개인정보 문서 노출과 가명정보 조합을 통한 개인정보 관련 피해 위험성 연구

김민주*, 김영은**, 이준민**, 이창현**, 하정희**,
정재완***, 강대명****, 김영철****, 허원석****
*서울여자대학교 정보보호학과
**한국정보기술연구원 Best of the Best
***광운대학교 정보융합학부
****한국정보기술연구원 Best of the Best 멘토
mjkim3715@swu.ac.kr

A Study on the Risk of Personal Information-related Damage through the Exposure of Personal Information Documents and the Combination of pseudonym Information

Min-Ju Kim*, Young-Eun Kim**, Jun-Min Lee**, Chang-Hyun Lee**, Jeong-Hee Ha**
Jae-Wan Jeong***, Dae-Myung Kang****, Yung-Chul Kim****, Won-Seok Heo****
*Dept. of Information Security, Seoul Women's University
**KITRI Best of the Best
***Dept. of Information Convergence, KwangWoon University
****KITRI Best of the Best Mentor

요 약

대부분의 공공기관과 기업에서 개인정보가 포함된 문서를 마스킹 처리하여 온라인상에 게재하고 있다. 이 때, 여러 검색 엔진에서 특정 키워드를 통한 검색 결과를 통해 개인정보가 포함된 문서들이 대량으로 노출되고 있으며 마스킹 처리가 된 정보라 하더라도 2 개 이상의 부가 정보들을 조합해서 개인을 특정할 수 있는 문제가 발생할 수 있다. 이를 통해 얻은 개인정보와 개인을 특정할 수 있는 정보는 다양한 범죄 피해를 발생시킬 우려가 있다. 따라서 본 논문은 검색 엔진과 온라인상에서 노출되고 있는 개인정보가 포함된 문서들을 탐지한다. 그 후 발견된 문서들의 통계와 조사를 통해 온라인상에 노출 중인 개인정보와 가명정보 등이 초래하는 피해의 심각성을 재고하고, 대안을 제시하고자 한다.

1 서론

제 4 차 산업혁명 시대의 도래에 따른 세계 각국의 데이터 경제 활성화 추진과 현대 사회 트렌드를 빠르고 정확하게 예측해 정보를 추출하는 일은 수많은 기업의 과제가 되고 있다 [1]. 이에 따라 텍스트, 이미지, 동영상 등 데이터에서 추출한 정보의 부가가치가 높아지고 있다. 최근 개인정보보호법, 정보통신망 이용촉진 및 정보보호 등에 관한 법률 그리고 신용정보의 이용 및 보호에 관한 법률 (이하 데이터 3 법) 은 데이터를 효율적으로 활용하고 안전하게 이용할 수 있도록 정책을 마련하기 위해 개정이 되었다. 데이터 3 법이 개정되면서 가명정보를 활용하고 이를 정보주체의 동의 없이 적절한 안전 조치 하에 통계, 연구, 기록, 보존 등 다양한 분야에서 활용 가능하게 되었다. 여기서 가명정보의 정의는 ' 가명처리함으로써 원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 정보(이하 가명정보)[2]이다.

일반적으로 개인정보의 유·노출 방지 및 보호를 위해 개인정보의 일부를 특정 문자나 기호로 치환하는 방식인 마스킹 처리 방법을 사용한다[3]. 마스킹 처리란, 주민등록번호의 경우 '200410-4******' 처럼 주민등록번호의 뒤 7 자리의 일부를 '*' 과 같은 특수 기호로 치환하는 방식으로 마스킹 처리할 수 있다. 마스킹 처리는 비교적 쉽게 개인을 식별할 수 없도록 하는 방법 중 하나로 많은 게시판이나 첨부파일에 활용되고 있다.

하지만, 마스킹 처리되지 않고 온라인상에 게시된 개인정보를 포함하는 문서(이하 개인정보문서)가 특정한 검색 키워드를 통해 검색 엔진에서 수집되어 노출되는 경우가 발생하고 있다. 이를 통해 취득한 개인정보 와 관련된 피해가 우려된다.

본 논문에서는 검색 엔진 상에서 개인정보문서와 마스킹 처리를 했음에도 불구하고 개인정보를 유추할 수 있는 문서를 탐지하고 문서들의 정보를 결합하여 개인 특정 가능 여부를 조사하고자 한다.

그 후 탐지된 문서들의 개인정보 노출 여부와 통계를 통해 심각성을 파악하고, 향후 데이터 3 법이 시행되었을 때 개인정보 유·노출 및 가명정보 조합으로 우려되는 피해를 줄일 수 있는 방안을 제시하고자 한다.

2 본론

2.1 현황 조사 및 개인정보 노출 현황

서론에서 언급한 개인정보 문서가 노출되는지 여부를 확인하기 위해 검색 엔진 선정을 진행하였다. 검색 엔진 점유율과 특성을 하단의 <표 1>과 같이 비교하여 대상 검색엔진을 선정하였다.

	.hwp 검색	국내 검색 엔진 점유 순위	고급 검색 지원 여부	검색 결과 마스킹 여부
네이버 (NAVER)	O	1 위	O	O
구글 (Google)	X	2 위	O	X
다음 (Daum)	O	3 위	△	X
줌 (ZUM)	X	4 위	X	X
빙 (Bing)	X	5 위	O	X

<표 1> 검색 엔진 비교[4]

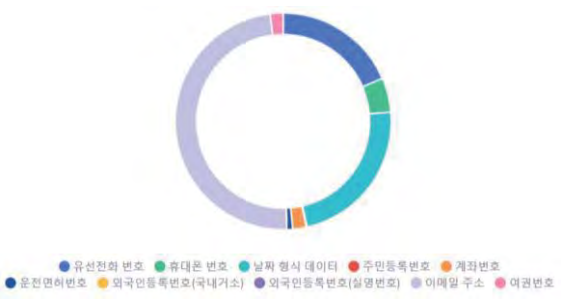
해당 논문에서는 아래와 같은 검색 엔진을 선정했는데, 그 이유는 국내 이용자들이 가장 많이 사용하는 검색 엔진인 네이버와 타 검색 엔진에 비해 방대한 정보를 캐시(Cache) 서버에 저장하는 구글(Google), 관련 직종 종사자와 경력자들의 조언을 바탕으로 기관과 기업에서 정보 유출이 다수 발견되는 빙(Bing) 또한 개인정보의 접근 가능성이 높아 대상으로 삼았다.

(그림 1) 검색 엔진 별 개인정보 노출 문서 통계표

상단의 (그림 1)은 위에서 언급한 3 가지 검색 엔진에 대해 유·노출 데이터를 수집한 결과이며, 네이버 1,868 건, 구글 412 건, 빙 142 건으로 총 2,422 건이 탐지되었다. 특히 엑셀 파일의 경우 단순 '숨기기' 기능이나 글자 색을 바꾸는 등의 처리만 하여 개인정보가 노출되는 경우가 다수 발견되었다.

이들을 대상으로 각 정규표현식을 통해 해당되는 개인 정보가 포함된 문서 및 게시글을 수집하여 개인정보 유·노출 현황 통계 작업을 진행했다. 개인정보종류는 '개인정보보호법' 및 KISA 에서 제공한 '홈페이지 개인정보 노출방지 안내서' 에 작성되어 있는 기준으로 고유식별정보(주민등록번호, 여권번호, 외국인등록번호, 운전면허번호), 개인식별정보(휴대폰 번호, 전화번호, 이메일, 생년월일(이하 날짜형식 데이터)), 신용정보(계좌번호) 등으로 9 가지를 선정하였다[3].

하단의 (그림 2)는 위에서 설명한 방법을 통해 36 시간가량 네이버와 구글, 빙에서 탐지한 문서 내의 개인정보의 총 개수이다.



(그림 2) 개인정보 노출 현황 통계 자료

위에서 탐지한 노출된 개인정보가 담긴 문서에서 아래의 <표 2>와 같이 개인 정보의 건 수를 추출할 수 있었다.

개인정보 종류	발견 건 수
이메일	2,714
날짜 형식 데이터	1,272
전화 번호	1,034
휴대전화 번호	286
계좌 번호	114
여권 번호	106
운전 면허 번호	48
주민 등록 번호	8
외국인 등록 번호	0

<표 2 > 탐지된 문서에서 발견한 개인정보 건 수

이 지표를 통해 마스킹 되지 않은 개인정보들이 상당수 온라인상에 노출되어 있으며, 이 개인정보들이 충분히 악용될 소지가 있다고 판단된다.

2.2 가명정보 조합을 통한 개인 특정

위 기법을 사용해 수집한 결과 가명정보들을 포함하는 문서들을 탐지할 수 있었다. 보통 개인정보가 포함된 문서를 게시하기 전 마스킹 처리를 실시한다. 일부를 가린 개인정보라고 해도 문서 내 함께 존재하는 소속 정보 및 직업 등과 같은 부가 정보와 별도의 문서에서 발견한 마스킹 된 개인정보들을 조합하면 충분히 개인을 특정할 수 있다.

다음은 가명정보를 포함하는 문서들 속에 존재하는 개인정보를 활용하여 개인을 특정한 예시이다.


선발캠프 대상자 명단(초등)

수험번호	이름	생년월일	휴대폰 번호
40	구*송	060312	010-****-5500
10	김*선	070803	010-****-1979
1	김*아	080508	010-****-8593
5	김*연	071026	010-****-1566
41	김*연	060708	010-****-7885
7	김*윤	080623	010-****-8133
27	김*진	071121	010-****-4676
42	김*진	060309	010-****-4676
28	노*빈	070521	010-****-6187
6	박*건	080529	010-****-1439
29	박*성	070603	010-****-4090
7	박*승	080825	010-****-2915
43	박*원	060311	010-****-4684
3	박*호	080510	010-****-7710
30	박*윤	070326	010-****-3416
15	백*영	071028	010-****-6022
8	백*원	081121	010-****-0617

수험번호	이름	생년월일	휴대폰 번호
11	오********스	080428	010-****-8884
44	서*재	060621	010-****-8420
12	성*빈	071031	010-****-1607
20	송*주	070108	010-****-0035
13	오*설	081209	010-****-8522
45	오*홍	060111	010-****-8829
31	윤*우	070702	010-****-7991
46	윤*우	061230	010-****-7719
16	윤*건	081015	010-****-7046
9	이*나	080302	010-****-8392
4	이*선	080521	010-****-8448
18	이*율	050628	010-****-3585
17	이*원	080605	010-****-7207
26	이*지	071231	010-****-2346
32	이*원	070719	010-****-3712
19	이*호	080911	010-****-1137
21	임*빈	081118	010-****-0254

수험번호	이름	생년월일	휴대폰 번호
47	임*우	060927	010-****-3580
14	임*윤	080121	010-****-2688
22	장*우	081130	010-****-1343
48	정*한	060611	010-****-7614
33	정*호	070207	010-****-6796
34	정*환	071016	010-****-6051
36	조*우	071015	010-****-8311
23	조*석	081019	010-****-2588
35	조*훈	070306	010-****-0420
24	진*윤	080415	010-****-3424
49	천*연	060222	010-****-9587
37	최*우	070720	010-****-4056
38	하**날	071018	010-****-5778
25	하*찬	080107	010-****-7079
39	홍*빈	080122	010-****-7220

(그림 3) 가명정보가 포함된 문서

(그림 3)의 경우 특정 기업에서 실시하는 선발 캠프 참가 대상자의 개인정보를 마스킹 처리하여 게시하고 있다. 이 중 한 명의 가명정보를 검색 엔진의 고급검색 기법 이용하여 검색을 진행하였다. 검색 결과는 다음 (그림 4)와 같다.

(그림 4) 고급 검색 기법을 통한 검색 결과

(그림 4)의 결과를 통해 같은 홈페이지에 게시되어 있는 것으로 확인되는 별도의 문서 한 건을 발견했다. 이 문서는 하단의 (그림 5)와 같다.

이름	학년	휴대폰 번호	아이디	비밀번호	이름	학년	휴대폰 번호	아이디	비밀번호
김*아	초4	010-****-8593	5101	S****@lg	이*원	초4	010-****-7207	5119	S****@lg
김*애	초4	010-****-7603	5102	S****@lg	이*호	초4	010-****-1137	5120	S****@lg
김*윤	초4	010-****-8133	5103	S****@lg	이*희	초4	010-****-5747	5121	S****@lg
박*건	초4	010-****-1439	5104	S****@lg	이*희	초4	010-****-1134	5122	S****@lg
박*승	초4	010-****-2915	5105	S****@lg	임*빈	초4	010-****-0254	5123	S****@lg
박*연	초4	010-****-2219	5106	S****@lg	임*윤	초4	010-****-2688	5124	S****@lg

(그림 5) 검색을 통해 얻은 별도의 문서

(그림 3)과 (그림 5)를 비교 분석했을 때 개인의 이름, 생년월일, 나이, LMS 아이디 및 비밀번호 일부를 수집할 수 있게 된다. 이 결과로 한 명을

특정할 수 있다는 것을 확인했다.

마스킹 처리가 된 가명정보라도 2 개 이상의 가명정보가 있거나, 직장, 나이, 소속 등 부가 정보들을 조합하면 개인을 특정할 수 있었다.

3 결론

기업과 기관에서는 개인정보 문서들을 게시하는 경우 검색 엔진 등이 게시판이나 첨부파일의 정보를 수집할 수 없도록 robot.txt 를 올바르게 설정할 필요가 있다. robot.txt 를 활용함으로써 기업이나 기관의 개인정보 문서를 일차적으로 보호할 수 있게 된다.

또한, 검색 엔진의 캐시 서버에 개인정보 문서가 존재하는지, 혹은 검색 엔진 상에 노출되어 있는 개인정보 문서가 없는지 주기적으로 점검할 수 있게 시스템 구축이 요구되며, 이미 노출된 문서들은 삭제 조치가 필요하다.

국내에서 가장 보편적으로 사용하는 검색 엔진인 네이버의 경우 타 검색 엔진에 비해 카페나 블로그의 게시글이나 첨부파일의 형태로 개인정보가 많이 노출될 가능성이 있으므로 관리상 주의가 필요하다.

타 검색 엔진과 달리 네이버는 개인정보가 존재하는 검색 결과를 탐지하여 마스킹 처리 후 검색 결과 페이지에서 보여주고 있다. 구글 등의 타 검색 엔진은 개인정보가 있더라도 검색 결과 페이지에서 마스킹 처리가 되지 않은 정보를 바로 확인할 수 있어 더 위험할 수 있다. 그러므로 국내에서 서비스하고 있는 검색 엔진들이 검색 결과에서 일차적으로 개인정보의 유출을 막을 수 있는 방안이 필요하다. 또한, 각 기관이나 기업, 개인은 개인정보를 온라인상에 게시하기 전 최소한 마스킹 처리를 하고 부가 정보를 삭제하는 것이 바람직하다.

위와 같은 사전 조치 방법을 시행했음에도 불구하고 개인정보의 유· 노출 여부가 탐지된 경우, 이를 탐지한 시점에서 최대한 그 정보와 관련 없는 타인이 악용할 수 없도록 해당 문서나 웹 서비스의 담당자(이하 담당자)는 빠르게 조치해야 할 필요가 있다.

따라서 개인정보 문서나 개인정보를 포함한 게시물이 탐지되면 이를 조치할 권한이 있는 담당자에게 즉시 연락을 통해 조치를 취할 수 있도록 한다.

탐지 결과를 레포팅하는 방식은 본 논문에서 활용한 프로젝트에 등록된 각 기관 담당자의 이메일 주소나, 온라인상에 게시된 담당자의 이메일 주소 등을 통해 개인정보 유· 노출 여부 탐지 시 바로 조치를 취할 수 있도록 탐지 및 통계 결과와 게시된 곳의 링크 등을 첨부하여 이메일을 발송한다.

또한, 웹 사이트의 경우 담당자의 이메일 주소 등의 연락처를 온라인 상에 게시하거나 개인정보 포함 내용을 빠르게 조치하기 위해 자체적으로 신고 게시판 제도를 운영하는 방법 등이 필요하다.

결과적으로 담당자는 연락처로 제공받은 정보와 신고 게시판을 통해 얻은 정보를 활용하여 보다 빠른 조치를 취해 타인이 해당 정보를 악용할 수 없도록 할 수 있다.

마지막으로 데이터 3 법이 개정된 시점에서 가명정보라 해도 부가 정보가 있거나, 두 가지 이상의 가명정보를 포함하는 별도의 문서가 함께 존재할 경우 개인이 특정될 가능성이 있다. 따라서 이와 관련한 법률과 명확한 가이드라인이 필요하다.

참고문헌

[1] 데이터 3 법 개정의 주요 내용과 전망, 2020 KISA REPORT 2 월호, KISA, 2020
[2] 개인정보보호법 제 2 조제 1 항다목
[3] 홈페이지 개인정보 노출방지 안내서, KISA, 2018
[4] http://www.internettrend.co.kr/trendForward.tsp
[5] 2010 인터넷상 개인정보 노출의 문제점 및 대응방안, KISA, 2018
[6] 개인정보 노출예방 교육, 개인정보보호 종합 포털, 2017
[7] 2019 년 개인정보 보호법 위반사례 및 대응방안, 개인정보보호 종합 포털(건양대 차건상 교수), 2019
[8] 홈페이지 개인정보 유노출 위반사례 및 후속조치, KISA, 2019