情 報 保 護 學 會 誌
第23卷 第3號, 2013. 6

# 빅데이터 개인정보 위험 분석 기술

최 대 선*, 김 석 현*, 조 진 만*, 진 승 헌*

요 약

본 논문은 온라인에 공개된 다양한 개인정보의 위험도를 분석하는 기술을 제안한다. 인터넷, SNS에 공개된 다양한 데이터를 수집, 분석하여 개인성향을 파악하고 타겟팅하는 가운데, 분산된 정보를 조합하고 추론하면 공개자의 의도와는 달리 신상이나 민감정보가 노출될 가능성이 크다. 본 논문에서는 이러한 데이터 수집 및 분석을 직접 수행하여 개인정보의 위험도를 분석할 수 있는 기술을 제안한다. 제안 기술이 개발되면, 개인정보 위험도에 따른 클라이언트, 웹사이트, 인터넷 전체 규모의 프라이버시 필터링이 가능해질 것으로 기대된다.

## Ⅰ. 서 론

빅데이터의 소스에 따라 기업이나 정부 등에서 이용자들을 대상으로 직접 수집하거나 자체 생성한 데이터를 분석하는 경우와 인터넷, SNS 상에 공개된 데이터를 수집하여 분석하는 경우로 나눌 수 있다. 전자의 경우 조직 내의 데이터 관리로 현행의 개인정보 보호 규제와 기술의 범주에서 데이터 양만 늘어난 것으로 커버할 수 있다고 본다. 후자의 경우 온라인 상에 공개된 데이터는 누구나 쉽게 수집하고 분석할 수 있다는 점에서 다른 문제가 된다.

이용자 스스로 데이터를 공개하는 것이므로, 개인정보보호는 이용자 책임으로 간주할 수도 있지만, 이용자는 공개 시점에서 개인정보의 위험도를 알 수가 없다. 추후 데이터가 누구에 의해 어떻게 쓰일 수 있는지 짐작할 수 없고, 익명이라고 생각하고 공개한 데이터라도 다른 데이터와 결합하여 신상을 유추할 수 있기 때문이다.

본 논문에서는 공개된 데이터에 포함된 개인정보를 분석하여 노출된 개인정보의 범위와 위험도를 평가하는 기술을 제안한다. 공개된 데이터를 수집하고, 대부분 비정형인 데이터를 분석하여 어떠한 정보가 어디에 얼마나 노출되어 있는지와 신상 유추의 위험, 민감 정보의 노출 여부 등을 분석하여 개인정보 위험도를 산정한다. 개인정보 위험도가 분석되면 이를 기반으로 기 공개된 정보를 삭제하거나 새로운 데이터 공개시 필터링을 할 수 있게 된다.

본 논문의 구성은 다음과 같다. II장에서는 현재 빅데이터 개인정보 보호 관련 현황에 대해 살펴보고, III장에서는 제안하는 빅데이터 개인정보 위험도 분석 기술을 설명한 뒤, IV장에서 결론을 맺는다.

## Ⅱ. 관련 현황

### 2.1 빅데이터 개인정보 노출 현황 및 문제점

한국인터넷진흥원은 국내 트위터 이용자 계정 200여 개를 대상으로 개인정보 노출 현황을 조사하여 상당수의 계정에서 개인정보가 공개되고 있음을 확인했다[1]. 이름(88%), 인맥정보(86%), 사진 등 외모정보(84%), 위치정보(83%), 관심분야 등 취미정보(64%), 스케줄 정보(63%), 가족 정보(52%) 등이 노출되어 있었으며 의료정보(29%), 정치성향 정보(19%) 등 민감 정보가 노출된 비율도 높았다. 페이스북은 이용자의 '댓글'을 분석해 이용자가 동성애자임을 분석해내고 이를 통한 타겟 광고를 한 사례도 있다[2].

[3]은 다양한 매체 및 데이터 분석을 통해 정보 제공자가 의도하지 않은 사생활 침해가 가능하다고 한다. 예

* 한국전자통신연구원 소프트웨어연구부문 사이버보안연구단 인증기술연구실{sunchoi,ksh4uu,zmzo,jinsh}@etri.re.kr

를 들어 트위터, 블로그 등에 따로따로 올린 내용을 통합분석하면 특정인이 혼자 사는지, 언제 집을 비우는지 등을 파악할 수 있다.

이상의 현황에서 도출된 빅데이터 개인정보 문제의 본질은 공개된 비정형 데이터의 추론, 조합에 의해 신원을 포함한 의도하지 않은 정보가 도출될 수 있는데, 공개자는 서비스 별, 시기 별로 분산된 정보 공개 현황을 기억할 수 없고, 따라서 개인정보 공개에 따른 위험을 알지 못하는데 있다.

### 2.2 개인정보 모니터링과 필터링

온라인 상의 개인정보 노출 자체를 방지하기 위한 조치로 개인정보를 모니터링하고 필터링하는 기술이 많이 보급되어 있다. PC 저장 파일에 포함된 개인정보를 탐지하는 솔루션[4]이 있고, 웹사이트 게시물 등을 모니터링하여 노출되는 개인정보를 차단하는 솔루션[5]도 있다. 또한, 인터넷 사이트 전체를 모니터링하여 개인정보 노출현황을 경고하는 시스템[6]도 운영되고 있다.

이러한 기술은 정규식을 이용하여 텍스트에서 주민번호, 계좌번호 등 정형적인 패턴을 갖는 단순한 개인정보 만을 탐지할 수 있다. 빅데이터에서 문제가 되는 추론/조합에 의해 도출되는 정보를 탐지할 수 없다. 또한 어떠한 정보가 탐지되더라도 그 정보가 누구의 것인지 판단할 수 없어, 특정 개인의 개인정보 위험도를 평가하는데 이용할 수 없다.

### 2.3 개인정보 모니터링과 필터링

표 1[7]에서는 빅데이터 분석의 각 단계 별로 발생할 수 있는 개인정보의 침해 가능성에 대응하기 위해 필요한 기술들을 명시하고 있다. 이 기술들은 대부분 데이터를 수집하여 분석하는 주체가 규제를 따라서 이러한 기술을 모두 적용하였을 때 의미를 갖게 된다. 공개된 데이터를 수집하여 프로파일링을 수행하는 경우에 대응하는데 관련있는 기술은 데이터 수집 거부 기술 정도로 볼 수 있다. 정상적 사용자에게 데이터를 제공하지만 로봇같이 자동으로 대량의 데이터를 가져하는 형태를 차단하는 기술[7]로서, 특정 ip로부터의 query가 일정량을 넘어서면 응답을 하지 않는 식으로 동작한다. 그런데 이러한 데이터 수집 거부 기술은 분산된 클라이언트 사용

〔표 1〕 빅데이터 환경의 개인정보보호 필요 기술

| 데이터 처리단계 | 필요기술 |
|---|---|
| 수집단계 | 데이터 수집시 동의 관련 기술 |
| | 데이터 수집시 법률적 위반사항 검토 기술 |
| | 데이터 수집 거부 기술 |
| 저장 및 관리 단계 | 데이터 암호화 기술 |
| | 데이터 접근 통제 기술 |
| | 데이터 필터링 및 등급 분류 기술 |
| 처리 및 분석 단계 | 익명화된 데이터 처리 기술 |
| | 암화화된 데이터 처리 기술 |
| 분석 결과 가시화 및 이용단계 | 이용자 동의와 관련된 기술 |
| | 분석정보의 이용 모니터링 기술 |
| 데이터 폐기 단계 | 데이터 폐기 모니터링 기술 |
| | 분산 환경에서 완전한 데이터 폐기 기술 |

등의 방법으로 충분히 우회가 가능하며, 인터넷에 한번 공개된 정보의 수집을 궁극적으로 차단하는 것은 불가능한 것으로 판단된다.

## Ⅲ. 개인정보 위험 분석 기술

본 장에서는 본 논문에서 제안하는 빅데이터 개인정보 위험 분석 기술에 대해 설명한다.

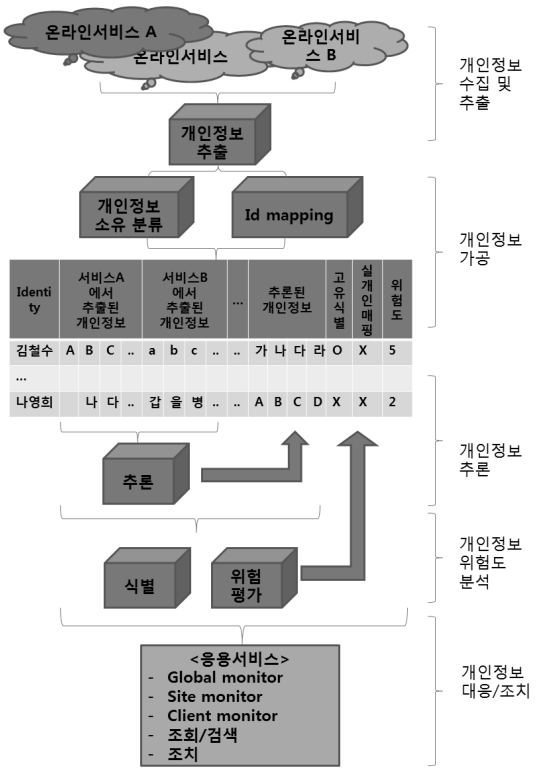
### 3.1 구성 및 작용

빅데이터 개인정보 위험분석 기술은 공개된 데이터를 수집하여 그 속에 포함된 개인정보를 추출한 다음, 그 위험도를 분석하여 대응 조치를 할 수 있게 해주는 기술이다. 그림1은 그 구성을 보여준다.

개인정보 수집 및 추출

- 개인정보 수집 : 통상의 검색 엔진에서 크롤링을 통해 인터넷 상에 공개된 정보를 수집하는 흐름과 동일하며 트위터 같은 SNS 공개 데이터가 추가된다.
- 개인정보 추출 : 수집된 온라인 데이터에서 개인정보를 추출하는 것이다. 기존의 정형적 개인정보에 더해 이름, 거주지, 직업, 나이와 같은 비정형

적 개인정보 추출이 포함된다. 텍스트 데이터에서 이름, 직업 등의 정보를 추출하는 기술은 개체명 인식 기술로, 정보검색이나 자연어 처리 분야에서는 잘 알려진 기술이다. 현재 일반적 개체명 인식 성능은 80% (f1-score)수준으로 알려져 있는데[8] 상기에 언급한 개인정보에 국한하여 인식할 경우 더 높은 성능을 보일 수 있을 것으로 예상된다.




〔그림 1〕 빅데이터 개인정보 위험분석기술 구성

## 개인정보 가공

개인정보 가공은 추출한 개인정보가 어떤 사람의 정보인지 판단하여 분류하고, 개인정보를 정규화 정형화 하는 과정이다.

- 개인정보 소유 분류 : 추출된 개인정보가 누구의 것인지 분류하는 과정이다. 블로그나 SNS 글 중에 나오는 개인정보가 저자의 개인정보일 수도 있고, 다른 사람의 개인정보일수도 있다. 저자의 정보로 분류된 정보는 저자의 id에 귀속시키고, 다른 사람의 소유가 밝혀진 경우는 해당인의 id로 귀속시키게 된다.
- Id mapping : 나의 온라인 서비스에서 추출한 개인정보는 그 온라인 서비스의 id에 귀속시키게 된다. 온라인서비스는 여러 가지이고 한 개인이 여러 온라인 서비스에 각기 다른 id를 보유할 수 있으므로, 각 온라인 서비스에서 획득한 정보를 통합하기 위해서는 서로 다른 온라인서비스에서의 id를 mapping시킬 필요가 있다.
- 정형화 : 추출되어 소유자 분류가 된 개인정보는 정규화된 후 저장된다. 여러 서비스에서 추출된 정보도 id mapping에 의해 하나의 id로 저장된다.

## 개인정보 추론

기 획득된 개인정보를 조합하고 추론하여 추가적인 개인정보를 도출할 수 있다. 추론이 필요한 이유는 개인정보를 수집하여 악용하려는 주체들도 추론을 하기 때문에 개인정보의 위험도를 판단하기 위해서는 마찬가지로 추론을 해야 한다. 가능한 추론의 유형은 다음과 같다.

- 단순 추론 : “누나”, “오빠” 등의 사용하는 표현을 보고 성별을 추론할 수 있다.
- 조합 추론 : A + B= C가 되는 추론이다. 구성 요소 정보는 각기 다른 게시물에 포함되어 있거나 심지어 다른 서비스에서부터 획득한 정보일 수 있다.
- 상호작용 추론 : 본인이 직접 언급한 정보는 아니지만 타인이 개인정보를 공개한 결과를 초래하는 경우이다. 예를 들어 “@kimcs 철수야 생일 축하해”라는 문장을 보면 이 날짜가 kimcs의 생일임을 알 수 있다.

이외에도 다양한 형태의 추론이 가능하며 에 따르면 페이스북의 “좋아요”를 누른 패턴을 바탕으로 이용자의 성별, 인종, 정치성향, 종교, 지수(IQ), 부모의 이혼 여부을 추론할 수 있다고 한다[9].

## 개인정보 위험도 평가

추출 및 추론에 의해 획득된 개인정보를 바탕으로 각 id 마다 공개된 정보에 따른 개인정보의 위험도를 평가할 수 있다. 공개된 개인정보 위험도 평가는 3가지 정도

로 구분할 수 있다.

- Re-identification 여부 : 상기에 언급한 id- mapping 즉, 타 서비스의 account를 특정하여 연결할 수 있다면 이를 re-identification이라고 하며[10] 온라인 상에서는 개인을 특정할 수 있는 것으로 볼 수 있다.
- 실 개인 매핑 여부 : 온라인 상에서 개인을 특정하는 것과 실 자연인을 특정하는 것과는 다르다고 할 수 있다. 한국의 경우 해당인에 대한 주민등록번호를 획득하면 명백히 실개인 매핑이 됐다고 볼 수 있다. 그런데, 주민번호 이외에도, 주소, 나이 등 알려진 정보를 조합하여, 실개인을 특정할 수 있다. 이는 DB 내에서의 동일한 속성 값의 조합이 몇 개 존재하는지를 의미하는 k-anonymity 와는 다른 개념이다.
- 민감 정보 노출 여부 : 개인정보 종류별 위험도 분류[11]에서 높은 등급의 개인정보로 분류된 정보들이 온라인 상에 노출되어 있는지를 의미한다. 민감정보여부는 사이버상이나 실개인 매핑으로 개인이 특정된 상황에서만 의미가 있다.

상기의 요소에 기반하여 종합적인 개인정보 위험도 스코어를 산정할 수 있다.

### 3.2 적용방안

제안하는 기술을 통해 빅데이터 상의 개인정보 위험도를 파악하면 위험도를 경고하고 대응 할 수 있게 된다. 대응 방법은 직접 정보 삭제도 가능하지만 이미 공개된 정보는 삭제해도 다른 곳에 수집되어 있어 효과가 없을 가능성이 크다. 앞으로 잊혀질 권리에 따른 규제 및 관련 기술이 개발되면 보다 실효성있는 삭제가 가능해질 전망이다. 한편, 새로운 개인정보 공개시 (예를 들어 SNS 게시물을 작성시) 더해지는 개인정보에 의한 위험도 증가 가능성을 분석하여 정보공개를 중단하는 방식은 즉시 효과를 거둘 수 있다.

제안 기술의 형태적 적용 방안으로는 한국인터넷진흥원의 개인정보노출 상시 감시 시스템[6]처럼 인터넷 환경 전체를 모니터링하여, 위험도를 파악하고 경고할 수 있는 시스템 구축도 가능하다. 웹사이트 개인정보모니터링 시스템에 포함될 수도 있고, 이용자 클라이언트에 설치되어 이용자가 작성/게시하는 게시물의 개인정보 위험도를 모니터링 해주는 에이전트 형태로 적용될 수도 있다.

## Ⅳ. 결 론

본 논문에서는 빅데이터 분석에 따른 프라이버시 침해 가능성과 이에 대응하기 위해 빅데이터 상의 노출된 개인정보 위험도를 분석하는 기술을 제안하였다. 제안 기술은 한국전자통신연구원에서 개발 중에 있으며 개발이 완료되면 빅데이터 상의 프라이버시 문제를 해결에 크게 기여할 것으로 기대한다

## 참고문헌


[1] “트위터를 통한 개인정보 유형별 노출 현황”, http://kcc.go.kr, 2011. 1

[2] “동성애자까지 알아낸 ‘페북’”, 머니 위크, 2013.5

[3] 정영수, “Big Data 시대의 프라이버시 보호”, NIA Privacy Issues, 제7호, 2012.12

[4] PCFilter, http://privacy.jiran.com

[5] 프라이버시스캐너, http://www.wdigm.co.kr

[6] 개인정보노출 상시 감시 시스템, http://www.kisa.or.kr

[7] 이재식, “빅데이터 환경에서 개인정보보호를 위한 기술”, Internet & Security Focus, pp.79-104, 2013.3

[8] 이창기, 장명길, “Structural SVMs 및 Pegasos 알고리즘을 이용한 한국어 개체명 인식”, 인지과학, 제21권 제4호, pp.665-667, 2010

[9] M. Kosinski, et.al. “Private traits and attributes are predictable from digital records of human behavior”, PNAS, vol 110 no.22, 2013.3

[10] A. Narayanan, V. Shimatikov. “Deanonymizing Social Networks”, 30th IEEE Symposium on Security and Privacy, pp. 173-187, 2009.5

[11] 한국인터넷진흥원, 2002년 개인정보보호백서, 2003.2

## 〈著者紹介〉


**최 대 선 (Daeseon Choi)**
1995년 : 동국대학교 컴퓨터공학과 졸업
1997년 : 포항공과대학교 컴퓨터공학과 석사
2009년 : 한국과학기술원 전산학과 박사
1997년~1999년 : 현대정보기술
1999년~현재 : 한국전자통신연구원 책임연구원
<관심분야> 인증, 개인정보보호, 빅데이터 분석

**김 석 현 (Kim Seok Hyun)**
비회원
2008년 : 충주대학교 전자통신학과 학사 졸업
2010년 : 전남대학교 정보보호협동과정 석사 졸업
2010년~현재 : 한국전자통신연구원 선임연구원
<관심분야> 통신공학, 정보보호, Social Network Security

**조 진 만 (CHO, JIN-MAN)**
1989년 : 충남대학교 계산통계학과 졸업
1991년 : 충남대학교 전자계산학과 석사
1991년~현재 : 한국전자통신연구원 책임연구원
<관심분야> 개인정보보호, 스마트 카드

**진 승 헌 (Seung-Hun Jin)**
종신회원
1993년 : 숭실대학교 전자계산공학과 졸업
1995년 : 숭실대학교 전자계산공학과 석사
2004년 : 충남대학교 컴퓨터과학과 박사
1994년~1996년 : 대우통신
1996년~1999년 : 삼성전자
1999년~현재 : 한국전자통신연구원 인증기술연구실장/책임연구원
<관심분야> 컴퓨터/네트워크 보안, 정보보호(PKI), ID 관리