected exposure. To do this, we developed a novel and advanced web security system for automatic personal information scanning and filtering.

Key Words : 개인정보, 정보보호, 웹, 스캐닝, 필터링

Ⅰ. 서 론

인터넷 및 모바일 기기가 정보사회의 핵심인프라로 기능하면서 사용자의 자기식별용 개인정보의 유출로 인한 피해가 급증하고 있다. 최근 온라인 포털이나 게임에서의 개인정보 유출 관련 사고가 잇따르는 등 개인정보의 노출로 인한 피해 사례가 급증함에 따라 인터넷상에서의 개인정보보호에 대한 사회적인 관심이 고조되고 있다. 개인정보 유출로 인한 기관 및 기업의 손실은 기관의 신뢰성 하락은 물론 손해 배상으로 인한 금전적인 손해까지 영향을 미치는 중대한 사유이다. 이에 따라 공공기관 및 개인 정보를 취급하는 금융, 홈쇼핑, 인터넷 포탈 등 민간 기업에서도 개인정보보호에 대한 적극적인 대응을 추진 중이다. 공공부문의 경우 행정자치부 전자정부제도 팀 및 정보보호진흥원 등을 통해 이러한 문제점을 적극 대응하고 있고 민간 기업들도 개인정보를 취급하는 대기업 중심으로 개별적인 대응 방안을 수립하고 있는 중이다.[1][2][4]

또한 개인정보보호에 대한 사회적인 관심이 고조되면서 2009년 7월 1일부터 시행된 [정보통신망 이용촉진 및 정보보호등에 관한 법률] 제67조는 통신사업자 외에 개인정보보호를 의무화해야 하는 준용사업자를 기존의 8개 업종에서 22개 업종으로 확장하여, 개인정보보호에 대한 법적 처벌 근거를 강화하였다. 즉, 기존 8개 업종(여행사, 호텔, 항공사, 학원, 휴양콘도, 대형마트, 백화점, 체인점)에서 추가하여 14개 업종을 새로 지정함으로써 주택건설, 주택관리, 건설기계대여, 자동차매매, 자동차대여,

*정회원, 한성대학교 멀티미디어공학과
**정회원, 한성대학교 정보시스템공학과 (교신저자)
접수일자 2011.9.20, 수정완료 2011.11.7
게재확정일자 2011.12.16

의료기관, 정유사, 체육시설, 직업소개소, 결혼중개업 등 약 22만여 업체가 준용사업자에 포함되었다. 개인정보보호 법적 의무화 준용사업자의 확대로 보다 많은 업종의 기업 및 사업자가 개인정보보호에 대한 대책과 자동화된 유출 방지 솔루션을 확보해야 할 필요성이 커졌으므로, 산업적인 면에서도 공공기관 및 교육기관 대상의 시장 외에 민간 기업 및 사업자를 대상으로 하는 개인정보보호시스템 시장이 형성될 수 있는 시점이다. 이 경우 주된 보호의 대상이 되는 개인정보는 주민번호, 핸드폰번호, 신용카드 번호, 계좌번호, 사업자번호, 법인번호, 이메일 등 일정한 패턴을 가지는 정보를 포함하며, 이는 개인의 프라이버시 보호 및 명의도용 방지를 위한 필수 항목이다. 이 외에도 명예 훼손, 유해 정보의 노출 등을 방지하기 위하여 운용자가 지정한 패턴을 탐지하고 보호하는 기능을 가짐으로써 특정 단어나 욕설 등을 필터링 할 수 있어야 한다.[8][9][10] 개인정보보호의 대상이 되는 데이터는 그림 1과 같다.

그림 1. 웹에서 유통되는 개인정보 데이터
Fig. 1. Personal Information on the Web.

개인정보의 주된 노출은 대국민 서비스 및 대고객 서비스의 접점인 홈페이지 상의 게시판 및 첨부파일을 통해 노출되고 있으며, 운영 중인 게시판과 게시물 첨부파일의 숫자가 워낙 방대하여 관리자의 수작업으로는 기존재하거나 최신 입력되는 개인정보를 진단하고 차단하는 데는 한계가 있다. 따라서 자동화된 개인정보 노출 진단을 위한 스캐닝 및 필터링 솔루션이 필요하다.[5][6][7] 본 논문에서 구현한 솔루션은 이러한 홈페이지 콘텐츠 상의 개인정보를 자동화 된 스캐닝 기능으로 찾아내고 입출력되는 내용에 대한 콘텐츠 필터링을 통해 홈페이지 상에 무분별하게 노출되는 개인정보를 효율적으로 관리하고 통제하기 위한 솔루션을 제공한다.

II. 본 론

1. 개발 시스템 구조

인터넷과 웹 게시판 등에서 범람하는 개인정보의 노출을 효과적으로 차단 및 치환하기 위하여 보안 프로토콜(SSL)을 지원할 수 있으며, 보다 빠른 성능을 보장하는 진보된 개인정보 보호 시스템을 개발하였다. 자동화된 개인정보 노출 진단을 위한 스캐닝 및 필터링 솔루션을 결합하여 홈페이지 상에 무분별하게 노출되는 개인정보를 효율적으로 관리하고 통제하기 위한 솔루션을 구현하였으며, 그 기능적인 구성은 그림 2와 같다.

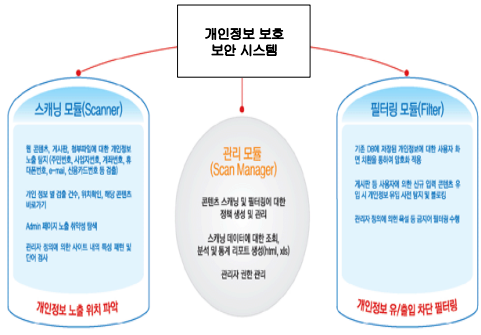

그림 2. 개인정보보호시스템 구조
Fig. 2. Sucture of Personal Information Protection System

전체 시스템은 스캐닝 모듈, 필터링 모듈, 그리고 관리 모듈로 구성되어 있다. 스캐닝 모듈은 관리자 정의에 의한 사이트 내의 특정 패턴 및 단어 검사를 통해 게시판, 첨부 파일, 웹 콘텐츠에 대한 개인정보 노출을 탐지하여 노출 건수, 위치를 확인하고 취약성을 탐색한다. 필터링 모듈은 스캐닝 모듈을 탐지 결과에 따라 기존 DB에 저장된 개인정보와 게시판등에 새로 신규입력된 콘텐츠에 대하여 사용자 화면 치환을 통한 암호화 적용 및 블로킹 작업을 수행한다. 관리 모듈은 사용자 권한관리와 함께 콘

텐츠 스캐닝 및 필터링에 대한 정책을 생성하고 관리하는 기능을 수행하며, 스캐닝 데이터에 대한 조회, 분석 등 통계리포트를 생성한다.

기본적인 서비스 모듈 이외에 웹서버의 부하를 줄이고 더 빠른 서비스가 가능하도록 정적인 파일(이미지 등)에 대한 캐싱 알고리즘을 통해 실제 서비스에서 개인정보 탐지 프로세스로 인한 성능 저하 요인을 보완하고 웹 가속 효과를 기대할 수 있도록 개발하였다.

Content 압축 기술은 파일에 대한 캐싱과 더불어 웹 가속을 이루는 주요한 기능중의 하나이다. 본 시스템에서는 텍스트 컨텐츠에 대해서 gzip 또는 deflate 형식으로 압축하여 전체적인 네트워크 전송량을 줄이는 방법을 적용하였다. 또한 Multi-Thread를 이용한 병행 처리로 엔진 성능을 극대화하고, 중복 스캔 배제, URL 및 패턴 등록, Connection Timeout 및 Thread 수 등을 사이트 상황에 맞게 조절하여 스캐닝 성능을 극대화함으로써 웹 가속 성능을 높이도록 구현하였다.

2. 개발 시스템 기능 구현

개인정보보호시스템의 주요 기능은 다음과 같다.

그림 3. 개인정보보호시스템 구현 기능
Fig. 3. Functions of Personal Information Protection System

- HTTP 1.1 지원(Request/Respose 패킷 완벽 처리)
- 다양한 전송 방식(multipart-form, chunking, gzip, deflate, ssl 등) 지원
- 개인 정보(주민번호, 핸드폰번호, 신용카드 번호, 계좌번호, 사업자번호, 법인번호, 이메일 등) 패턴 및 사용자 지정 패턴 탐지 지원
- 정적 캐싱 및 컨텐츠 압축 기술을 통한 웹 가속 기능 지원
- 멀티도메인 지원
- 웹UI 형식의 관리자 모듈을 통한 모니터링 및 통계 기능 지원

위와 같은 기능을 지원하기 위한 세부 구현 과정과 내용은 다음과 같다.

- 전체 시스템 구조 설계

스캐닝 모듈, 필터링 모듈, 관리 모듈의 개발 범위에 따라 구체적으로 S/W 구성을 설계하고, H/W 설계를 위해 어플라이언스를 구성하기 위한 장비를 선정하고, 해당 장비에 맞는 네트워크 카드 등을 탑재하고 테스트를 수행하였다.

- HTTP Proxy 모듈

정보보호시스템의 기본 기능을 수행하면서도 성능저하가 없도록 탑재한 시스템의 안정성 및 응답 속도가 중요한 요소이다. 안정적인 서비스를 위한 멀티쓰레드 환경 구성과 HTTP 1.1 프로토콜 및 Proxy 기술을 이용하여 안정성 및 성능의 보완을 구현하고 기본적인 Throughput Test를 실시하였다.

- Protocol 모듈

Transfer-Encoding, Content-Encoding, Connection 옵션 등으로 인해 다양한 전송 방식(Close, Keep-Alive, chunked, gzip, deflate 등)을 이해하고 특성별 정확한 데이터 분석이 가능할 수 있도록 파싱 모듈을 개발, 구현하였다.

- 파일필터 적용 모듈

MS-Office 문서, Hwp 문서, Pdf 문서 등 서식이 있는 파일로부터 텍스트를 추출하는 프로그램을 가진 파일 필터를 필터링 모듈과 연계할 수 있는 인터페이스 모듈을

개발하였다.

- 개인정보 탐지 모듈

관리자가 결정한 개인정보 패턴을 컨텐츠의 차단 및 치환에 적용할 수 있도록 반영하고, 이 패턴에 대한 탐지를 수행하며 그 정확성을 테스트 할 수 있도록 통계 기능을 가진 모듈을 개발하였다.

- Caching 모듈

실제 서비스에서는 안정성과 더불어 빠르게 응답하고 웹서버의 부하를 줄일 수 있는 형태로 본 제품이 역할을 하는 것이 보다 중요하다. 따라서, 기본적인 서비스 모듈 이외에 웹서버의 부하를 줄이고 더 빠른 서비스가 가능하도록 정적인 파일(이미지 등)에 대한 캐싱 알고리즘을 통해 실제 서비스에서 개인정보 탐지 프로세스로 인한 성능 저하 요인을 보완하고 웹가속 효과를 기대할 수 있도록 개발하였다.

- 콘텐츠 압축 모듈

콘텐츠 압축 기술은 파일에 대한 캐싱과 더불어 웹가속을 이루는 주요한 기능중의 하나이다. 이 기술은 텍스트 컨텐츠에 대해서 gzip 또는 deflate 형식으로 압축하여 전체적인 네트워크 전송량을 줄이는 기술로 컨텐츠에 대한 압축 기술 확보와 HTTP 헤더 정보 변경 등을 다룰 수 있도록 개발하였다.

- 통합 UI 관리툴

웹 UI를 이용하여 스캐너 및 필터에 대한 통합 관리가 가능한 통합 관리 GUI 툴을 개발한다. 구현은 Tomcat jsp 컨테이너를 이용해서 기본 적으로 웹 구성이 가능한 환경을 구성하였으며, 각 모듈간 인터페이스가 원활하도록 하여 전체적인 완성도를 높이도록 구현하였다. 개발된 개인정보보호 시스템은 어플라이언스 형태의 기기로 두 가지 조건으로 구현되었으며 구체적인 사양 및 기능은 아래 표 1과 같다.

표 1. 시스템 기능 및 사양
Table 1. System Function and Specification

Product	1	2
Architecture	Appliance/Package	Appliance/Package
Chassis Height	1U	2U
Redundant AC Power	X	i
Processor Type	Intel	Intel Zeon
Processor Clock Speed	2.4 GHz	3.0 GHz
Number of Processors	1	2
Default Memory	1 GB	1 GB
Memory Upgradeable to	-	2 GB
Hard Drive	160 GB	160 GB
Maximum Throughput	100 Mbps	300Mbps
connections/sec	1,000 cps	4,000 cps
VLAN tagging(802.1q)	지원	지원
syslog	지원	지원
SNMP v1/2/3	지원	지원
NTP(Network Time Protocol)	지원	지원

III. 실험 및 결과

개발된 시스템은 실제 사이트에 적용하여 개인정보 보호를 위해 활용될 수 있다. 따라서, 기본적인 서비스가 정확하게 안정적으로 유지되는지를 판단하는 것이 무엇보다 중요한 단계이다. 이를 위해, 개인정보에 대한 탐지(차단 및 치환)가 정확하게 이루어지는지, 서비스가 안정적으로 유지되는지를 종합적으로 테스트할 필요가 있다. 테스트 방식은 개인정보의 등록 및 기 등록된 데이터에 대한 모의 테스트 시나리오를 통해서 체계적인 검증 단계를 거칠 수 있도록 다양한 형태의 테스트를 진행하였다.

자체 테스트 시스템을 통해 수집된 스캐닝 모듈 및 필터링 모듈을 탐지 통계를 분석하여 표 2에 나타난 바와 같이 탐지 성능을 얻을 수 있었다. 이 결과에 따르면 총 5715건의 검출 패턴에 대하여 29건의 오탐건수를 확인할 수 있으므로 전체적으로 1% 미만의 오탐율을 보였다. 실험 데이터에서는 사업자 번호와 계좌번호의 오탐율이 비교적 높게 나타났다. 이는 사업자 번호의 경우는 검출빈도 자체가 작기 때문이고, 계좌번호의 경우 그 정규식 패턴이 다양한 이유에서 기인한 것으로 보인다. 이메일과 핸드폰 번호와 같이 정규식 패턴이 정형화 된 경우는 오탐율이 극히 낮은 것으로 나타났다.

표 2. 개인정보 탐지정확성
Table 2. The accuracy of Information Detection

패 턴	검출수	검출률	오탐수	오탐율
주민번호 노출 검사	1	0%	0	0%
사업자번호 노출 검사	132	2%	2	2%
법인번호 노출 검사	9	0%	1	11%
신용카드 번호 노출 검사	0	0%	0	0%
핸드폰 번호 노출 검사	3040	53%	18	1%
계좌번호 노출 검사	52	1%	5	10%
이메일 노출 검사	2479	43%	3	0%
사용자정의 번호 노출 검사	2	0%	0	0%
합계	5715	100%	29	0.5%

또한 정보보호시스템을 탑재한 웹 홈페이지의 성능을 알아보기 위하여 Inktomi Traffic Server를 이용한 트래픽 저리 성능실험을 수행하였다. 성능평가 결과 그림 5에서 그림 7까지 나타난 바와 같이 구현된 시스템은 트래픽 기준으로 40~190Mbps인 웹 상황에 적합하며, 세션 기준으로 2000~4000cps인 웹 상황에서, 구성 HTML의 크기가 1~10KB인 웹 상황에 적합한 것으로 나타났다.

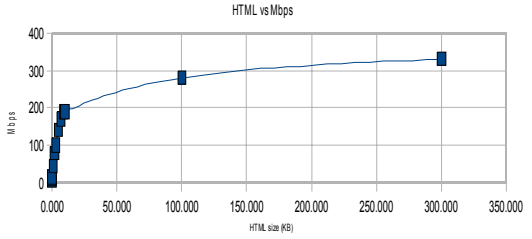

그림 4. HTML 크기와 네트워킹 성능 비교
Fig. 4. HTML size vs. Throughtput

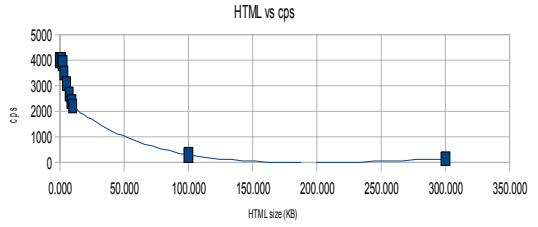

그림 5. HTML 크기와 세션 성능 비교
Fig. 5. HTML size vs. Session performance

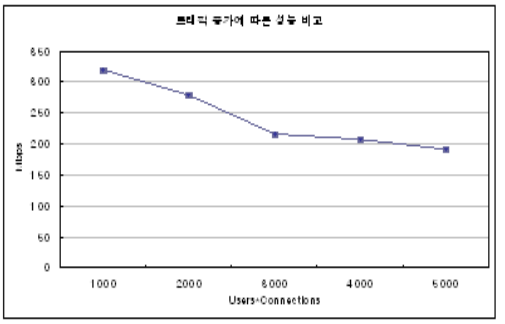

그림 6. 트래픽 증가에 따른 세션 성능 비교
Fig. 6. Traffic size vs. Session performance

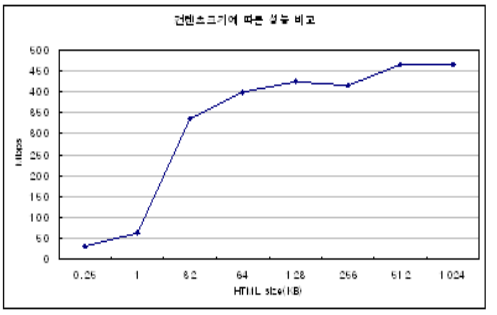

그림 7. 콘텐츠 크기에 따른 세션 성능 비교
Fig. 7. Contents size vs. Session performance

Ⅳ. 결 론

본 논문에서는 인터넷 웹상에서의 개인정보보호를 위한 스캐닝 및 필터링 모듈을 통한 개인정보보호시스템을 구현하였다. 개인정보의 주된 노출은 대국민 서비스 및 대고객 서비스의 접점인 홈페이지 상의 게시판 및 첨부파일을 통해 노출되고 있으며, 운영 중인 게시판과 게시물 첨부파일의 숫자가 워낙 방대하여 관리자의 수작업으로는 기 존재하거나 최신 입력되는 개인정보를 진단하고 차단하는 데는 한계가 있다. 따라서 자동화된 개인정보 노출 진단을 위한 스캐닝 및 필터링 솔루션이 필요하며, 본 논문에서 개발한 솔루션은 이러한 홈페이지 콘텐츠 상의 개인정보를 자동화 된 스캐닝 기능으로 찾아내고 입출력되는 내용에 대한 콘텐츠 필터링을 통해 홈페이지 상에 무분별하게 노출되는 개인정보를 효율적으로 관리하고 통제하기 위한 솔루션이다. 본 논문에서는 검색엔진 기술을 기반으로 파일 캐싱 및 압축 기법을 활용한 웹 가속 기능을 통하여 보다 신속하게 기업 내부의 전반

적인 데이터 및 취약점을 스캐닝 하고 데이터 유출을 차단할 수 있는 시스템을 구현하였다. 기업에서 보호대상 정보의 진단 및 필터링에 소요되는 인력감소 및 생산성 향상, 그리고 만일의 개인정보 유출로 인한 손해배상 비용까지 고려하면 이러한 솔루션의 도입은 기술적인 부분 이외에 경제적인 부분에서도 커다란 가치를 가질 수 있을 것이다.

참 고 문 헌

[1] 김정덕, 개인정보보호를 위한 관리체계와 거버넌스, 정보보호학회지 제18권 제6호, pp. 1-5, 2008. 12.

[2] 남기효, 박상중, 강형석, 남기환, 김성인, 개인정보보호기술의 최신 동향과 향후 전망, 정보보호학회지 제18권 제6호, pp. 11-19, 2008. 12

[3] 조영임, 개인정보보호와 지능형 에이전트 기술, 한국정보기술학회지, 제6권 제1호, pp. 29-35, 2008. 12

[4] 송유진, 이동혁, 남택용, 장종수, 유비쿼터스 환경에서 개인정보보호의 기술동향, 정보보호학회지, 제16권 제3호, pp. 75-86. 2006. 6

[5] 홍승필, 이철수, 유비쿼터스 컴퓨팅 환경내 개인정보보호 프레임워크 적용 방안, 정보보호학회논문지, 제16권 제3호, pp.157-164, 2006. 6

[6] 장은영, 김형종, 시스템 명세화 기법 기반의 개인정보보호 모바일 알람 시스템 설계 및 구현, 정보보호학회논문지 제20권 제1호, pp. 113-121, 2010. 2

[7] 장현미 외 5인, 개인정보보호엔진 설계 방안, 한국인터넷정보학회 학술발표대회 논문집, 제9권 제1호, pp. 167-172, 2008. 5

[8] 최상호, 이은옥, 정미란, 유비쿼터스 환경의 개인정보 보호를 위한 법률, 제도적 방안, 한국정보과학회 한국컴퓨터종합학술대회 논문집(A), pp. 121-123, 2005. 7

[9] 김경진, 홍승필, 개인정보보호 법규에 준한 시스템 대응 방안, 한국인터넷정보학회 학술발표대회 논문집, 제9권 제2호, pp. 125-130, 2008. 11

[10] 강신범 외 4인, 개인정보보호 관점에서의 웹 트래픽 수집 및 분석 서비스에 대한 타당성 연구, 정보보호학회논문지 제19권 제6호, pp. 121-134, 2009. 12

[11] 나희성, 고일석, RwO-캐싱 : 연관 웹 객체 기반의 웹 캐싱 기법 연구, 한국전자거래학회지 제13권 제4호, pp. 161 ~ 171, 2008. 11

※ 본 연구는 한국연구재단의 일반연구자지원사업(2009-0075209) 및 한성대학교 교내연구비 지원으로 수행되었음.

저자 소개

황 호 영(정회원)

• 학위
- 서울대학교 컴퓨터공학 학사
- 서울대학교 컴퓨터공학 석사
- 서울대학교 전기컴퓨터공학 박사

• 경력
- 한성대학교 멀티미디어공학과 부교수

<주관심분야 : 정보통신, 유무선 네트워크, 센서네트워크, 멀티미디어>

김 남 윤(정회원)

• 학위
- 서울대학교 컴퓨터공학 학사
- 서울대학교 컴퓨터공학 석사
- 서울대학교 전기컴퓨터공학 박사

• 경력
- 삼성전자(주) 무선사업부 책임연구원
- 한성대학교 정보시스템공학과 부교수

<주관심분야 : 멀티미디어 통신, 웹 검색, 모바일 통신 및 응용>