페이스북과 트위터에 노출된 개인정보 현황
Personal Information Exposed on Facebook and Twitter

최대선(한국전자통신연구원), 안은영(한밭대학교)

차 례

1. 서론

인터넷에는 많은 개인정보가 노출되어 있다. 게시판, 카페, 블로그 등에 자신의 신상정보와 관심분야, 정치성향, 동태정보가 포함된 글을 올리는 이용자가 많다. SNS 이용이 활성화되면서 더욱 많은 이용자들이 더욱 빈번히 자신의 정보를 페이스북이나 트위터같은 SNS에 게시하고 있다. 그런데 이용자들은 자신의 정보가 공개되는 범위를 잘 알지 못하는 경우가 많다. SNS의 경우 친구들만 볼 수 있다고 생각하고 공개한 정보가 전체 이용자들에게 노출되는 경우를 들 수 있다.

주민번호, 계좌번호, 전화번호 등과 같이 그 자체로 개인 식별에 사용될 수 있는 식별정보를 공개할 경우에는 당연히 개인을 식별할 수 있다. 그런데, 이러한 식별 정보 외에도 이름, 거주지, 나이, 직업, 직장, 학교 등의 비식별 정보도 개별적으로 혹은 여러 개의 정보를 조합하면 개인을 특정하고 식별할 수 있는 경우가 많다. 이용자들은 이러한 추가적인 식별 가능성을 잘 알지 못한다. 기존에 어떤 정보를 공개했는지 기억하지 못하기 때문에, 지금 공개하는 정보가 기존에 공개한 정보와 조합이 되어 자신이 식별될 수도 있다는 생각을 하지 못한다.

본 논문에서는 많이 사용되는 SNS인 트위터와 페이스북의 계정들을 조사하여 노출된 개인정보의 종류와 개수를 파악하였고, 노출된 정보를 통한 위험을 분석하였다. 2장에서 조사 방법을 제시하고, 3장에서는 조사 결과를 제시하며 4장에서는 노출된 개인정보로 인해 생기는 문제점 및 대응 방안에 대해 기술한 뒤, 5장에서 결론을 맺는다.

2. 컨텐츠 수집 및 가공

2.1 컨텐츠 수집

개인정보 노출을 분석하기 위해 한국인 이용자의 페이스북과 트위터 계정의 공개 정보를 크롤링하였다. 페이스북은 657만 개, 트위터는 277만 개의 한국인 이용자 계정이 파악되었다(이 계정들이 한국인 이용자 계정의 전부는 아니다). 계정 목록을 획득하는 방법은 페이스북의 경우, 연구자의 친구 중 한국인 추정 이용자 목록을 획득하여 이를 최초 목록으로 하고 획득된 이들의 친구관계들 중 한국인을 목록에 추가하고, 다시 이들의 친구 중 한국인을 추가하는 식으로 계정 목록을 확장했다. 트위터는 연구자의 팔로어 들을 최초 이용자 목록으로 하여 이들과 RT와 멘션을 주고 받은 이용자 중 한국인으로 추정되는 이용자를 획득하고, 다시 이들과 RT/멘션을 주고 받은 사람으로 목록을 확대하는 방식으로 이용자 계정 목록을 획득했다. 이용자가 한국인인지 추정하는 방법은 이름 필드가 한글로 되어 있거나, 영문이름이지만 한글 성을 갖는 경우로 한정했다.

이용자 계정에서 컨텐츠를 획득하는 방법으로, 페이스북은 http://facebook.com/*userid*로 표시되는 타이틀 페이지를 스크랩한다. 이때 타이틀 페이지에 노출된 정보는 이용자가 페이스북 이용자들에게 공개한 정보이다.

트위터는 이용자가 공개한 프로파일 정보와 위치 정보를 획득한다. 프로파일 정보는 프리텍스트로 구성되어 있으며, 위치정보는 이용자가 게시한 트윗에 포함된 위치정보 중 최종 위치를 사용한다.

2.2 데이터 추출 및 가공

페이스북에서 수집된 컨텐츠인 타이틀 페이지에는 여러 가지 개인정보가 필드화되어 포함되어 있다. 필드화된 정보들은 필드 별로 추출하여 획득할 수 있지만, 이렇게 추출된 정보라도 바로 사용할 수는 없다. 이유는 이용자가 자유롭게 입력한 정보이므로, 같은 정보도 다양한 형태로 나타나기 때문이다. 또한 하나의 필드에 여러 가지 데이터가 포함된 경우가 있다. 이러한 점을 고려한 전처리가 필요하다. 특히, 어떤 정보를 공개함으로서 개인이 특정될 수 있는지 알기 위해서는 해당 정보 값을 갖는 개인의 숫자가 중요한데, 이를 위해서는 정보 값이 정규화되어야만 한다. 이름의 경우를 예로 들면, '김길동', '김 길동', 'Kil-dong Kim', 'Gildong Kim'을 모두 '김길동'으로 정규화해야만 정확한 산정이 가능하다. 영어 이름의 경우는 영어-한글 이름 변환기[6]를 사용하여 한글로 변환하였고, 다른 필드의 경우, 별도의 목록을 구축하고 이 목록으로 정규화 하였다.

트위터의 경우는 프로파일 정보가 프리텍스트로 되어 있어 여러 가지 정보가 포함되어 있다. 그 중에서 이름, 나이, 학교, 직업, 직위, 이메일, 전화번호 정보를 추출하였다. 이메일, 전화번호 같은 정형 정보는 정규식[1]을 사용해서 추출이 가능하다. 기존의 개인정보 스캐닝/필터링 제품[2][3]들도, 주민번호와 계좌번호를 포함한 이러한 정형 개인정보를 탐지할 수 있으나 이름, 나이, 학교, 직업, 직위, 등의 비정형 개인정보는 추출이나 탐지가 불가능하다. 본 연구에서는 나이, 학교는 패턴 규칙을 사용해서 추출하였고, 이름, 직업, 직위는 ETRI에서 개발한 개체명 인식기[4]를 사용하여 추출하였다. 개체명 인식기의 경우 미인식과 오인식이 있으므로 이를 이용해 추출한 개인정보에는 오류가 포함될 수 있으며, 모든 노출 정보가 포함되지 않을 수 있다. 트위터 프로파일에는 상기 정보 이외에도 다양한 개인정보가 포함되어 있으나 추출한 정보는 상기 정보에 한정하였다.

3. 개인정보 노출 현황

페이스북 타이틀 페이지에서 필드화된 정보를 추출하여 가공한 결과로 [표 1]은 페이스북 계정에 추출된 개인정보 현황을 보여준다. 이름은 모든 계정에서 필수 공개로 되어 있는 항목이다.

표 1. 페이스북 개인정보 노출 현황

개인정보	노출계정수	비율
이름	6,575,571	100.0%
성별	6,059,339	92.1%
고등학교	3,139,450	47.7%
혈액형	2,686,130	40.9%
대학교	2,335,233	35.5%
직장/직업	1,624,908	24.7%
관심사	1,299,364	19.8%
음악	933,056	14.2%
TV프로그램	574,500	8.7%
영화	558,446	8.5%
책	467,490	7.1%
게임	457,165	7.0%
스포츠 선수	447,492	6.8%
스포츠	365,782	5.6%
스포츠 팀	359,812	5.5%
활동	357,078	5.4%
인용구	294,686	4.5%
직책/직위	256,027	3.9%
대학원	199,508	3.0%
웹사이트	117,819	1.8%
구사 언어	52,938	0.8%
종교관	43,635	0.7%
전화번호	41,900	0.6%
이메일	24,469	0.4%
대화명	22,296	0.3%
정치관	17,548	0.3%
주소	12,834	0.2%

전체의 92%의 계정이 성별을 노출하고, 이어 학교, 혈액형, 직장/직업 등의 신상 정보가 노출된 비율이 높았다. 관심사, 좋아하는 음악 등이 노출된 비율은 10% 이상이었다. 이러한 정보를 공개한 것은 이용자의 선택인데, 모든 이용자가 자신의 정보가 모든 페이스북 이용자에게 노출된다는 사실을 알고 공개한 것이라고 보기는 어렵다. 예를 들어 관심사를 공개한 129만 명 모두가 자신의 관심사를 모든 페이스북 이용자에게 의도적으로 공개했다고 할 수는 없을 것이다. 그 외의 TV, 영화, 게임, 스포츠 등이 노출된 비율은 10% 미만이었다. 직책/직위를 공개한 비율은 직장/직업을 공개한 비율보다 많이 낮았다.

종교관, 정치관 등이 공개된 비율은 1% 미만이었다. 이메일, 전화번호, 주소 같이 종래에 개인정보로 생각되던 식별 정보를 노출한 계정의 비율은 0.5% 미만 이었다. 식별정보가 적게 노출되어 있으므로 SNS의 개인정보 노출은 심각하지 않은 것으로 볼 수도 있으나, 식별정보가 아닌 비식별정보로도 개인을 특정할 수 있음을 다음 장에서 밝힌다.

트위터에서 수집된 정보에서 추출한 개인정보 현황은 [표 2]와 같다.

표 2. 트위터 개인정보 노출 현황

개인정보	노출 계정 수	비율
이름	1,929,407	69.4%
지역	1,252,289	45.2%
직업	933,056	33.7%
학교	558,446	20.2%
직위	457,165	16.5%
나이	92,291	3.3%
전화번호	5,960	0.2%
이메일	2,376	0.1%

트위터에서 이름이 노출된 계정은 69% 이고, 지역(위치)가 노출된 경우는 45%이었다. 직업이 노출된 경우는 33%, 학교 정보는 20%의 계정에서 노출되었다. 트위터에서도 전화번호, 이메일 같은 식별정보가 노출된 경우는 1% 미만이었다. 프로파일 프리텍스트 분석은 미인식된 경우가 많이 있기 때문에 분석기술이 고도화되면 더욱 많은 정보 노출을 확인할 수 있을 것으로 예상된다.

4. 노출된 개인정보 위험 분석

SNS에 노출된 정보에는 전화번호, 이메일 등 직접 식별에 사용될 수 있는 정보도 있다. 이러한 정보들은 종래에도 개인정보로 분류되어 보호 대상으로 하고 있다. 한편, 이러한 식별 가능 정보 이외에도 어떤 개인정보 값은 조사 대상 계정들 가운데서 그 값을 갖는 계정이 한 개만 존재하는 경우가 있다. 예를 들어 페이스북 이용자 중에서 이름의 '김길동'인 계정이 1개만 존재하는 경우다. 이것은 k-anonymity[5]에서 k=1인 경우를 의미한다. 이를 통해 이용자를 특정할 수 있다. 즉 이름이 모두 식별정보는 아니지만, 특정한 이름의 경우는 이름을 통해 식별이 되는 경우가 있는 것이다. 이렇게 어떤 값을 갖는 계정 수가 몇 개인지 정확히 알기 위해서는 각 필드 값을 정규화해야 한다. 본 연구에서는 2장에 기술한대로 페이스북의 필드값을 정규화하였다. 이렇게 개인정보마다 유일 값을 통해 특정할 수 있는 사람 수가 [표 3]에 나타나 있다.

표 3. 페이스북에 공개된 개인정보로 특정되는 사람 수

개인정보	노출 계정 수	유일 값 개수	유일 값 비율	특정 비율
이름	6,575,571	207,027	3.1%	3.1%
전화번호	41,900	41,821	99%	0.6%
이메일	24,469	24,469	100%	0.4%
주소	12,834	8,499	66%	0.1%
별명	22,296	21,469	96%	0.3%
직책/직위	256,027	0	0.0%	0.0%
성별	6,059,339	0	0.0%	0.0%
혈액형	2,686,130	0	0.0%	0.0%
대학교	2,335,233	0	0.0%	0.0%
고등학교	3,139,450	36	0.0%	0.0%

이름의 경우는 207,027개의 유일 값을 가지며, 전체 노출된 이름 중 3.1%가 유일하다. 모든 계정이 이름을 포함하고 있으므로, 이름으로 특정되는 계정의 비율 또한 3.1% 이다. 전화번호와 이메일은 종래에도 식별정보로 분류되어 온 것처럼 유일한 값의 비율이 100%에 육박한다. 하지만 노출된 계정의 수가 적으므로 이를 통해 특정되는 계정의 비율은 1% 미만이다. 직책/직위, 성별, 혈액형은 유일 값을 갖는 경우가 전혀 없었다. 직책명, 성별, 혈액형 등을 유일하게 갖는 사람은 없을 것이기에 이 결과는 당연한 것이다. 한편, 대학교의 경우 국내의 350개의 대학만을 인식하였다. 이때 유일한 값을 갖는 계정은 없었다. 이는 350개의 대학교 중 이를 명시한 이용자가 1명뿐인 학교는 없었다는 의미이다. 고등학교의 경우 36개 학교는 해당 학교를 명시한 이용자가 1명뿐인 경우가 있었다. 이를 통해 36명을 특정할 수 있다는 의미이다. 트위터의 경우 이름을 통해 특정되는 이용자를 조사하였는데 145,040명이 특정되었다. 이름 하나의 필드 값만으로 두 SNS 서비스에서 특정할 수 있는 개인이 35만 명에 달한다.

두 개 이상의 필드를 조합한 값의 경우는 그 값의 조합이 유일한 경우가 훨씬 많다. 이름과 대학교, 고등학교를 조합하여 특정할 수 있는 사람의 수를 조사한 결과가 [표 4]에 표시되어 있다.

표 4. 페이스북에 공개된 개인정보의 조합으로 특정되는 사람 수

개인정보 조합	유일 값 수	특정 비율
이름-고등학교	2,262,410	34.4%
이름-대학교	1,169,170	17.7%
이름-고등학교-대학교	2,975,399	45.2%
고등학교-대학교	109,397	1.6%
고등학교-대학교-혈액형	194,860	2.9%

이름-고등학교 정보 조합을 이용해 특정할 수 있는 사람의 수는 226만 명이었으며 이는 전체 이용자중 34.4%에 달한다. 이름-대학교 조합은 고등학교 보다 특정성이 떨어지는데 이는 특정 대학교를 명시한 동명이인의 수가 특정 고등학교를 명시한 동명이인의 수보다 적다는 걸 의미한다. 이름-고등학교-대학교 조합의 경우 297만 명을 특정할 수 있어 매우 높은 특정성을 보였다. 이름을 제외하고, 고등학교와 대학교만을 가지고 특정할 수 있는 사람은 10만명 수준이다. 대학교 만으로는 0명, 고등학교 단독으로 36명인데 비해 조합을 통한 특정 가능성이 크게 증가함을 확인할 수 있다. 여기에 혈액형을 추가

하면 특정할 수 있는 사람 수는 2배 가까운 19만명으로 증가한다. 혈액형은 4가지 값 중 하나로 그 자체로는 특정할 수 있는 사람이 전혀 없으나, 다른 정보와의 조합으로 특정 확률을 높일 수 있다

5. 결론

본 논문에서는 페이스북 657만 개와 트위터 277만 개의 한국인 이용자 계정의 개인정보 노출 현황을 조사하기 위해 각 서비스의 데이터를 수집하여 분석하였다. 곧바로 식별에 이용될 수 있는 식별정보는 1% 미만이 노출되어 있으나, 기존에 비식별 정보라고 생각되던 정보로 개인을 특정할 수 있는 경우가 3% 이상이고 다른 정보와 조합을 통해 개인을 특정할 수 있는 경우가 최대 45%에 달하였다. 이로서 페이스북과 트위터에 공개된 개인정보의 현황과 정보를 공개한 이용자는 알지 못했던 조합을 통한 특정 가능성과 그 정도를 확인할 수 있었다.

참고문헌

[1] http://en.wikipedia.org/wiki/Regular_expression
[2] 프라이버시스캐너, http://wdigm.com
[3] SafePrivacy, http://www.nicstech.com/new/sub_02_01_03.html
[4] 이창기, 장명길, “Structural SVMs 및 Pegasos 알고리즘을 이용한 한국어 개체명 인식,” 인지과학, 21(4), pp.665-667, 2010년 12월
[5] Latanya sweeney, “k-Anonymity : A Model For Protecting Privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, Oct. 2002.
[6] 김석현, 최대선, 진승헌, “비정형 사용자 이름의 정형화된 한글 이름 변환 방법 연구,” 정보과학회 2013 추계학술발표회 논문집, 2013년 11월

저자소개

● 최 대 선(Daeseon Choi)

▪ 1997년 2월 : 포항공과대학교 컴퓨터공학과 (공학석사)
▪ 2009년 1월 : 한국과학기술원 전산학과(공학박사)
▪ 1997년 1월 ~ 1999년 6월 : 현대전자/정보기술 연구소 선임
▪ 1999년 7월 ~ 현재 : 한국전자통신연구원 책임연구원
<관심분야> : 개인정보보호, 빅데이터 분석, 컨텐츠 관리

● 안 은 영(Eunyoung Ahn)

▪ 1989년 2월 : 동국대학교 전자계산학과 학사
▪ 1991년 2월 : 동국대학교 컴퓨터공학 석사
▪ 2000년 8월 : 동국대학교 컴퓨터공학 박사
▪ 2000년 3월 ~ 2006년 3월 : 백석대학교 정보통신학부 조교수
▪ 2006년 ~ 현재 : 한밭대학교 정보통신.공학과 부교수
<관심분야> : 컴퓨터그래픽스, 가상현실, HCI, 유체 가시화