Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.28, NO.3, Jun. 2018 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2018.28.3.665

온라인 상에 공개된 부분 익명화된 빅데이터의 프라이버시 침해 가능성 분석*

정 강 수,[1†] 박 석,[1] 최 대 선[2‡]
[1]서강대학교, [2]공주대학교

Analysis of Privacy Violation Possibility of Partially Anonymized Big Data*

Kang-soo Jung,[1†] Seog Park,[1] Dae-seon Choi[2‡]
[1]Soagang University, [2]Kongju University

요 약

정보통신의 발전, 특히 무선인터넷 기술과 스마트폰의 보급에 따라 디지털 데이터가 증가하면서, 온라인 빅데이터 개인정보 문제 즉, 개인 민감정보의 온라인 노출과 이로 인한 프라이버시 침해에 대한 우려 역시 높아지고 있다. 본 논문은 포털 서비스를 중심으로 국내 인터넷 환경에 공개된 온라인 빅데이터의 개인정보 침해 현황에 대한 분석을 수행하고 프라이버시 침해 가능성을 평가하기 위한 척도를 제시하였다. 이를 위하여 본 연구팀은 포털 사이트에서 제공하는 서비스 컨텐츠 중 약 5천만건의 사용자 게시글을 수집하여 개인정보에 해당하는 정보를 추출하고, 추출된 개인 정보를 기반으로 각 사용자의 ID가 부분 익명화 되었음에도 개인을 특정할 수 있는 신상 정보가 노출될 수 있음을 확인하였다. 또한 부분 익명화된 ID를 사용하여 서비스간 개인 정보의 연결 가능성과 개인 신상 정보 노출 수준을 반영한 위험도 측정 척도를 제안하였다.

ABSTRACT

With the development of information and communication technology, especially wireless Internet technology and the spread of smart phones, digital data has increased. As a result, privacy issues which concerns about exposure of personal sensitive information are increasing. In this paper, we analyze the privacy vulnerability of online big data in domestic internet environment, especially focusing on portal service, and propose a measure to evaluate the possibility of privacy violation. For this purpose, we collected about 50 million user posts from the potal service contents and extracted the personal information. we find that potal service user can be identified by the extracted personal information even though the user id is partially anonymized. In addition, we proposed a risk measurement evaluation method that reflects the possibility of personal information linkage between service using partial anonymized ID and personal information exposure level.

Keywords: Privacy analysis, Anonymization, Data crawling, Inference attack

I. 서 론

정보통신의 발전, 특히 무선인터넷 기술과 스마트폰의 보급은 디지털 데이터의 폭발적인 증가를 가져왔으며, 대용량 데이터의 신속한 검색 및 저장, 분석 기술이 발달함에 따라 데이터는 '21세기의 원유'

Received(02. 13. 2018), Modified(05. 28. 2018), Accepted(05. 28. 2018)
* 본 연구는 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No. 2017-0-00498, 차분 프라이버시 기반 비식별화 기술 개발)
† 주저자, azure84@naver.com
‡ 교신저자, sunchoi@kongju.ac.kr (Corresponding author)

라고 불리며 새로운 부가가치를 창출하는 자원으로써 각광받고 있다. 기업 및 정부 조직은 데이터를 활용하여 조직의 의사 결정 및 문제 해결에 도움이 되는 정보를 얻고자 노력하고 있으나, 이는 동시에 개인의 프라이버시 침해 가능성을 내포하고 있다.

개인 정보의 프라이버시 침해 가능성이 정보 활용에 있어 심각한 문제로 부상하면서 현실의 프라이버시 침해 수준을 평가하고 이에 대한 대처 방안을 찾는 다양한 연구들이 이루어져 왔다. 한국인터넷진흥원은 국내 트위터 이용자 계정 200여개를 대상으로 개인정보 노출 현황을 조사하여 이름, 인맥정보, 사진 등 외모 정보와 위치정보, 관심 분야 등의 취미정보를 비롯하여 의료정보나 정치성향 등의 민감한 정보가 노출되어 있음을 분석한 바 있다[2]. 또한 [5]은 트위터, 블로그 등에 게시한 내용을 통합분석하면 특정인의 민감 정보를 추론하여 사생활 침해가 가능함을 보인 바 있다. [6, 7]의 연구는 인터넷, SNS 상에 노출된 국내 사용자 정보를 수집하여 개인성향을 파악하고 개인정보의 위험도를 분석할 수 있는 기술을 제안하였다. 특히 [6]의 경우, 페이스북과 트위터 한국인 이용자 계정 934만개를 대상으로 개인정보 노출 현황을 분석하여 이름, 학교같은 비식별 정보를 통해 개인을 특정할 수 있는 경우가 다수 존재함을 확인하였다. [12]의 연구는 국내 대학생을 상대로 페이스북에 공개하는 정보의 범위와 유형, 그리고 이용시간에 대한 분석을 수행하고 정보 공개에 영향을 미치는 요인을 분석하였다. [19]의 연구는 페이스북 계정에서 수집한 미국 대학 전체 대학생의 프로필 데이터를 통해 개인 식별 정보를 숨기려는 노력에도 불구하고 데이터를 통해 개인을 식별할 수 있음을 보였으며, [20]의 연구는 온라인 개인 정보 보호에 대하여 각 웹사이트들의 개인정보보호 수준과 행동 양식에 대한 분석을 분석하였다.

관련 연구에서 나타난 바와 같이 웹상에 공개된 개인정보는 게시자 본인의 동의하에 공개된 합법적인 데이터임에도 서비스 제공자와 게시자 모두의 의도를 넘어서는 개인정보 노출이 발생할 수 있다. 따라서 의도치 않은 프라이버시 침해에 대한 대응 기법의 개발이 절실하나 이를 위해서는 먼저 개인정보 침해 가능성에 대한 심도있는 분석이 선행되어야 한다.

본 연구는 빅데이터 환경의 개인정보 활용과 보호의 조화를 위해 웹 상에 공개된 정보 간 연계를 통한 개인정보 노출 위험 현황을 분석하였다. 세부 연구 항목은 다음과 같다.

(1) 포털 사이트를 중심으로 온라인 상에 공개된 정보를 수집한 뒤, 수집한 정보로부터 개인정보를 추출 (2) 포털 사이트에서 자체적으로 제공하는 개인정보보호 기법인 id의 부분 익명화 적용 시 발생 가능한 개인 식별 가능성 분석(3) 제안 식별 공격에 따른 위험도 분석 지수 개발

본 논문의 공헌점은 다음과 같다.

(1) 국내 인터넷 사용자 대다수가 사용하는 포털 서비스에서 5천만 건에 달하는 개인 정보를 수집하고, 개인 정보를 추출하여 웹 상의 공개된 정보가 지닌 개인 정보 노출 위험도 수준을 분석함

(2) 포털 서비스에서 자체적으로 제공하는 id의 부분 익명화에도 불구하고 해당 id의 유일성 수준과 포털 서비스 내 타 서비스와의 연계를 통해 특정 개인을 식별할 수 있음을 보임

(3) 부분 익명화된 id의 포털 서비스 내 서비스 간 연계 수준 및 사용자가 게시한 글로부터 추출 가능한 개인 정보의 민감도에 기반한 위험도 평가 지수 개발

II. 공개된 정보의 위험 분석

2.1 온라인 상 공개 정보 수집 및 개인 정보 추출

우리는 전체 인터넷을 대상으로 정보를 수집하는 대신, 포털 사이트에서 제공하는 서비스 상에 이용자들이 남긴 게시글을 수집 대상으로 정의하였다. 포털 사이트에서 정보를 수집하기로 결정한 이유는 국내 웹 이용자의 대부분이 포털 사이트에서 제공하는 뉴스, 카페, 블로그, 웹툰, 영화 등의 서비스를 사용하고 있기 때문이며, 각 서비스의 특성에 따라 이용자들이 개인정보를 드러내는 양상도 다르기 때문이다. (예: 관심 분야, 성별, 연령, 전화번호, etc).

또한, 포털 사이트에서 제공하는 서비스들의 종류가 너무 많은 이유로, 포털 사이트에서 제공하는 서비스 중 뉴스, 영화, 웹툰, 그리고 중고 물품의 거래를 목적으로 하는 카페(이하 중고카페)만을 대상으로 게시글을 수집하였다. 수집한 데이터에 대한 세부 사항은 아래의 Table 1과 같다.

데이터 수집은 파이썬 크롤러인 BeautifulSoup를 사용하였으며, 저장은 MongoDB를 사용하였다.

이후 수집한 정보로부터 개인정보에 해당하는 정보를 정의하고, 추출하였다. 추출 방법은 각각 별도의 가공없이 수집한 그대로 개인정보로 사용할 수 있는 단순 추출 정보와 정규표현식 등의 후처리를 통해

Table 1. Crawling data specification

Service name	Collection target	Collection scope	Data volume
News service	Comments from users on the news	Comments on news from 2016.5 to 2017.4	49,809,574
Movie rate service	Comments from users on the movie	Movies rated 1,000 or more from 2014.5 to 2017.4	3,164,724
Webtoon service	Comments from users on the webtoon	Comments on webtoon from 2014.5 to 2017.4	69,710,203
Cafe service	Article from users on the cafe	Each 50,000 posts on the 10 bulletin board	407,607

얻은 필터링에 의한 추출, 그리고 성별 및 연령대 등 명시적으로 드러나 있지 않지만 추론을 통해 확보 가능한 추론을 통한 추출을 사용하였다. 본 연구팀이 추출한 개인 정보는 Table 2와 같다.

Table 2. extracted personal information from crawling data

Service name	Extracted personal information attributes
News service	Partially protected user ID, news category, comment time, comment contents, gender, age
Movie rate service	Nickname, partially protected user ID, movie genre, comment time, movie rate, comment contents, full ID, gender, age
Webtoon service	Nickname, partially protected user ID, webtoon category, comment time, comment contents
Cafe service	Nickname, partially protected user ID, full ID, article, article time, phone number, email, gender

2.2 추출된 개인정보를 통한 개인정보 노출 위험도 분석

추출된 개인정보로부터 개인정보 침해 위험도를 분석하기에 앞서, 본 논문에서 수집 대상으로 삼은 포털 서비스의 개인정보보호 정책에 대해 설명하고자 한다. 수집 대상으로 삼은 포털 서비스는 이용자의 ID의 전위 4자리를 제외한 나머지 부분을 4개의 '*' 문자로 표시함으로써 정확한 ID가 노출되는 것을 방지하는 부분 익명화를 자체적으로 수행하고 있다. 이처럼 포털 서비스 이용자의 고유값인 ID 정보를 부분 보호 처리(Masking)하여 완전한 ID가 드러나지 않게 함으로써, ID 추적에 의해 이용자의 성향이 타인에게 과도하게 드러나는 가능성을 줄임과 동시에 서비스 사용 시 개별 이용자를 구분할 수 있는 정보를 제공한다.

또한 ID 부분 보호 처리는 ID 정보를 드러나지 않게 만드는 것에 더해 서로 다른 ID의 이용자도 동일 ID로 표현되도록 하는데, 이에 따라 표시된 ID가 동일하더라도 정확히 그 이용자를 식별할 수 없게 된다. 예를 들어 carmen74라는 ID를 지닌 이용자와 carmechanic82란ID를 지닌 이용자는 서로 다른 ID의 이용자이지만 포털에서 제공하는 부분 보호 처리 과정을 거치게 되면 carm****라는 동일한 ID로 표현되게 된다. 즉, 포털 서비스의 ID 부분 보호 처리는 둘 이상의 이용자가 동일한 이용자로 인식됨으로써 특정인이 식별될 가능성을 낮게 하는 효과를 갖게 한다.

그러나 포털 서비스에서 제공하는 개인정보보호 정책에도 불구하고 개인 식별과 개인정보 노출의 위험은 존재한다. 실제 ID의 종류에 따라 ID 부분 보호 처리 이후에도 고유하게 식별될 수 있는 정보가 게시물 내에 존재하기 때문이다.

예를 들어, 영화, 웹툰, 카페 서비스에서는 해당 서비스에서 사용하는 닉네임을 통해서 부분 보호 처리된 ID가 고유한 것인지 아닌지 구분할 수 있고 뉴스 서비스에서는 웹페이지 소스에 포함되어 있는 프로필 이미지 URL을 통해 부분 보호 처리된 ID가 고유한 것인지 구분할 수 있다. 본래 영화, 웹툰, 카페 서비스에서는 댓글의 닉네임을 통해 이용자를 식별하고 해당 이용자의 댓글을 연결할 수 있는 기능을 제공하고 있으므로 서비스 내에서 새로이 노출되는 개인정보는 없다. 하지만 뉴스 댓글의 경우 ID 부분 보호 처리를 통해 댓글 목록을 공개하지 않은 이용자

는 댓글 목록을 연결할 수 없었지만, 프로필 이미지 URL을 통해 이용자를 식별하고 댓글들을 연결하는 것이 가능해진다.

또한, 부분 보호 처리된 ID가 각 서비스 내에서 고유한 것으로 판단할 수 있는 경우 이를 이용해서 서비스 간의 ID 연계를 할 수 있다. 각 서비스 내에서 고유하지 않더라도, 중복성 정도를 알 수 있으면 서비스간 ID 연계의 확률을 제고할 수 있다. 또한 연령이나 성별 등에 추론에 있어서도 ID의 중복성 정도가 많은 영향을 준다.

포털 서비스에서 개인을 구분할 수 있는 수단은 닉네임과 부분 보호 처리된 ID이다. Fig.1에서 보듯, 본 연구에서 수집 대상으로 삼은 서비스 중 영화, 웹툰, 카페 서비스는 닉네임을 부분 보호 처리된 ID와 함께 제공하므로 이를 통해서 이용자를 식별할 수 있다.

뉴스 서비스의 경우 닉네임 없이 부분 보호 처리

Movies, webtoons, and cafes which nicknames exist

Nick	ID	Comment	Time
Younee	love****	Just look	05.24.02:05
Kong	love****	I feel like I had forgotten my childhood	05.24.03:17
Jungji	love****	beyond the touch of the original	05.24.04:16
Jinyoung	aunt****	I was happy for a long time.	05.24.04:48
easy	aunt****	It is a masterpiece at all times.	05.24.05:36

There are multiple comments starting with love ***, but it is possible to identify that the nickname is the comment made by three different users

News that does not have a nickname

ID	Comment	Time
love****	It's a policy for an aging society	05.24.02:05
love****	Let's make a school for the elderly	05.24.03:17
love****	We need aging measures	05.24.04:16
aunt****	We need a policy to prepare for a population cliff.	05.24.04:48
aunt****	That's a nice story	05.24.05:36

There is hard to tell if a user using love *** id is 3 different users, or a user using one love *** id left 3 comments

Fig. 1. Difference of ID exposure Possibility depends on nickname existence

된 ID만을 공개한다. 따라서, 뉴스 서비스의 경우, 웹 상에 표현되는 정보로는 개별 이용자를 구별할 수 없으나 개발자 도구를 통해 확인 가능한 HTML 소스 상에 댓글을 단 이용자의 프로필 이미지의 URL이 고유값으로 존재하는 것을 확인할 수 있었다. 따라서 크롤링 단계에서 각 이용자의 프로필 이미지의 URL을 개별 이용자를 식별하는 수단으로 삼아 각 이용자를 구분할 수 있다.

우리는 부분 보호 처리된 ID의 연결 가능성 분석을 위해, 각 서비스별로 부분 보호 처리된 ID의 중복 수준을 분석하였다. Table 3은 부분 보호 처리된 ID를 지니는 이용자 중 동일한 부분 보호 처리 ID를 지니는 이용자의 수(이를 중복지수 k로 표기한다)를 나타낸다. 여기서 중요한 정보는 부분 보호 처리된 ID 임에도 다른 이용자와의 중복 없이 유일하게 존재하는 ID들, 즉 동일한 부분 보호 처리 ID 수를 지닌 이용자의 수가 1인 이용자들이다. 이 ID들은 서비스 내에서 유일하게 존재하는 ID이므로 타 서비스와의 연계에서 높은 확률을 제공한다.

부분 보호 처리된 ID가 유일하게 존재하는 이용자의 경우, 상대적으로 높은 노출 위험에 더하여 포털 서비스 내의 다른 서비스들과의 연계도 가능한 취약점이 존재한다. 포털 사이트에서 제공하는 서비스들은 각 서비스마다 고유 닉네임을 새롭게 지정할 수 있다. 따라서 닉네임으로는 서로 다른 서비스에 존재하는 이용자가 동일 이용자임을 추론하기 어려우나, 부분 보호 처리된 ID가 각 서비스마다 유일하게 존재할 경우, 해당 이용자들을 동일 이용자로 추론할 수 있는 가능성이 높아진다. 우리는 다음 장에서 유일하게 존재하는 부분 보호 처리 ID를 사용한 다른 서비스와의 연계를 통한 개인정보의 확장에 대해 설

Table 3. Number of identical protected ID users

k	News	Movie	Webtoon	Cafe
1	156,623	112,585	133,356	33,131
2	99,626	76,712	115,842	12,586
3	78,036	54,075	98,544	7,047
4	61,776	42,136	85,684	4,924
5	50,005	33,695	72,990	3,480
6	44,100	28,320	64,050	3,108
〉7	1,444,766	797,631	2,241,059	39,726
Total number	1,935,232	1,145,154	2,811,525	104,002

명한다.

2.3 성별, 연령별 통계 자료를 통한 부분 보호 처리된 ID의 중복성 변화

본 연구에서 수집한 뉴스와 영화 서비스는 각 뉴스 및 영화에 댓글 및 평점을 작성한 이용자들의 회원정보를 사용하여 해당 뉴스 및 영화에 댓글을 단 이용자의 성별 및 연령의 통계 정보를 제공한다. 이 통계 정보를 사용하여 각 이용자의 성별 및 연령 정보를 추론할 경우, 위에서 보인 동일한 ID를 지닌 이용자의 분포가 변화할 수 있다.

Fig. 2에서 보이듯이 통계 정보를 사용한 추론에 의해 각 이용자의 개인정보에 추론된 성별 정보를 추가를 할 경우, ID의 중복 수준이 변경됨에 따라 동일한 부분 보호 처리 ID를 지닌 이용자의 분포가 변화한다. 물론 성별 및 연령 정보는 정확한 값이 아닌 추론에 의한 정보이므로 각 이용자를 확정적으로 구분할 수는 없지만, 일정 임계값을 두어 이를 초과하는 경우에 한해 각 이용자를 구분할 수 있는 수단으로 사용할 수 있다.

우리는 성별 및 연령 정보를 추론하기 위하여 추론을 위해 사용한 확률값의 일정 임계값을 정한 뒤, 해당 임계값을 초과하는 경우에만 성별 및 연령 정보를 개인정보로 추가하였다. 성별 및 연령 정보 추가를 위해서는 베이지안 확률[1]을 사용하였다.

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)} \quad (1)$$

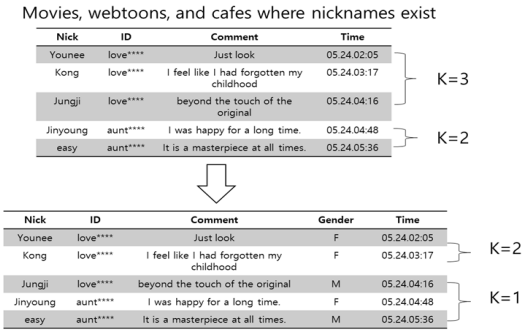

Fig. 2. Duplication degree change by adding gender information

1) 사전확률 p(A)과 우도확률 p(B|A)을 사용하여 사후확률 p(A|B)를 얻는 확률 계산

위 수식에서 사전 확률에 해당하는 P(H)는 전체 뉴스 및 영화에서 작성된 댓글 및 평점을 통한 성별 및 연령 분포이고, 우도에 해당하는 P(D|H)는 각 뉴스 및 영화에서 제공되는 성별 및 연령의 통계 정보이다. 이를 통해 각 이용자의 성별 및 연령대에 해당하는 P(H|D)에 대한 값을 얻을 수 있다. 우리는 5번 이상 댓글 및 평점을 작성한 이용자들의 값이 일정 임계값을 넘는 경우에 대하여 연령 및 성별 정보를 개인정보로 추가하였다. 성별 및 연령대 정보 추론에 의한 뉴스에서의 성별 및 연령 정보 추론에 의한 동일 ID 빈도수 분포 변화는 Table. 4와 같다.

영화의 경우, 뉴스에 비해 통계 정보를 제공하는 영화의 빈도수 자체가 적어 큰 변화는 없으나 뉴스의 경우는 k=1인 부분 보호 처리된 ID의 동일 이용자 수 분포가 크게 변화한 것을 관찰할 수 있다. 이는 곧 성별과 연령대 추가로 인해 서비스 간 연결될 수 있는 가능성이 크게 늘어났음을 의미한다.

Table 4. identical protected ID users number change by adding gender/age information in News service

k	initial number	Adding gender	Adding age
1	156,623	182,645	187311
2	49,963	89,276	90567
3	26,012	44,862	45099
4	15,444	26,295	26355
5	10,001	17,050	17406
6	7,350	12,309	12363
〉7	42,885	70,647	70,853

III. 공개 데이터 연계 위험 분석

앞 장에서 언급한 바와 같이 포털 사이트에서 제공하는 각 서비스들은 id와 닉네임으로 각 이용자를 구분한다. 이 중 닉네임은 서비스 별로 설정을 달리할 수 있으며, 닉네임의 교체도 자유로우나 포털 사이트 가입 시 설정한 id는 변경이 불가능할 뿐 아니라 모든 서비스들에서 공통적으로 사용되는 고유 정보이다. 따라서 id를 통해 포털 사이트에서 제공하는 각 서비스의 이용자들의 정보들을 하나로 통합하는 것이 가능하다.

그러나 이와 같은 연계 가능성은 특정 이용자의 정보가 집적되어 과도하게 많이 노출되는 결과를 가져온다. 즉, id를 사용하여 특정 이용자가 뉴스 서비스에 작성한 댓글과 영화 서비스에 작성한 평점, 중고거래 카페에서 거래 목적으로 작성한 게시글과 이로부터 추출한 개인정보를 연계하여 확장된 개인정보를 구축할 수 있게 되는 것이다. 이를 방지하기 위하여 포털 사이트에서는 이용자의 전위 4자리만을 남겨둔 id를 공개하고 있으나 3장에서 살펴본 바와 같이 부분 보호 처리된 상태에서도 유일한 id들이 존재한다.

본 장에서는 부분 보호 처리에도 불구하고 고유하게 구분되는 id를 사용한 연계 방식을 보인다.

3.1 관계형 연결지수

서비스 연계 방법과 결과를 설명하기에 앞서, 관계형 연결지수이라는 개념에 대해 설명하고자 한다. 전체 서비스에서 부분 보호 처리된 id가 사용되더라도 영화, 웹툰, 카페의 경우는 닉네임에 의해, 뉴스는 이용자 프로필의 고유 url을 통해 부분 보호 처리된 id의 중복 정도를 알 수 있다. 이를 통해 부분 보호 처리된 id가 해당 서비스에서 고유하다면 서비스간의 부분 보호 처리된 id를 연계할 수 있다. 부분 보호 처리된 id가 고유하지 않은 경우 서비스 간 id가 연결될 수 있는 경우의 수가 증가한다. 예를 들어 Fig. 3에서 부분 보호 처리된 id 값이 tige***인 이용자의 경우, 웹툰과 카페에 해당 id가 유일하게 존재하므로 웹툰과 카페 서비스 간 id 연계 시 가능한 조합의 경우의 수는 1가지 경우이다.

그러나 부분 보호 처리된 id값이 서비스 내에서 중복성을 갖는 경우, 서로 다른 서비스 간 연계 시 가능한 경우의 수는 2가지 이상의 경우이다. 예를 들어 Fig. 4에서 부분 보호 처리된 id 값이

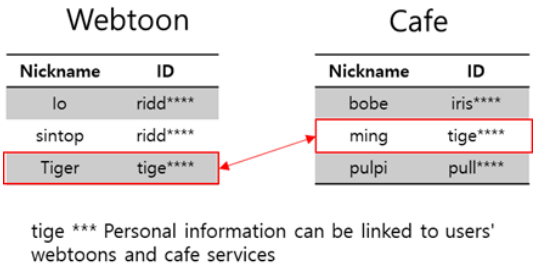

Fig. 3. Linkage between unique protected ID

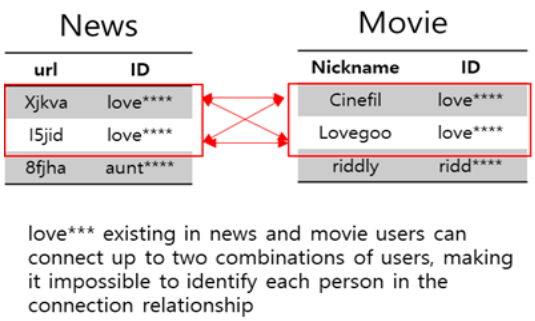

Fig. 4. Linkage among multiple identical protected ID

love****인 이용자의 경우, 뉴스와 영화에 해당 id가 각각 2개씩 존재하므로, 뉴스와 영화 서비스 간 id 연계 시 가능한 조합의 경우의 수는 2가지 경우이다.

이는 각 서비스에서 개별 이용자를 고유하게 구분해줄 수 있는 닉네임이나 url이 각 서비스에 국한되어 사용되기 때문이다. 서로 다른 서비스 간의 연계 시에는 부분 보호된 id가 고유하게 존재하지 않는 경우에는 개별 이용자를 정확하게 연결할 수 있는 확률이 1/가능한 조합의 수가 된다. 우리는 이를 관계형 연결지수(KC) 라고 정의한다

정의 1. 관계형 연결 지수 (K-Connectivity) : 각 서비스에서 부분 보호 처리된 id를 사용하여 연결 가능하고, 이 때 각 서비스간의 연계를 통해 연결될 수 있는 경우의 수이며 다음과 같이 기술될 수 있다.

$$\text{연결 경우의 수} = \prod_{i=1}^{n} u_i \quad (u_i = \text{서비스 I에 존재하는 부분 보호 처리된 id의 중복 수}) \qquad (2)$$

서비스 간 연계 확률은 1/연결 경우의 수이며, 이 경우의 수가 관계형 연결지수에서 KC값이 된다.

3.2 포털 서비스 내 서로 다른 서비스간 연계

우리는 포털 사이트의 뉴스, 영화, 웹툰, 중고 거래 카페에서 수집한 데이터 중 고유하게 구분되는(k값이 1인) 부분 보호 처리된 id만을 대상으로 서비스 간 연계를 수행하였다. k값이 2 이상인 id에 대해서도 연계를 수행할 수 있으나 이는 너무 많은 조

합을 발생시키므로 여기서는 k값이 1인 경우, 즉 KC=1인 경우의 연계만을 분석한다.

3.2.1 연결 단계가 2단계인 경우의 조합

먼저, 2개의 서비스 간의 조합이 이루어진 경우를 보인다. 이 경우의 조합은 각각 뉴스&영화, 뉴스&웹툰, 뉴스&카페, 영화&웹툰, 영화&카페이다.

Table. 5는 각 서비스에서 부분 보호 처리된 id의 중복지수 k=1인 이용자들 중, 해당 id가 서로 다른 서비스간에도 공통으로 존재하는 이용자들의 조합의 수이다. 부분 보호 처리된 id의 k값이 1이고, 해당 id가 두 서비스에 공통으로 나타난다고 하더라도 두 서비스에 존재하는 부분 보호 처리된 id의 이용자가 동일한 이용자일 것이라는 보장은 없다. 예를 들어 이용자 A가 뉴스 서비스만을 사용하고 이용자 B는 영화 서비스만을 사용한다고 할 때, 두 이용자의 부분 보호 처리된 id의 값이 같을 수 있기 때문이다. 그러나 우리는 언급한 경우가 확률적으로 많지 않을 것이라는 전제 하에 서로 다른 서비스에서 동일한 부분 보호 처리 id가 존재할 경우 동일한 이용자로 여기기로 한다.

Table 5. The number and ratio of users with KC = 1 between different services when the linkage step is 2 steps

	news&movie	news&webtoon	news&cafe	movie&webtoon	webtoon&cafe	movie&cafe
user number	31,165	29,193	3,513	24,553	2,013	3,807
ratio to news	0.1989	0.1863	0.0224	-	-	-
ratio to movie	0.2768	-	-	0.218	-	0.0293
ratio to webtoon	-	0.2189	-	0.1841	0.015	-
ratio to cafe	-	-	0.106	-	0.0607	0.1149

Table. 5의 이용자 수 외의 수치는 각 서비스에 존재하는 k=1인 부분 보호 처리된 id 이용자의 수와 2개의 서비스 간 조합으로 존재하는 공통 id를 지닌 이용자 수의 비율이다. 수치에서 알 수 있듯이, 서비스의 조합에 따라 비율상의 차이는 존재하나, 약 20%가량의 이용자들이 2개 이상의 서비스에 공통으로 존재함을 확인할 수 있었다. Table. 6는 연결 단계가 2인 경우에 확장된 개인정보 테이블에서 확보한 속성의 정보량이다.

서비스 간 id 연계가 개인정보 측면에서 갖는 의미는 개인정보의 집적이다. 한쪽 서비스에서 획득할

Table 6. The amount of extended personal information when the linkage step is 2 steps

Service name	Attribute	Data volume
news&movie	news gender	12819
	news age	12797
	movie gender	5
	movie age	16
news&webtoon	news gender	12263
	news age	12238
	wetboon nickcname	23669
news&cafe	news gender	1795
	news age	1794
	phone number	712
	email	344
movie&webtoon	movie gender	2
	movie age	7
	webtoon nickname	18880
webtoon&cafe	phone number	440
	email	234
	webtoon nickname	2015
movie&cafe	phone number	444
	email	875
	movie gender	5
	movie age	0

수 있는 개인정보와 다른 서비스에서 획득할 수 있는 개인정보가 집적되면 그 개인에 대해 더욱 많은 정보가 노출됨을 의미한다. 특히, 한쪽 서비스는 익명성을 위주로 한 서비스이고 다른 한쪽은 실명성을 특징으로 하는 서비스인 경우 예를 들어, 한쪽은 뉴스 댓글이고 다른 쪽은 중고 물품 거래를 위해 실명, 전화번호, 이메일 등을 공개하는 서비스인 경우 연결을 통해 익명성 서비스에서 실명이 드러나는 경우가 발생할 수 있다.

3.2.2 연결 단계가 3단계인 경우의 조합

3개의 서비스간의 조합은 각각 뉴스&영화&웹툰, 뉴스&영화&카페, 뉴스&웹툰&카페, 영화&웹툰&카페이다.

Table 7. The number and ratio of users with KC = 1 between different services when the linkage step is 3 steps

	news& movie& webtoon 웹툰	news& movie& cafe	news& webtoon & cafe	movie& webtoon & cafe
user number	7,482	1,032	658	560
news	0.0477	0.0065	0.0042	-
movie	0.0664	0.0091	-	0.0049
webtoon	0.0561	-	0.0049	0.0041
cafe	-	0.0311	0.0198	0.0169

연결 단계가 3단계인 경우, 2단계에 비해 공통으로 존재하는 이용자의 수가 감소한다. 카페의 경우 상대적으로 수집한 데이터가 적어 카페와 함께 조합되는 경우 공통으로 존재하는 id의 수가 급격히 줄어들게 됨을 확인할 수 있었다.

Table 8. The amount of extended personal information when the linkage step is 3 steps

Service name	Attribute	Data volume
news&movie&webtoon	news gender	575
	news age	574
	movie gender	0
	movie age	1
	phone number	193
	email	102
news&movie&cafe	news gender	3160
	news age	3155
	movie gender	2
	movie age	3
	webtoon nickname	6059
news&webtoon&cafe	news gender	364
	news age	364
	phone number	145
	email	77
	webtoon nickname	651
movie&webtoon&cafe	movie gender	0
	movie age	0
	phone number	110
	email	68
	webtoon nickname	578

3.2.3 연결 단계가 4단계인 경우의 조합

4개의 서비스간의 조합은 각각 뉴스&영화&웹툰&카페의 한 경우이다.

Table 9. The number and ratio of users with KC = 1 between different services when the linkage step is 4 steps

	news&movie&webtoon&cafe
user number	197
news	0.001258
movie	0.00175
webtoon	0.001477
cafe	0.005946

Table. 10은 연결 단계가 4단계인 경우의 이용자 중 한 명의 정보에 대한 예시이다.

각 항목에 대한 설명은 다음과 같다.

- id: 서로 다른 서비스를 연계하는 역할을 수행하는 부분 보호 처리된 id
- news_cate: 해당 이용자가 가장 많은 댓글을

Table 10. Extended personal information with 4 steps of linkage

attribute	personal information	value
id	partially protected ID	aszc****
news_cate	news cagegory	defense/diplomacy
news_age	inferred news age	null
news_gender	inferred news gender	null
cafe_id	cafe full id	[redacted][3]
cafe_cate	cafe category	woman clothes
cafe_gender	cafe gender	woman
cafe_phone	phone number	[redacted]
cafe_email	email	[redacted]
cafe_nick	cafe nickname	[redacted]
movie_id	movie nickname	[redacted]
movie_cate	movie genre	drama
movie_age	inferred movie age	null
movie_gender	inferred movie gender	null
webtoon_id	webtoon ID	null
webtoon_cate	webtoon category	drama

작성한 뉴스의 카테고리

- news_age: 뉴스 연령대 통계로부터 추론한 연령대 정보로, 해당 이용자는 성별 정보를 추론하지 못 했음
- news_gender: 뉴스 성별 통계로부터 추론한 성별 정보로, 해당 이용자는 연령대 정보를 추론하지 못 했음
- cafe_id: 카페에서 크롤링을 통해 추출한 포털 사이트 전체 id로 해당 이용자는 [redacted][2]을 전체 id 정보로 지님
- cafe_cate: 중고거래 카페 내에서 가장 많은 게시글을 작성한 게시판 카테고리로 해당 이용자는 여성상의 게시판에 가장 많은 게시글을 작성하였음
- cafe_gender: 카페 내의 성별이 분화된 게시판 정보로부터 추론한 성별 정보로 해당 이용자는 여성 관련 카테고리에 가장 많은 게시글을 올렸으므로, 여성으로 성별을 추론함
- cafe_phone: 거래를 목적으로 게시한 전화번호 추출 내역으로 해당 이용자의 전화번호는 [redacted]
- cafe_email: 거래를 목적으로 게시한 이메일 주소 추출 내역으로 해당 이용자의 이메일 주소는 [redacted]임
- cafe_nick: 카페 상에서 사용하는 닉네임으로 해당 이용자의 닉네임은 [redacted]임
- movie_id: 영화 서비스 상에서 사용하는 닉네임으로 해당 이용자의 닉네임은 [redacted]임. 포털 서비스에서 수정을 하지 않는다면 닉네임은 이용자 본인의 실명으로 초기 설정되어 있으므로 해당 닉네임은 [redacted] 이용자의 실명일 가능성이 높음
- movie_cate: 영화 내에서 가장 많은 평점을 작성한 카테고리로 해당 이용자는 드라마 카테고리의 영화에 평점을 가장 많이 남겼음
- movie_age: 영화 연령대 통계로부터 추론한 연령대 정보로, 해당 이용자는 연령대 정보를 추론하지 못 했음
- movie_gender: 영화 성별 통계로부터 추론한 성별 정보로, 해당 이용자는 성별 정보를 추론하지 못 했음
- webtoon_id: 웹툰 서비스 상에서 사용하는 닉네임으로 해당 이용자는 닉네임을 등록하지 않아 닉네임을 추출하지 못 했음
- webtoon_cate: 웹툰 내에서 가장 많은 댓글을 작성한 카테고리로 해당 이용자는 드라마 카테고리의 영화에 가장 많은 댓글을 작성하였음

위의 표에서 알 수 있듯이, 공개된 정보로부터 추출한 개인정보에 더하여, 서로 다른 서비스간 부분보호 처리된 id를 사용함으로써 확장된 개인정보를 구축할 수 있다. 또한 이렇게 확장된 개인정보는 보다 높은 수준의 식별 위험과 개인정보 침해 위험을 가지게 된다.

2) 개인정보보호를 위해 마스킹 처리하였음

3) 개인정보보호를 위해 마스킹 처리하였음

IV. 위험도 분석

이전 장에서 우리는 개인정보에 해당하는 값을 추출한 결과와 서로 다른 서비스 간 연계를 통한 개인정보의 확장에 대해 설명하였다. 수집 및 추출한 개인정보들은 수집 수단과 목적, 정보 자체가 지닌 민감 수준에 따라 서로 다른 위험도를 지닌다. 본 장에서 우리는 개인정보를 연결 수준과 민감 수준에 따라 평가하는 방법을 제안하고, 그 평가 결과를 보인다.

4.1 차등화 된 위험도 분석

포털 사이트에서 제공하는 뉴스, 영화, 웹툰, 카페에서 수집한 정보들은 추출한 개인정보의 속성에 따라 서로 다른 프라이버시 침해 수준을 지닌다. 그러나 현재 프라이버시 침해 수준을 평가할 수 있는 객관적 기준이 수립되어 있지 않다. 우리는 수집/추출/추론한 정보들을 각 속성의 특성에 따라 식별 속성과 민감 속성으로 분류하였다.

정의 2. 식별 속성 : 다른 정보와의 연결을 통해 확장된 개인정보를 추출할 가능성이 있는 정보
정의 3. 민감 속성 : 그 자체로 프라이버시 침해를 야기할 수 있는 정보

Table 11. Identification and sensitive attributes of collected/extracted/inferred personal information

category	attribute
Identification Attribute	• Partially protected ID • Full ID • Nickname • Phone number • Email • Gender • Age
Sensitive Attribute	• User preference information • Phone number • Email • Residential Area • Gender • Age

우리는 수집/추출/추론한 개인정보 속성들을 각각 다음과 같이 식별 속성과 민감 속성으로 분류하였다.

우리는 차등화 된 위험도를 분석하기 위하여 식별 정보에 기반 한 연결 가능성과 민감 속성에 기반 한 정보 민감성, 그리고 두 요소를 함께 고려한 위험도의 3가지 측정 기준을 세웠다.

4.2 연결 가능성

앞서 설명한 바와 같이 연결 가능성은 식별 속성에 기반하여 평가된다. 연결 가능성에 대해 설명하기에 앞서 식별 정보의 연결 정확도에 대해 설명한다. 식별 정보는 연결 정확도에 따라 전체 ID, 전화번호, 이메일 등 정확히 이용자간의 연결을 이룰 수 있는 정보와 일정 수준의 부정확성을 담보하는 부분 보호 처리된 ID를 사용한 연결로 분류된다. 본 연구에서는 일정 수준의 부정확성을 전제하는 연결에 대해 다룬다.

부분 보호 처리된 ID를 사용한 연결은 다시 부분 보호 처리에 의해 중복으로 존재하는 이용자의 수에 따라 1:1 연결과 n:n 연결로 분류할 수 있다. 1:1 연결은 k=1인 부분 보호 처리 ID간의 연결을 의미하고 n:n 연결은 서로 다른 서비스 간에 적어도 어느 한 쪽이 k=2 이상의 부분 보호 처리 ID를 지닌 채 연결되는 것을 의미한다. 예를 들어, carm**** 라는 ID가 뉴스에서 2명, 영화에 3명 등장한다면 carm**** ID는 2:3 연결을 이루며 총 6가지 연결 가능한 경우의 수를 지니게 된다.

우리는 식별 정보에 기반 한 연결 가능성을 측정하기 위하여 다음과 같은 식을 제안한다.

$$\text{연결 가능성} = f(\text{연결 단계, 엔트로피}) = \text{Max}(\sum_i(\alpha\cdot\text{연결단계}_i/\beta\cdot\text{엔트로피}_i)),\ (\alpha,\ \beta\text{는 임의의 실수 }) \quad (3)$$

- 연결 단계: 해당 이용자가 다른 서비스와 연결되는 정도를 나타낸다. 예를 들어 carm****이라는 부분 보호 처리된 ID를 지니는 이용자가 뉴스와 영화에 2번 나타난다면 해당 ID의 연결 단계는 2가 된다.
- 엔트로피: 서비스간 연결을 통해 결정된 이용자 연결의 경우에 수에 근거하여 계산된다. 엔트로피 계산 수식은 아래와 같다.

$$Entropy = -p_i(x) \cdot \log p_i(x) \quad (4)$$

($p(x)$= 1/해당 부분 보호 처리 ID가 지닌 연결 가능한 경우의 수)

엔트로피 계산에서 1:1 연결의 경우 엔트로피의 값은 존재할 수 없으므로, 1:1 연결에는 가장 작은 엔트로피 값보다 작은 값을 지니는 임의의 값을 설정한다.

4.3 정보 민감도

연결 가능성이 프라이버시 침해의 확장 가능성을 보여주는 지표라고 한다면, 실제 프라이버시 침해 수준에 대한 측정은 민감 속성과 민감 속성을 바탕으로 도출된 정보 민감성에 의해 이루어질 수 있다.

우리는 정보 민감성을 민감 속성에 대한 등급과 민감 속성의 다양성을 나타내는 l-diversity의 2가지 요소로 분류하여 측정하였다. 또한 다양한 속성들이 민감 속성이 될 수 있으나, 모든 속성들은 정보의 쓰임새에 따라 다양한 수준의 민감도를 지닌다. 우리는 민감 속성을 등급화하여 수치로 나타내었다. 민감 속성 등급은 Table. 12와 같다.

- 3등급: 이용자의 신원을 정확히 특정할 수 있으며, 웹 상의 정보를 사용한 스토킹이나 스팸, 사이버 괴롭힘 등에 사용될 수 있음
- 2등급: 이용자를 특정할 수 없으나 해당 이용자의 정치, 문화적 성향 등의 개인정보를 추정할 수 있음
- 1등급: 이용자 특정에 보조적인 역할을 하며 개인정보가 거의 포함되지 않은 정보

Table. 12의 민감 속성 등급표는 포털 서비스의 뉴스, 영화, 웹툰, 카페 서비스에서 추출 가능한 개인정보 속성을 기준으로 작성하였다. 다른 분야에 해당 도메인의 전문가의 판단에 의거하여 민감 속성 등급표를 작성할 수 있을 것이며, 이에 따라 민감도를 산정하는데 사용 할 수 있다.

Table 12. Sensitivity Rating Tables

Rank	attribute
3	Phone number, email, full ID
2	User preference information
1	Gender, age partial protected ID, nickname

우리는 정보 민감도를 측정하기 위해 다음과 같은 정보 민감도 측정식을 제안한다.

$$\text{정보 민감도} = f(\text{민감 등급, l-다양성}) = \sum_j(\alpha \cdot \sum \text{연결단계}_i / \beta \cdot \sum \text{l-다양성}_i)$$

(α, β는 임의의 실수, i,j $\in$ n, j=연결된 서비스 수 n=이용자가 지닌 항목의 수)

- $\sum$민감 등급: 각 이용자가 지닌 개인정보 속성들의 민감 등급의 총합을 나타낸다.
- $\sum$l-다양성: 각 이용자가 지닌 민감 항목들의 l-diversity의 총합이다.

l-다양성은 k-익명화 시 발생 가능한 동질성 공격(같은 그룹에 속한 데이터의 민감 속성이 전부 동일한 경우일 때, 익명화에도 불구하고 민감 속성이 추론될 수 있는 공격)을 방지하기 위해 그룹 내의 민감 속성이 최소 l개의 다양한 민감 정보를 가지고 있어야 한다는 조건을 의미한다. 본 연구에서는 l-다양성이 적용되는 민감 속성을 각 서비스별 이용자 성향으로 제한하여 평가하였다. l-다양성을 측정하는 수단으로는 entropy l-다양성을 사용하였다. entropy l-다양성은 그룹 내에 존재하는 민감 속성의 가짓수를 분모로 하여 entropy를 계산하는 방식이다.

4.4 정보 위험도

정보 위험도는 연결 가능성과 정보 민감도를 함께 고려하여 해당 이용자의 개인정보 노출 위험도를 최종적으로 수치화한 개념이다. 정보 위험도 측정식은 다음과 같다.

$$\text{정보 위험도} = f(\text{연결 가능성, 정보 민감성}) = \alpha \cdot \text{연결 가능성} \times \beta \cdot \text{정보 민감성}$$

(α, β는 임의의 실수)

정보 위험도 분석은 개별 서비스의 정보 위험도와 k=1인 경우의 조합, 그리고 k=2 이상인 경우의 조합으로 나누어 분석하였다.

우리는 공개 데이터로부터 정보 위험도를 계산하여 연결 단계에 따른 각 조합에 따른 위험도의 평균

Table 13. Information Risk Average according to each combination

	News	Movie	Webtoon	Cafe
connection 1	13.93	17.32	12.13	77.79
Connection 3	News&Movie&Webtoon	News&Movie&Cafe	News&Webtoon&Cafe	Movie&Webtoon&Cafe
	88.7	160.71	185.45	172.39
connection 4	News&Movie&Webtoon&Cafe			
	217.1			

값을 측정하였다.표 13에서 보이듯 서비스간의 연결이 늘어날수록 위험도의 평균값이 증가하였다. 이는 연결된 서비스가 늘어날수록 개인정보의 집적 수준이 높아져 개인정보 침해 가능성이 증가한다는 직관을 뒷받침한다. 또한 표 10에서 제시한 사용자의 경우 1,944라는 평균치를 상회하는 위험도 값을 보인다. 반면 연결 단계가 1단계에 머무르는 사용자들은 대부분 낮은 위험도를 지닐뿐 아니라, 해당 사용자로부터 본인을 특정하거나 본인의 개인정보로 여겨지는 정보를 얻어낼 수 없음을 확인하였다. 이는 제안 위험도 측정 기준이 포털 서비스 환경에서 사용자의 프라이버시 침해 위험도를 평가하는 하나의 측정 기준으로 적용될 수 있음을 의미한다.

V. 위험도 분석 프로토타입

우리는 제안한 위험도 평가에 기반하여 이용자가 자신의 위험도를 시각적으로 판단할 수 있는 위험도 평가 시스템의 프로토타입을 구현하였다.

Fig. 5는 구현한 위험도 프로토타입의 스크린샷이다. 스크린샷의 각 부분에 대한 설명은 다음과 같다.

- User name: 조합에 존재하는 이용자들이 목록
- 조합: 각 서비스 별 가능한 조합의 경우의 수
- 위험도 분포표: 조합에 존재하는 이용자들의 위험도 분포표 및 선택된 이용자의 위험도 위치
- 이용자의 위험도: 선택한 이용자의 위험도 값
- 조합의 위험도 평균: 현재 조합에 존재하는 이용자들의 위험도 평균 값

해당 프로토타입은 각 서비스 및 조합에서의 이용자 ID를 선택하면 해당 조합에서의 전체 이용자의 위험도 분포와 분포에서의 선택된 이용자의 위험도 위치를 붉은 점으로 화면에 표시한다. 또한 하단에

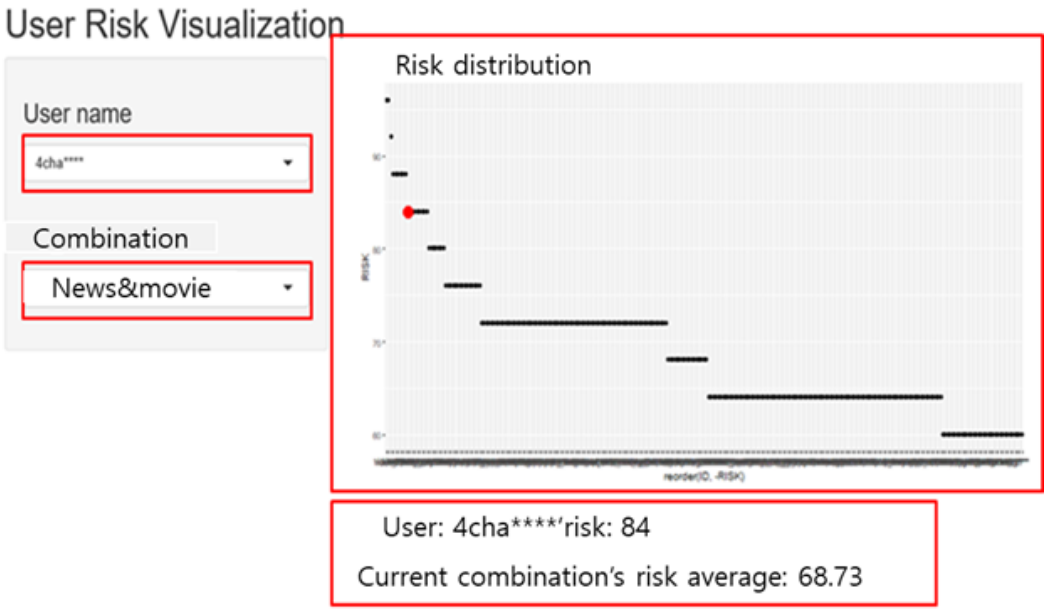

Fig. 5. Prototype of risk assessment system

해당 이용자의 위험도와 현재 조합의 위험도 평균값과 선택한 이용자의 위험도를 텍스트로 나타낸다. 이용자가 제안하는 프로토타입과 같은 위험도 평가 시스템을 사용할 수 있다면, 자신의 동의 하에 공개된 정보라 할지라도 해당 정보가 지니는 위험도를 직관적으로 이해하여 자신의 정보 공개 수준을 스스로 조절할 수 있는 기준으로 삼을 수 있을 것이다.

VI. 결 론

본 논문은 통신 및 모바일 기술의 발달에 따른 웹&앱 서비스 활성화에 따른 웹상에 공개된 디지털 데이터 증가 현상에 주목하여, 각 개인이 서비스를 사용하기 위하여 자발적으로 정보를 게시한 공개 정보의 개인정보 노출 위험도를 분석하고자 하였다. 이와 같은 공개 정보는 서비스를 사용하는 이용자는 누구든 열람/수집 할 수 있다는 점에서 서비스 제공자의 서버에 저장된 개인정보와는 차이점을 지닌다.

이러한 개인정보의 샘플로서 본 연구팀은 국내 환경에서 3,000만명이 넘는 이용자를 확보하고 있는 포털 서비스와 그 포털 서비스 내에서 제공하는 뉴스, 영화 평점, 웹툰, 카페 서비스를 대상으로 이용자가 게시한 컨텐츠를 수집하여 연구를 수행하였다. 수집한 공개 정보의 양은 뉴스의 경우 뉴스 댓글 총 49,809,574건, 영화 평점 3,164,724건, 웹툰 댓글 69,710,203건, 카페 게시글 407,607건이다.

수집한 정보는 개인정보 노출 위험도를 수치화하여 평가하기 위하여 개인정보에 해당하는 속성을 보호된 이용자 ID, 전체 ID, 전화번호, 이메일, 성별, 연령대, 서비스별 카테고리, 작성 시간 등으로 정의하고 해당하는 정보를 추출 및 추론하였다. 이를 통해 전화번호 15,862건, 이메일 8,185건, 성별 추론 정보 약 73만 건, 연령대 추론 정보 약 73만 건의 개인정보를 확보할 수 있었다.

또한 포털 서비스 자체에서 제공하는 ID 부분 보호 처리(Masking)에도 불구하고 이용자를 유일하게 구별할 수 있는 닉네임과 프로파일 URL 같은 정보가 있었다. 이를 통해 부분 보호 처리 상태에서도 유일하게 구분되는 ID들이 있음을 확인하고, 이러한 ID가 각 서비스별로 4%에서 30%수준으로 존재함을 알 수 있었다(뉴스 서비스 156,623명, 영화 112,585명, 웹툰 133,356명, 카페 33,131명).

이와 같은 유일하게 존재하는 부분 보호 처리된 ID들은 서비스 간의 ID 연결에 활용될 수 있었다. 우리는 각 서비스에서 유일하게 식별되는 부분 보호 처리된 ID 들의 교집합이 존재함을 확인하였고, 이를 통해 ID 연계를 시도했다. ID 연계를 통해 서로 다른 서비스에 노출된 개인정보들을 집적한 빅테이블을 구축할 수 있었다. 이용자들은 부분 보호 처리된 ID를 통해 개별 서비스들 간의 연결이 안될 것이라는 인식을 갖고 있으므로 각 서비스의 성향에 따라 다양한 인격으로 활동하고 있으나, 본 연구를 통해 서비스 간 연결이 가능함을 보였으며 이는 이용자가 의도하지 않은, 또는 의도한 수준 이상의 개인정보 노출이 가능하다는 사실을 의미한다.

우리는 자신이 공개한 정보로 인한 개인정보 노출 수준을 이용자가 직관적으로 이해할 수 있도록 연결 가능성과 정보 민감도라는 개인정보 노출 평가 척도와, 연결 가능성과 정보 민감도 모두를 고려한 위험도라는 평가 척도를 제안하고 이를 실제 데이터에 적용하여 분석하였다. 분석 결과 각 서비스 간 연결 단계가 증가할수록, 연결 가능성과 정보 민감도, 그리고 그 둘의 조합인 위험도 수치가 증가하는 경향을 확인할 수 있었다.

마지막으로 각 서비스 내에 존재하는 이용자의 ID와 서비스 간 가능한 조합을 선택하였을 때, 본 연구에서 제안한 위험도를 사용하여 해당 조합 내 전체 이용자의 위험도 분포와 위험도 분포에서 선택된 이용자의 위험도가 어느 수준에 위치하는지를 시각적으로 표시하는 위험도 평가 프로토타입을 구현함으로써 이용자들이 자신의 위험도를 직관적으로 이해하고 자신이 공개한 정보 수준을 조절할 수 있는 정량화된 위험도 평가 기준을 제안하였다.

References

[1] H. Lee, J. Song, “A Research on De-identification Technique for. Personal Identifiable Information”, pp.1-51, Aug. 2016

[2] Press release, “How much my information expose in twitter?,” Korea Communications Commission, May. 2011

[3] H. Lim, “Analysis of personal information De-identification processing in big data environment”, E-finance and financial security, No. 8, pp.1-37,

Apr. 2017
[4] H. Kim, "Privacy protection technology for statistical anonymity", NIA Privacy Issues. June. 2012.
[5] L. Yahui, "Privacy protection in Big Data era", Journal of Computer Research and Development, vol.52, no. 1, pp.229-247, Jan.2015
[6] D. Choi, et al. "Personal Information Exposure on Social Network Service", Journal of The Korea Institute of Information Security and Cryptology, 23(5), pp.977-983, Oct. 2013
[7] D. Choi, et al. "Big Data Privacy Risk Analysis Technology", Journal of The Korea Institute of Information Security and Cryptology, 23(3), pp.56-60, June. 2013
[8] KISA, "In-depth report on overseas situation and case of personal information de-identification" KISA Power Review, pp.1-12, May. 2016
[9] J. Park, "Standardization of de-identification technology", Telecommunications technology association, ICT Standard Weekly, June. 2017
[10] H. Lee, J. Song, "A Research on De-identification Technique for. Personal Identifiable Information", SPRi, pp.1-65, Aug. 2016
[11] S. Garfinkel, "De-Identification of Personal Information", National institute of standards and technology, pp.1-46, Oct. 2015
[12] M. Lee, et al. "Analysis of the Facebook Profiles for Korean Users: Description and Determinants", Journal of Korean Society for Internet Information, 15(2) pp.73-85, Apr. 2014.
[13] C. Casper, "Hype Cycle for Privacy", Gartner, Jul. 2017
[14] J. Bambauer, K. Muralidhar, R. Sarathy, "Fool's Gold: An Illustrated Critique of Differential Privacy", Vanderbilt Journal of Entertainment & Technology Law, Vol. 16 No. 4, pp. 701-755, Aug. 2014
[15] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy", International Journal on Uncertainty, Fuzziness and Knowledge-based System, Vol.10, No. 5, pp. 557-570, Oct. 2002.
[16] A.Machanavajjhala, J.Gehrke, D.kifer. "l-Diversity:Privacy Beyond k-anonymity", Proceedings of the International Conference on Data Engineering, pp. 24-24, Apr. 2006
[17] N.Li, T.Li, S.Venkatasubramanian. "t-Closeness:Privacy Beyond k-anonymity and l-diversity," Proceedings of the International Conference on Data Engineering, pp. 106-115, Apr. 2007
[18] S.Lodha, D.Thomas. "Probabilistic Anonymity," Privacy, Security, and Trust in KDD: First ACM SIGKDD International Workshop, pp.56-79, Jan. 2007
[19] M. Zimmer, "But the data is already public": on the ethics of research in Facebook. Ethics and information technology, vol. 12, no. 4, pp. 313-325, Dec. 2010
[20] A. Ginosar, Y. Ariel, Y. "An analytical framework for online privacy research: What is missing?",. Information & Management, vol. 54, no.7, pp. 1948-957, Nov. 2017

〈저자소개〉

정 강 수 (Ksngsoo Jung) 정회원

2007년 8월: 서강대학교 컴퓨터공학과 졸업
2009년 8월: 서강대학교 컴퓨터공학과 석사
2017년 2월: 서강대학교 컴퓨터 공학과 박사
관심분야: 개인정보보호, 접근제어

박 석 (Seog Park) 종신회원
1978년 2월: 서울대학교 계산통계학과 졸업
1980년 2월: 한국과학기술원 전산학과 석사
1989년 8월: 한국과학기술원 전산학과 박사
1983년 9월~현재: 서강대학교 컴퓨터공학과 교수
관심분야: 개인정보보호, 접근제어

최 대 선(Dae-Seon Choi) 종신회원
1995년: 동국대학교 컴퓨터공학과 졸업
1997년: 포항공과대학교 컴퓨터공학과 석사
2009년: 한국과학기술원 전산학과 졸업 박사
1997년~1999년: 현대정보기술 선임
1999년~2015년: 한국전자통신연구원 인증기술연구실 실장/책임연구원
2015년~현재: 공주대학교 의료정보학과 부교수
2016년~현재: 정보보호학회 이사
관심분야: 인증, 개인정보보호, 이상거래탐지, 의료정보보안, 머신러닝