

국내 검색엔진을 이용한 개인정보 해킹에 관한 연구

양형규*, 이강호**, 최종호***

A Study on Personal Information Hacking using Domestic Search Engines

Yang Hyung Kyu *, Lee Kang Ho **, Choi Jong Ho ***

요 약

네트워크 기술과 인터넷 인프라의 발전은 빠르고 편리한 정보화 사회를 구축하였다. 그러나 개인정보에 대한 노출 피해와 검색엔진을 이용한 사용자 아이디나 비밀번호, 또는 신용카드 번호 같은 개인정보 해킹 등의 문제점도 발생시켰다. 따라서 각 검색 엔진에서 발생하는 이러한 문제점들을 파악 분석하고, 이러한 문제점을 해결하기 위한 대응책이 요구된다. 본 논문에서는 해외 검색엔진 구글과 국내 검색엔진 네이버, 엠파스를 이용하여 개인정보 해킹을 시도하고 해킹한 결과를 비교 분석한 후, 각 검색엔진에 대한 문제점을 파악하고, 그 피해를 최소화하고 대응 할 수 있는 해결책을 제안한다.

Abstract

Recent advances of network technologies and internet infrastructures construct a fast and useful information-oriented society. However, these may infringe on privacy and expose sensitive information such as user id, secret number and credit card number. Therefore, we need countermeasures for solving these problems. In this paper, we try to hack personal information using Google and domestic search engines, Naver and Empas. After analyze the result, we suggest solutions to prevent personal information hacking based on these search engines.

▶ Keyword : 검색엔진(search engine), 개인정보(personal information), 해킹(hacking)

• 제1저자 : 양형규

• 접수일 : 2007.5.1, 심사일 : 2007.7.12, 심사완료일 : 2007. 7.23.

* 강남대학교 컴퓨터미디어정보공학부 교수 ** 한국재활복지대학 컴퓨터정보보안과 교수

*** 강남대학교 전자시스템정보공학부 교수

1. 서론

검색엔진의 발전으로 정보의 빠른 검색이 가능해졌지만 악의적인 공격자가 개인정보 해킹을 목적으로 검색엔진을 이용하는 것 또한 가능하게 되었다. 해외의 경우 이미 구글을 이용한 사용자 아이디나 비밀번호 또는 신용카드번호 같은 개인정보 해킹 피해들이 보고된 바 있다[1,2,3]. 구글을 이용한 해킹 방법은 많은 자료들이 존재하지만 대부분의 자료들이 해외의 실정에 맞게 제작되어있다. 또한 국내 검색엔진을 이용한 해킹 방법에 대한 자료는 거의 없을뿐더러 카페나 블로그와 같이 국내 인터넷이 가지고 있는 특성들도 있기 때문에 기존의 구글을 이용한 해킹 방법을 그대로 국내에 적용시키기에는 어려움이 있다.

따라서 본 논문은 해외 검색엔진 구글과 국내 검색엔진인 네이버, 엠파스를 이용하여 개인정보 해킹을 시도하고 그 결과를 비교 분석한 뒤, 그 피해를 최소화하기 위한 대응책을 제안하고자 한다.

본 논문은 2장에서 검색엔진을 이용한 개인정보 해킹을 수행하여 그 결과를 검색엔진별로 비교 분석하고 3장에서 그 대응책을 제안하며 4장에서 결론을 맺는다.

II. 관련연구

2.1 개인정보 검색

사용자 아이디, 비밀번호, 웹 디렉토리, 신용카드번호, 주민등록번호, 전화번호 등 총 6개의 항목에 대하여 구글과 네이버, 엠파스를 이용하여 검색하였다(표 1. 참조). 그리고 검색 결과를 평가하기 위해 상위 50위내 검색 결과 중에서 적합 문서의 수를 백분율로 나타내었다[4].

검색 결과를 보면 전체적으로 구글이 가장 높은 검색율을 보인다. 이는 사용자 아이디와 전화번호가 웹 문서에서 쉽게 드러남을 뜻한다(그림 1. 참조). 특히 사용자 아이디와 전화번호 검색에서 구글의 검색율이 국내 검색엔진보다 월등히 높은 것을 보면 구글의 기본적인 웹 문서 검색 성능이 국내 검색엔진보다 더 뛰어남을 알 수 있다.

비밀번호나 신용카드번호의 검색율은 구글도 국내 검색엔진과 크게 다르지 않다. 사용자 아이디나 전화번호 같은

정보와는 달리 비밀번호나 신용카드번호 같은 정보는 대부분 워드프로세서 파일이나 엑셀 파일로 문서화되어 저장된다. 따라서 웹 문서 검색에 잘 드러나지 않지만 구글을 이용하면 몇 개의 검색 결과를 얻을 수 있다. 그림 2는 구글 검색을 통하여 찾은 웹문서로 이름과 사는 지역, 신용카드번호 16자리 중 12자리와 주민등록번호 뒷자리 7자리가 드러나 있다.

보다 쉽게 문서화된 정보를 검색하기 위해서는 웹 디렉토리 검색이나 검색엔진의 문서 영역 검색 옵션을 이용해야 한다. 그림 3은 엠파스에서 문서 영역 검색 옵션을 이용하여 찾은 웹 문서로 이름과 신용카드번호, 카드 만기일 등이 드러나 있다.

웹 디렉토리 검색에서는 세 검색엔진 모두 100%에 가까운 검색율을 보인다(그림 4. 참조). 복잡한 검색어 없이 웹 디렉토리의 특성을 나타내는 간단한 단어만으로 웹 디렉토리를 쉽게 검색할 수 있고 이는 곧 웹 디렉토리가 개인정보 해킹에 있어서 가장 편하고 보편적인 목표가 된다는 것을 뜻한다.

표 1. 구글, 네이버, 엠파스의 개인정보 검색율
Table. 1 Personal information search rate in Google, Naver, and Empas

| 검색항목 | 구글 | 네이버 | 엠파스 |
|---------|------|------|-----|
| 사용자 아이디 | 42% | 4% | 2% |
| 비밀번호 | 12% | 8% | 6% |
| 웹 디렉토리 | 100% | 100% | 89% |
| 신용카드번호 | 12% | 6% | 2% |
| 주민등록번호 | 10% | 2% | 6% |
| 전화번호 | 64% | 30% | 4% |

| 검색ID | 이름 | 원래 | 이메일주소 | 검색율 |
|------|----|----|----------------|-----|
| 1 | 이름 | 이름 | @data | 03 |
| 2 | 이름 | 이름 | @hanmail.net | 05 |
| 3 | 이름 | 이름 | @naver.com | 016 |
| 4 | 이름 | 이름 | @hanmail.net | 031 |
| 5 | 이름 | 이름 | @hwanhob.co.kr | 05 |
| 6 | 이름 | 이름 | @hwanhob.co.kr | 011 |
| 7 | 이름 | 이름 | @hwanhob.co.kr | 051 |
| 8 | 이름 | 이름 | @hwanhob.co.kr | 016 |
| 9 | 이름 | 이름 | @hanmail.net | 017 |
| 10 | 이름 | 이름 | @hwanhob.co.kr | 01 |

그림 1. 구글 웹 문서 1
Fig. 1 Web-Page 1 of Google

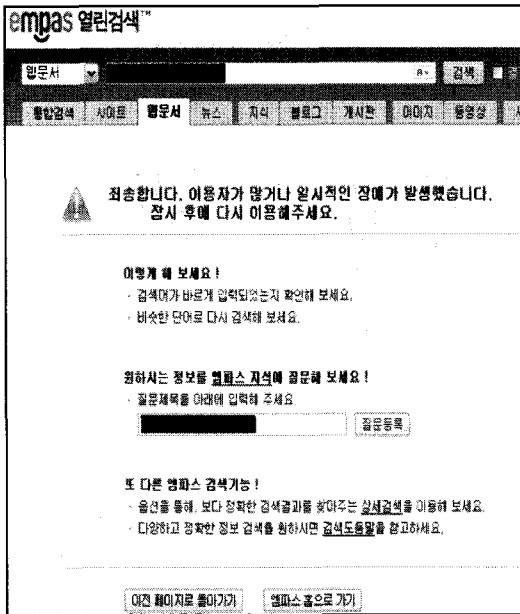


그림 5. 엠파스에서 검색 영역을 'co.kr'로 설정했을 때의 검색 결과
Fig. 5 Searching result in Empas / Area : co. kr

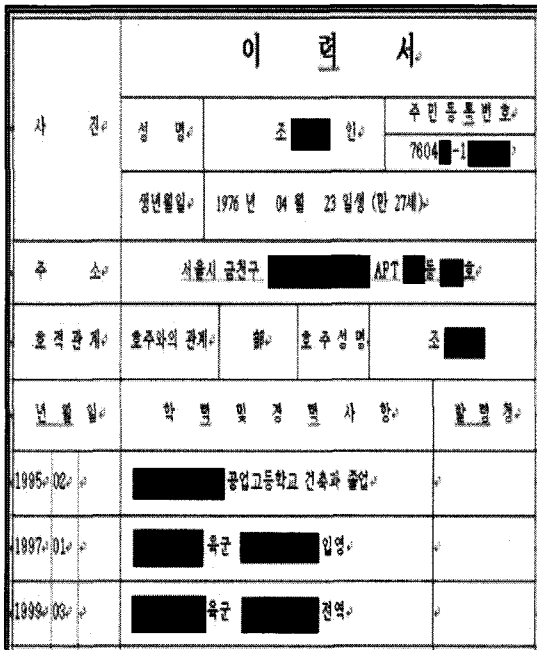


그림 6. 엠파스 검색 문서
Fig 6 Searching document in Empas

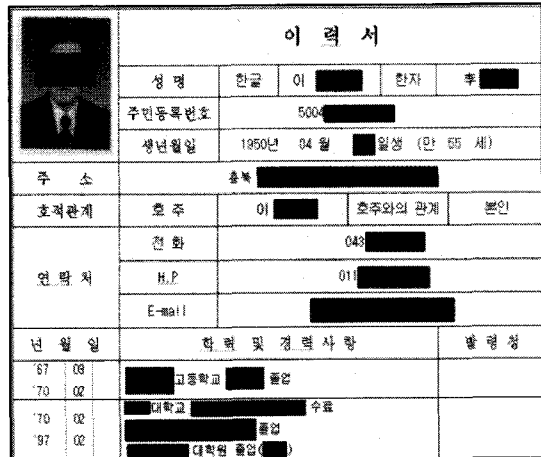


그림 7. 구글 검색 문서
Fig. 7 Searching document in Google



그림 8. 네이버 검색 패스워드 파일
Fig. 8 Password file in Naver

2.3 포털 사이트의 카페와 블로그

한국인터넷진흥원에 따르면, 2006년 말 기준 국내 인터넷 인구수는 약 3,412만 명이며 그 중 2,000만 명에 달하는 이용자들이 네이버, 다음, 네이버 등 매주 한 번 이상 방문하고 있다(6). 2007년 현재 가장 큰 국내 포털 사이트인 네이버의 경우 주간 방문자 수가 23,424,953명에 이른다(그림 9). 이렇게 국내 인터넷 이용자의 절반 이상이 이

용하고 있는 포털 사이트의 카페와 블로그에 저장된 개인정보를 구글과 네이버, 엠파스를 이용하여 검색하였다(표 3. 참조).

검색 결과를 보면 구글과 엠파스의 카페와 블로그에 대한 검색율은 0%이고 네이버만 2%의 검색율을 보인다. 게다가 네이버는 다른 검색엔진의 접근을 차단하고 있기 때문에 엠파스나 구글은 네이버의 블로그와 카페에 접근하지 못한다. 이는 카페나 블로그에서 노출되는 개인정보는 거의 없다는 것을 뜻한다. 그러나 일부 네이버 카페와 블로그의 개인 정보는 네이버를 이용하여 검색할 수 있다. 그림 10은 네이버의 카페 검색을 통해 한 가족 모임 카페에서 찾아낸 것으로 신용카드번호, 계좌번호, 운전면허번호, 통장번호가 드러나 있다. 그림 11은 '신용카드 매출전표 수취 시 부가세매입세액 공제유무'에 대하여 설명해놓은 블로그의 예시 그림 파일이다. 신용카드번호와 신용카드회사명이 그대로 드러나 있다.

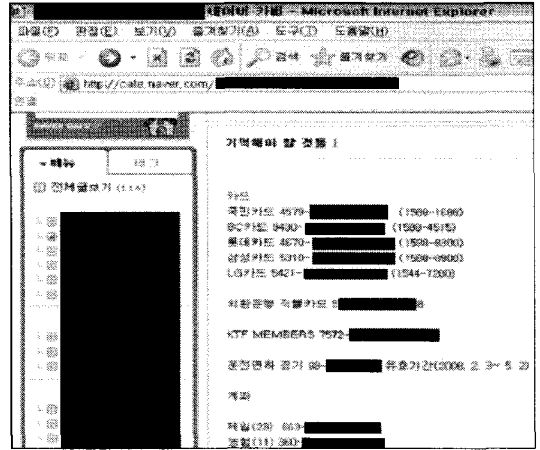


그림 10. 네이버 검색 신용카드번호 및 계좌번호
Fig. 10 Credit card and Account number in Naver

표 3. 구글, 네이버, 엠파스의 블로그, 카페 검색율
Table. 3 Blog and cafe search rate in Google, Naver, and Empas

| 검색엔진 \ 항목 | 네이버 | | 엠파스 | 구글 | |
|-----------|-----|-----|-----|----|-----|
| | 카페 | 블로그 | 블로그 | 카페 | 블로그 |
| 비밀번호 | 2% | 0% | 0% | 0% | 0% |
| 신용카드번호 | 2% | 2% | 0% | 0% | 0% |
| 주민등록번호 | 6% | 0% | 0% | 0% | 0% |

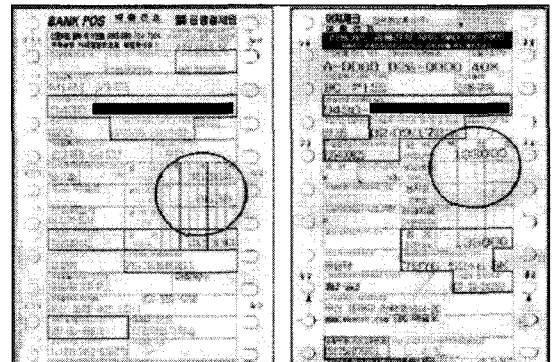


그림 11. 네이버로 검색한 카드전표 그림파일
Fig. 11 Check image file of credit card in Naver

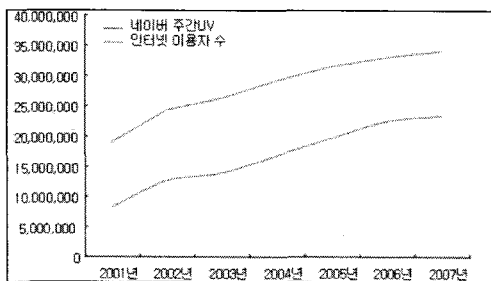


그림 9. 인터넷 이용자 수
Fig. 9 Number of Internet user

III. 검색엔진을 이용한 개인정보 해킹에 대한 대응책

3.1 웹 디렉토리 검색에 대한 대응책

웹 디렉토리는 서버 관리자가 의도하지 않은 비공개 파일을 노출시키므로 위험하다. 구글과 네이버의 웹 디렉토리 검색의 주민등록번호 검색율은 70%가 넘기 때문에(표 4. 참조) 중요한 정보를 다루는 공공기관이나 학교는 웹 디렉토리 사용을 자제해야한다. 따라서 파일 서비스를 제공하는 경우가 아니라면 웹 디렉토리 기능은 사용하지 않는 것이 안전하

다. Apache 서버의 경우 웹 디렉토리 기능을 사용하지 않기 위해서 httpd.conf 파일을 다음과 같이 수정한다[7].

```
Options -Indexes FollowSymLinks MultiViews
```

부득이하게 웹 디렉토리 기능을 사용해야 할 경우, 모든 디렉토리가 인덱스 파일을 포함해야 한다. 또한 서버 관리자는 자신이 관리하는 서버를 대상으로 검색엔진을 이용한 개인정보 해킹을 시도하여, 노출된 정보를 찾아 차단해야 한다 [8]. 웹 디렉토리 기능을 사용하는 경우 검색엔진의 검색로봇의 접근을 막는 것이 필요하다. 다음과 같이 robots.txt 파일을 수정하여 검색로봇의 접근을 차단할 수 있다.

```
User-agent: *
Disallow: /
```

만약 세부 디렉토리별로 검색로봇의 접근 권한을 설정하고 싶다면 검색로봇정책을 참고하여 robots.txt 파일을 수정한다[9]. 그림 12는 robots.txt의 실 사이트 적용 예이다 [10].

웹 서버 사용자가 관리자가 아니라면 메타 태그를 이용하여 검색 로봇의 접근을 막는 방법이 있다(그림 13. 참조).

표 4. 웹 문서 검색과 웹 디렉토리 검색의 개인정보 검색율
Table. 4 Personal information search rate of web-page and web directory

| 검색항목 | 검색엔진 | 구글 | MSN | 엠펙스 |
|-----------|--------|-----|-----|-----|
| 웹 문서 검색 | 비밀번호 | 12% | 8% | 6% |
| | 신용카드번호 | 12% | 6% | 2% |
| | 주민등록번호 | 10% | 2% | 6% |
| 웹 디렉토리 검색 | 비밀번호 | 0% | 0% | 0% |
| | 신용카드번호 | 4% | 6% | 0% |
| | 주민등록번호 | 78% | 80% | 4% |

```
http://www.you.com/robots.txt
# Robots.txt file from http://www.searchengineaid.com
#
# Built from text file http://info.webcrawler.com/mak/projects/robots/active/all.txt
#
# This restricts access to only known and registered robots.
#
User-agent: Mozilla/3.0 (compatible; Miner; mailto:miner@miner.com.br)
Disallow:
User-agent: is_archiver
Disallow: /
User-agent: WebFerret
Disallow:
User-agent: Due to a deficiency in Java it's not currently possible to set the User-agent.
Disallow:
User-agent: no
Disallow:
User-agent: 'Mey! The Homepage Finder'
Disallow:
```

그림 12. robots.txt파일의 실 사이트 적용 예
Fig. 12 Example of real site on robots.txt file

3.2 카페와 블로그 검색에 대한 대응책

```
<HTML>
<HEAD><TITLE>ROBOT META TAG SAMPLE</TITLE>
<META NAME="ROBOTS" CONTENT="NOINDEX,NOFLOW">
</HEAD>
<BODY>
:
</BODY>
</HTML>
```

그림 13. 검색 로봇의 접근을 막는 메타 태그의 적용 예
Fig. 13 Example of Meta-tag protecting access in search robot

블로그나 카페의 사용자는 서버의 기능을 변경할 권한이 없으므로 서비스를 제공 중인 포털 사이트가 개인정보 보안을 위한 정책을 수립해야 한다. 먼저 개인정보가 노출되는 것을 방지하기 위해 글 작성자에게 반드시 검색 허용 여부를 물어야 하며, 허용 여부 문구는 따로 경고 창을 띄우거나 눈에 잘 띄는 곳에 배치해야 한다. 또한 작성된 글에 신용카드번호나 비밀번호와 같은 민감한 키워드가 포함되어 있거나 웹 디렉토리를 사용 중인 카페나 블로그에 대하여 자동적으로 검색로봇의 접근을 막는 정책이 필요하다.

IV. 결론

검색엔진이 해킹의 도구로 이용되는 것이 이제는 낯선 일이 아니다. 구글 해킹이라는 신조어를 만들며 개인정보해

킹의 대명사로 떠오른 구글 못지않게 국내 검색엔진도 충분히 해킹 도구로 이용될 수 있음을 본 논문에서 보였다.

본 논문의 연구 결과 구글뿐만 아니라 네이버나 엠파스를 이용해도 비밀번호나 주민등록번호, 신용카드번호 등의 개인정보를 쉽게 얻을 수 있었다. 개인정보 검색 성능은 구글이 전체적으로 뛰어났으나 카페와 블로그 검색에 있어서는 네이버가 더 뛰어났다. 검색되는 개인정보의 종류에 대해 살펴보면 검색 가능한 개인정보의 대부분이 웹 디렉토리에서 노출되는 문서였다. 이에 대하여 웹 디렉토리 사용 정책 변경과 포털 사이트의 보안 경고 문구 삽입, 특정 키워드에 대한 검색로봇 접근 방지 정책을 통한 대응책을 제시하였다. 이를 통하여 검색엔진 해킹 피해의 사전 예방이 가능하다. 나아가 본 논문이 검색엔진해킹 관련 연구에 도움이 될 것으로 기대된다.

참고문헌

- [1] 김성락, "통합 보안관리 에이전트를 확장한 웹 어플리케이션 공격탐지 연구", 한국컴퓨터정보학회논문지, 제12권 제1호, pp. 161-168, 2007.
- [2] 박대우, 서정만, "TCP/IP 공격에 대한 보안방법 연구", 한국컴퓨터정보학회논문지, 제10권 제5호, pp. 217-225, 2005.
- [3] CNet News.com, Google queries provide stolen credit cards, http://news.com.com/구글+queries+provide+stolen+credit+cards/2100-1029_3-5295661.html
- [4] 이재운, "질문 유형에 따른 인터넷 검색엔진의 성능 비교", 한국정보관리학회 제10회 학술대회 논문집, pp. 185-192, 2003.
- [5] 한국인터넷진흥원, <http://isis.nida.or.kr>
- [6] 네이버 검색 블로그, 네이버 검색 시장 점유율이 78%?, 42%? 왜 다를까, http://blog.naver.com/naver_search/20036714276
- [7] 강유 역저, 구글 해킹, 에이콘 출판사, 2005.
- [8] Lisa Lancor and Robert Workman, "Using google hacking to enhance defense strategies", 38th SIGCSE technical symposium on Computer science education SIGCSE '07, pp. 491-495, 2007.
- [9] robotstxt, robotstxt.org, <http://www.robotstxt.org>

[10] 안랩코코넛, 안랩코코넛 시큐레터,

http://www.coconut.co.kr/board/view.php?id=secutip&no=1283&pg_no=1&searchtype=&searchdata=&view_no=48, 2007.

저자 소개



양형규

1995년 2월 : 성균관대학교 정보공학 박사

1995년 ~ 현재 : 강남대학교 컴퓨터 미디어정보공학부 교수

관심분야 : 정보보호, 네트워크 보안 등



최종호

1987년 2월 : 중앙대학교 전자공학 박사

2002년 ~ 현재 : 강남대학교 전자시스템정보공학부 교수

관심분야 : 정보통신정책, 영상정보통신, 컴퓨터시각, 제스처인식



이강호

1986년 중앙대학교 전자공학과 공학 석사

1991년 중앙대학교 전자공학과 공학 박사

1990년 ~ 2000년 대덕대학 사무자동화과 교수

2000년 ~ 2003년 송호대학 정보산업 계열 교수

2003년 ~ 현재 국립한국재활복지대학 정보보안과 교수

(관심분야 : 정보보안, 디지털 영상처리)