데이터와 알고리듬 프라이버시, 우리는 어디에 서 있는가

강 송 희*

Abstract

개인정보 유출, 소수자 혐오와 차별, 성희롱 등 다양한 윤리적 논란으로 서비스를 중단한 인공지능(AI) 챗봇 이루다 사건이 또다시 발생하지 않도록 하기 위한 방법은 무엇일까? 그 첫걸음은 이러한 상황을 방지하기 위한 그간의 노력과 현재 수준의 한계를 명확히 아는 것이다. 근본적인 제약은 데이터의 정확도와 신뢰성을 희생할수록 프라이버시·공정성 보장의 수준은 강화된다는 이율배반적인 상충관계에 있다. 이 글에서는 데이터를 연구개발이나 산업현장에서 활용할 수 있도록 허용할 때, 적절한 수준의 프라이버시를 보장할 수 있도록 하는 기술적, 통계적 방법과 이슈들을 사례와 함께 살펴본다.

I. 서 론

지난 1월 개인정보 유출, 성소수자 차별 등 윤리적 논란의 중심에 있다가 이제는 데이터셋까지 흔적 없이 사라진 이루다 챗봇 서비스는 연애 콘텐츠 앱 '연애의 과학'(2016년 6월)의 비식별화 데이터에 기초하여 2020년 6월부터 1,500명 규모로 베타테스트를 시작했고, 2020년 12월 22일에 정식 출시했다고 알려졌다. 2021년 1월 초에는 단기간에 누적 사용자 수 40만 명(10대 85%, 20대 12%)에, 누적 대화량 9천만 건을 기록하며 인기를 끌다, 사용자들의 성희롱을 포함한 성차별적 대화를 학습한 이루다가 성소수자 혐오 발언을 하면서 윤리적 논란이 점화됐다. 몇 가지 트리거가 될 수 있는 키워드를 포함한 대화로 개인정보가 고스란히 유출된다는 제보도 있었다. 이루다와 관련한 논란의 쟁점은 개인정보보호법

* 소프트웨어정책연구소 선임연구원, dellabee@spri.kr

위반 소지, 인공지능의 윤리적 문제 등이 얽혀 있었지만, 이루다 챗봇 서비스가 중지되고 데이터 셋을 폐기하게 된 주요 근거는 인공지능의 윤리적 문제가 아닌 개인정보보호법 이슈에 있었다고 할 수 있다. 논란의 중심에 있었을 1월 11일에 발표된 이루다 챗봇 개발사인 스캐터랩의 입장문에는 다음과 같은 내용이 있다.

> "데이터 활용 시 사용자의 닉네임, 이름, 이메일 등의 구체적인 개인 정보는 이미 제거돼 있습니다. 전화번호 및 주소 등을 포함한 모든 숫자 정보, 이메일에 포함될 수 있는 영어 등을 삭제해 데이터에 대한 비식별화 및 익명성 조치를 강화해 개인을 특정할 수 있는 정보는 유출되지 않았습니다."

이 글은 이루다 이슈를 '인공지능' 기술 개발의 문제로 보지 않는다. 산업 현장에서 개인정보보호법을 위반했다는 의혹을 사게 된 SW기업이 출시 서비스에 공교롭게도 인공지능을 활용하고 있었을 뿐이다. 중요한 것은 이러한 상황이 재발하지 않도록 하기 위한 방법을 모색하는 일이며, 그 첫걸음은 우리의 개인정보보호를 위한 기술적 수준과 한계를 명확히 파악하는 데 있다. 그리하여 이 글은 국내 개인정보 비식별 조치 가이드라인(2016)에 공식적으로 제시된 k-익명성이라는 솔루션과, 더 발전된 방식인 차등 프라이버시에 대해 알기 쉽게 설명하고, 이들이 해결할 수 없는 이슈들을 지적한 후 시사점을 도출해 본다.

Ⅱ. 본 론

1. 넷플릭스 공모 대회 사례와 k-익명성의 한계

2006년 넷플릭스가 아직 스트리밍 서비스가 아니라 우편 주문 DVD 대여 서비스를 제공할 때다. 서비스의 핵심 기능이라 할 수 있는 사용자 취향에 맞는 영화를 추천해주는 기능을 제공하기 위한 영화 추천 엔진의 개선을 목적으로, 넷플릭스는 공개적으로 협업 필터링 알고리듬[1]을 공모하기 시작했다. 넷플릭스가 공개한 데이터는 일부 사용자의 기존 영화 평가 점수(별점 1~5점)와 평가 날짜 목록이었고, 이 대회에서는 특정 사용자가 아직 보지 않은 영화를 평가하는 방법을 예측한 후, 가장 높은 평가를 받을 것으로 예상되는 영화를 추천할 수 있도록 구상하여 제출하면 되었다. 기존 추천 엔진의 시스템 정확도를 1% 개선하는 것은 어려운 문제이기 때문에 넷플릭스는 다년간에 걸쳐 100만 달러를 걸고 10% 이상

1) 협업 필터링(Collaborative Filtering)이란 유사한 사용자가 잘 평가한 항목을 기반으로 사용자에게 구매를 권장하도록 설계하는 기계 학습 문제이다.

정확도를 개선하는 것을 목표로 대회를 추진하게 됐다. 이 대회를 위해 넷플릭스는 많은 데이터를 공개하게 되었고, 그 데이터 셋은 약 50만 명의 사용자가 평가한 약 18,000개의 영화 정보를 담고 있는 1억여 건의 레코드로 구성돼 있었다. 당시 적용되던 美 비디오 개인 정보 보호법에 의하면 기록 유출 등의 사고에 대해 고객 당 최고 2,500달러의 배상 책임이 비디오 대여 업체에게 있는 상황이었기 때문에, 넷플릭스는 사용자 식별자를 모두 제거하고 각 사용자에 대해 고유하지만 의미가 없는 숫자 ID를 붙였다. 인구 통계적 정보(성별, 주소 등) 역시 모두 삭제됐다. 데이터는 사용자의 영화 평가 정보만을 담고 있었다. 그러나 이렇게 비식별화된 넷플릭스 데이터 셋이 공개된 지 2주 만에 사람들이 자신의 이름으로 영화 평가 점수를 공개하는 인터넷 영화 DB(IMDB, Internet Movie Database)를 상호 참조함으로써, 사용자가 영화를 6개 이상 평가했을 경우, 평가한 대략적인 날짜만 알고 있다면 99% 식별 가능하다는 연구[2]가 발표됐다. 실제로 자신의 성적 취향을 공개하지 않았던 성소수자인 어머니 한 사람이 넷플릭스를 고소했는데, 사람들이 넷플릭스에서 그녀가 본 영화를 알게 되면 그녀의 성적 취향이 공공연해지고, 그녀의 생계와 가족 부양에 지장을 줄 것이라고 주장한 것이었다. 그녀는 소송에서 넷플릭스 200만 명 이상의 구독자에 대한 법정 최대 벌금을 요구했고, 넷플릭스는 협의를 통해 이를 해결하는 한편 미국연방거래위원회의 시정 권고에 따라 두 번째 넷플릭스 공모 대회를 공식적으로 취소했다. 넷플릭스의 데이터 관리자는 비식별화된 데이터 셋을 공개할 때 IMDB와 같은 타 데이터 소스와의 상호 참조는 미처 고려하지 못했을 것이고, 실질적으로 모든 잠재적인 타 데이터 소스와의 결합을 고려한다는 것은 불가능한 일이다. 그래서, 비식별화, 특히 익명화된 데이터는 사실상 익명이 아니거나, 혹은 쓸 만한 데이터가 될 수 없다는 주장이 나온다[3].

재식별을 방지하기 위한 솔루션 중 하나가 바로 k-익명성인데, 이는 단순히 설명하자면 개별 레코드 정보를 수정하여 다른 특성 집합들이 서로 매칭되지 않도록 하는 것이라 할 수 있다. 개별 특성 집합은 민감한 정보와 민감하지 않은 정보를 모두 포함할 수 있는데, 민감하지 않은 정보와 민감한 정보를 연결하기 어렵게 만드는 게 k-익명성의 핵심 아이디어이다. 데이터 셋에 나타나는 민감하지 않은 정보의 조합이 공개된 데이터의 최소 k명의 개인과 일치하는 경우 이를 k-익명성이라 부르는 것이다. 이 k-익명성은 재식별을 효과적으로 방지할 수 있지만, 개인정보보호에 대한 위협은 재식별에서 그치지 않는다는 것이 이

2) Narayanan, A., & Shmatikov, V. (2006). "How to break anonymity of the netflix prize dataset". arXiv preprint cs/0610105.

3) Kearns, M., & Roth, A. (2019). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press.

슈이다. 예를 들어 광범위한 환자 데이터 셋이 있다고 할 때, 나흡연이라는 사람이 70대의 남성 흡연자인데, 알고보니 데이터 셋의 70대 남성 흡연자 레코드들은 모두 폐암에 걸렸거나 심장병에 걸렸다는 정보를 나타낸다면, 나흡연이라는 사람은 폐암에 걸렸거나 심장병에 걸렸음을 쉽게 유추할 수 있고, 이는 심각한 개인정보 침해가 된다. 이렇게 레코드가 모이면 더 많은 특성 정보가 보인다는 것은 k-익명성으로도 어떻게 해결할 수 없는 이슈이다. 더 심각한 문제는 k-익명성으로 공개된 여러 데이터 셋이 상호 참조될 때다. 개별적으로 각 데이터 셋이 k-익명성을 만족하더라도, 상호 참조하게 되면 이럴 것이라고 추정되던 특성 정보에 확실성이 더해진다. 첫 번째 데이터 셋에서 나흡연이라는 사람이 HIV, 폐암, 또는 심장병에 걸렸음을 알 수 있고, 두 번째 데이터 셋에서 HIV, 또는 대장암에 걸렸음을 알 수 있고, 이런 식으로 데이터 셋의 교집합을 찾아가다 보면 나흡연이라는 사람이 HIV에 걸렸음을 확신할 수 있게 되는 경우가 있다. 재식별만을 방지하는 것이 목적이라면 데이터를 집계하여 공개하면 된다고 생각할 수 있지만, 이러한 상호 참조 등을 통한 집계 데이터로부터도 개인 정보가 유출될 수 있음을 확인할 수 있다. 최신 연구에 따르면 훈련된 기계학습 모델에 대한 입력-출력 정보만 주어지면 훈련 데이터 셋에서 사용된 데이터를 높은 확률로 식별해 낼 수도 있다.

2. 차등 프라이버시와 애플, 구글의 사례

이러한 배경 하에서 등장한 것이 신시아 드워크라는 저명한 학자가 동료 세 명과 함께 주창한 차등 프라이버시[4]다. 차등 프라이버시란 특정 개인의 데이터가 데이터 셋에 포함된 경우와 그렇지 않은 경우 추론할 수 있는 특성 정보를 비교해야 한다는 아이디어에서 출발한 수학적 개념을 구체화한 것이다. 사용자가 데이터를 분석할 때는 일반적으로 원시 데이터를 사용하려 하지만, 원시 데이터를 사용하면 개인정보를 침해할 수 있기 때문에 문제가 된다. 차등 프라이버시는 사용자가 개별 데이터 요소를 식별할 수 없도록 데이터에 "노이즈" 또는 임의성을 추가하여 이 문제를 해결하려고 한다. 엡실론이라고 알려진 값은 데이터의 노이즈 또는 비공개 정도를 측정하는 지표라고도 할 수 있다. 엡실론은 노이즈나 프라이버시와 반비례 관계가 있어, 엡실론 값이 낮을수록 데이터의 노이즈 및 비공개 정도가 높다. 개인정보 보호가 목표이기는 하지만 데이터의 유용성 및 신뢰성과는 서로 상충 관계가 있다. 데이터 분석에서 정확도는 샘플링 오차로 인해 야기된 불확실성의 척도라고 생각

4) Dwork, C. (2006). "Differential Privacy, In inference Control in Statistical Databases". Springer.

할 수 있는데, 이러한 불확실성은 특정 범위에 속하는 경향이 있다. 차등 프라이버시 측면에서의 정확도라는 개념은 데이터의 신뢰성을 의미하며, 이것은 차등 프라이버시 메커니즘에 의해 야기된 불확실성에 영향을 받는다. 즉, 노이즈나 프라이버시 수준이 높을수록 엡실론, 정확도 및 신뢰성이 낮은 데이터로 변환된다고도 할 수 있다.

신시아 드워크[5]는 위와 같은 내용을 담아 2016년에 차등 프라이버시에 대해 수학적 정의를 내렸다. κ 를 어떤 랜덤화 함수(randomized function)라고 정의하고, 어떤 두 개의 데이터 셋 D1, D2간에 오직 한 명의 개인정보만 다르고 다른 개인들의 정보는 모두 동일하다고 가정한다. 만약 모든 경우의 집합 S $\subseteq$ Range(κ) 에 대해서 아래 식을 만족하면, 랜덤화 함수 κ 는 ε -차등 프라이버시를 보장한다. ε 의 값이 0에 가까울수록 강한 수준의 정보보호 상태가 보장된다.

$$Pr[\kappa(D1) \in S] \leq \exp(\varepsilon) \times Pr[\kappa(D2) \in S]$$

차등 프라이버시를 위한 최초의 대규모 상업적 배포 사례는 구글과 애플을 들 수 있다. 2014년에 구글은 보안 블로그를 통해 크롬 브라우저가 사용자 컴퓨터의 멀웨어에 대한 특정 사용 통계를 차등 프라이버시 방식을 적용하여 수집하기 시작했다고 발표했다. 크롬 사용 현황의 질문에 대한 사용자들의 예/아니오 형태의 응답에 노이즈를 더해 데이터베이스에 저장하여 개인정보를 관리하기 시작한 것이다. 2016년에는 애플이 아이폰에서 이모티콘 사용 내역을 수집할 때 이 차등 프라이버시 방식을 사용한다고 발표했는데, 이들은 모두 무작위 응답 방식의 알고리듬을 기반으로 하는 로컬 신뢰 모델을 채택하여 관련 개인 데이터 자체를 수집하지 않는다는 공통점이 있다. 그래서 구글은 '예/아니요'의 실제 비율이 아닌 비율의 추정량 및 분산추정량을 분석에 이용한다[6]. 두 번째 공통점은 레거시 시스템에 적용한 것이 아니라 이전에 전혀 수집하지 않았던 데이터를 처음 수집하면서 이러한 차등 프라이버시 방식을 적용했다는 것이다. 종합하면 이러한 차등 프라이버시 방식은 더 많은 데이터를 개인정보 침해 위협을 현재 수준에서 최소화하면서 수집할 수 있는 방법이기도 한 것이다.

한편, 로컬 신뢰 모델과 대조되는 중앙 집중식 차등 프라이버시 방식은 2017년 9월 미국

5) Dwork, C., & Rothblum, G. N. (2016). "Concentrated differential privacy". arXiv preprint arXiv: 1603.01887.

6) Park, M-J., & Kim, H. J. (2016). "Statistical disclosure control for public micro data: present and future". The Korean Journal of Applied Statistics, 29(6), 1041-1059

인구 조사국에서 2020년 인구통계 조사 시부터 적용하겠다고 발표했는데, 이는 데이터를 정확히 수집한 후 집계 통계에 차등 프라이버시를 적용하는 방식이다.

그런데, 이러한 차등 프라이버시는 사실상 소규모 그룹의 개인정보 보호에만 유용하다. 개인에 대해 3-차등 프라이버시를 보장하는 알고리듬이 있다고 할 때, k명의 개인의 데이터 셋에 대한 3k-차등 프라이버시를 제공한다고 말할 수도 있는데, 개념상 다시 생각해보면 이 k값이 크면 아무 의미 없는 보장 수준이 된다. 더구나, 공개적으로 사용 가능한 평균치 등의 상식적 데이터에서 개인 정보를 추론하는 것을 원천적으로 차단할 수는 없는 일이다. 신시아 드워크가 든 예를 살펴보면, "Terry Gross는 리투아니아 여자의 평균 키보다 2인치 작다는 정보"가 있을 때, 리투아니아 남녀의 평균 키를 제공하는 자료가 존재한다면 Terry Gross의 키는 노출될 수 있는 것이다.

나아가, 우리가 주목할 것은 차등 프라이버시의 개념 자체에서 내포하고 있는 제약, 혹은 한계이다. 이는 데이터의 정확도, 신뢰성과 개인정보 보호 수준 간의 관계가 서로 상충 관계라는 것이다. 개인정보 보호가 어느 수준으로 달성되어야 하는가는 지역적, 문화적 맥락에 따라 다를 수 있고, 같은 지역권이라 할지라도 산업별, 활용상의 요구사항에 따라 달라질 수 있다.

Ⅲ. 결 론

소프트웨어 및 통신 기술의 발달로 방대한 정보가 생성됨과 동시에 이를 검색하고 활용하는 수단도 첨단을 달리고 있다. 관련 기술의 발전과 산업적 활용을 위해 데이터의 공개와 활용의 중요성이 핀조명을 받게 됐고, 잠들어 있는 데이터를 문제의 소지 없이 공개하고 활용하기 위한 개인정보 보호 수단들도 빠르게 발전하고 있다. 이 글에서는 개인정보 비식별화 기법(가명화, 익명화 등) 중 k-익명성이라는 솔루션과 차등 프라이버시와 같은 발전된 수단의 개념과 사례, 그리고 그 한계를 다뤘다. 원론적으로 이야기하자면 개인정보를 포함한 데이터를 활용해야 하는 경우 데이터의 이용자는 개인정보 침해 사고가 발생하면 책임을 져야 한다. 그럼에도 불구하고 시대적 요구에 맞게, 데이터의 수집이나 보존이 아닌 사용에 목적을 두어 데이터를 보호해야 한다. 동시에, 데이터가 수집되고 보급되는 원리와 개인정보보호 원칙을 투명하게 공개할 필요도 있고, 개인정보 침해 위협을 관리하는 방안을 자체적으로 마련해야 한다. 정부에서 마련한 가이드라인은 사실상 최소한의 조치인데, 데이터를 활용하는 입장에서는 정확도, 신뢰도와 개인정보보호 수준 간의 상충 관계를

고려했을 때 최소한의 조치만을 채택할 개연성이 높다. 하지만 그럼에도 불구하고, 다루는 개인정보의 양이 많은 경우에는 오히려 법제도적 리스크, 즉 개인정보보호법 위반 관련 위험을 최소화하기 위한 수단을 지속적으로 연구하고 적용하려는 기업 관점의 노력이 절실하다. 예를 들면 구글, 애플과 같이 로컬 신뢰 모델 차등 프라이버시와 같은 방법을 고려할 수도 있다. 차등 프라이버시는 개념이 명확하기에 이미 오픈 소스로 구현된 프로젝트도 많이 존재하며, 요구되는 목표 수준에 따라 다양한 사례를 참조하여 개선하여 활용할 수 있다. 정부가 마련한 가이드라인을 지킨다는 것은 기업이 현실에서 맞닥뜨리는 법제도적 리스크와, 사용자·브랜드 신뢰 구축에 대한 위협을 모두 완벽하게 헷지해주지는 않는다. 기술적 측면에서, 사실상 엣지를 달리고 있는 혁신 주체는 정부가 아니라 기업이며, 그러하기에 개인정보보호를 위한 수단들도 기업 자체의 이익을 위해, 한 박자 느린 정부보다 기업이 더 그 첨단의 수준과 한계에 대해 잘 알고 있어야 하지 않을까 생각해 보며 부족하지만 글을 마친다.

참고문헌

"Narayanan, A., & Shmatikov, V. (2006). How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*."

Kearns, M., & Roth, A. (2019). The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press.

"Dwork, C. (2006). Differential Privacy, In inference Control in Statistical Databases. Springer."

"Dwork, C., & Rothblum, G. N. (2016). Concentrated differential privacy. arXiv preprint *arXiv:1603.01887*."

"Park, M-J., & Kim, H. J. (2016). Statistical disclosure control for public micro data:present and future. *The Korean Journal of Applied Statistics*, 29(6), 1041-1059"