물류 시스템에서 생산된 식별 가능한 비정형 데이터를 활용한 공개 가능한 비식별화 알고리즘 설계

(Design of Disclosed Unidentifiable Algorithms Using Identifiable Unstructured Data Produced by Logistics System)

정 철 우, 권 오 언, 김 성 호*
(재)경북IT융합산업기술원 연구개발부
(Cheol-Woo Jung, Oeon Kwon, and Sungho Kim)
(Research Development Division, Gyeongbuk Institute of IT Convergence Industry Technology (GITC))

Abstract : In this paper, we designed a de-identification algorithm for personal information in logistics systems. The proposed de-identification algorithm consists of four steps. The first stage collects the large amount of logistics data based on transportation systems. The second stage performs the definition and classification of logistics data. In the third stage, a machine learning is performed in a GPU-based high performance system. The last step is to perform de-identification. In future research, we will proceed with implementation and verification based on the proposed algorithm design.

Keywords : de-identification, natural language processing, machine learning, bigdata

Ⅰ. 서 론

최근 정부에서는 '포스트 코로나 시대의 디지털 정부혁신 발전 계획'과 '한국판 뉴딜 종합계획'의 일환으로 2025년까지 행정, 공공기관의 정보시스템이 클라우드 기반 통합관리 운영 환경으로 전환하여 민간에게 공개하는 것을 진행 중에 있다. 또한 한국지능정보사회진흥원(NIA)에서는 민간에서 활용하고 있는 주요 생산 데이터(물류 운송, 제조, 물성 등)를 공개하고자 하는 노력도 진행 중에 있다. 그러나 공공기관 및 민간에서는 대외적으로 공개하기 어려운 데이터(주민번호, 전화번호, 주소 등)로 인해 법적 분쟁이 끊이지 않은 실정이다[1].

이러한 문제를 해결하고자 정부에서는 2016년도에 '개인정보 비식별 조치 가이드라인'을 발표하였다. 해당 가이드라인에서는 개인정보를 식별자와 속성자로 분류하여 정의하고 있다. 식별자는 개인 또는 관련한 사물에 고유하게 부여된 값 또는 이름을 뜻하며, 속성자는 개인과 관련된 정보로서 다른 정보와 쉽게 결합하는 경우 특정 개인을 알아볼 수도 있는 정보를 의미한다[2]. 따라서 공공데이터는 식별자와 속성자에 대한 비식별 조치를 진행한 이후 데이터를 공개해야한다. 그러나 기존에 활용중인 데이터를 기반으로 공공데이터 공개를 진행할 경우 대용량의 데이터를 하나하나 수정하기도 어려우며, 그에 따른 많은 비용이 초래된다.

이러한 문제를 해결하기 위해 많은 연구자들은 식별자와 속성자의 데이터를 자연언어처리, 데이터 마이닝 기법 등을 활용하여 외부에 공개할 수 있도록 자동으로 처리할 수 있는 많은 기법들에 대한 연구가 진행 중에 있다[3]. 그러나 기존에 다양한 기법들은 정형화된 데이터를 생산하고 추출하는 시스템에 적합하도록 설계되어 있다. 이는 정형화되지 않은 데이터에서 비식별화를 수행할 경우 정상적으로 수행되지 않을 수 있다. 특히 물류 데이터를 생산하는 시스템은 체계적으로 데이터를 관리하고 있지 않기 때문에 비정형화 데이터를 활용한 비식별화 기법이 절실하다.

본 논문에서는 물류 운송 시스템에서 생산된 비정형 텍스트 데이터에서 등장하는 다양한 단어들에 대한 정보를 바탕으로 개인정보에 대한 비식별화 알고리즘을 설계를 진행하고자 한다. 이는 물류 운

* 교신저자(Corresponding Author)
정철우, 권오언, 김성호 : (재)경북IT융합산업기술원
※ 본 연구는 중소벤처기업부의 규제자유특구혁신사업육성 지원에 의한 연구임 [P0020333]

송 시스템에서 데이터 특성에 따라 개인정보 데이터 비식별에 활용할 수 있을 것이며, 이를 기반으로 공공데이터를 활용할 수 있는 방향성을 제시할 수 있을 것 이다.

Ⅱ. 데이터 특성분석 및 알고리즘 설계

1. 물류데이터 특성분석

최근 대기업을 중심의 물류 업체에서는 물류와 관련한 주요 데이터(배송, 송장 등)를 공공에 활용 가능한 연구 목적으로 온라인에 공개하고 제공하고 있다[4]. 이러한 공개 데이터 정보를 기반으로 이 절에서는 비정형 데이터 비식별 알고리즘 설계를 위한 물류 주문 데이터 분석을 진행하고자 한다.

표 1은 물류 주문 데이터 관련한 데이터 셋을 보여주고 있다. 표 1에서 물류 주문 조회 데이터 셋은 이름, 주민등록번호, 휴대폰번호, 주소가 식별자, 속성자의 특성 범주 부합한 도메인이다. 따라서 물류 주문 데이터를 공공에게 제공하기 위해서는 표 1의 이름, 주민등록번호, 휴대폰번호, 주소 데이터를 비식별화를 통해 공개하는 것이 정보제공 및 개인정보 유출을 방어할 수 있는 필수적인 요소라고 할 수 있다.

표 1. 물류 주문 목록 관련 데이터 셋
Table 1. Dataset of Logistics order list

구분	항목	설명
검색	회원구분	전체/회원/비회원
	이름	회원 이름
	주민등록번호	회원 주민등록번호
	휴대폰번호	회원 휴대폰번호
	주소	배송지 주소
	검색어	최대 10개까지 동시 검색 가능(콤마로 구분),주문번호(디폴트)/주문자정보/수령자정보
	기간	주문일/메모작성일/입금확인일
	상품	상품정보
	희망배송업체	고객이 주문 시 선택한 배송업체/방식
	주문 상태	상품 준비 중/배송 준비 중/ 배송보류/배송대기/배송 중/배송완료
	CS상태	취소/교환/반품/환불
	입금/결제상태	입금전/추가입금대기/입금완료(수동)/입금완료(자동)/결제완료

2. 비식별화 알고리즘 설계

이절에서는 본 논문에서 제시하는 비식별화 알고리즘에 대한 설계를 진행하고자 한다. 그림 1은 제안하는 알고리즘의 구조도를 보여주고 있다.

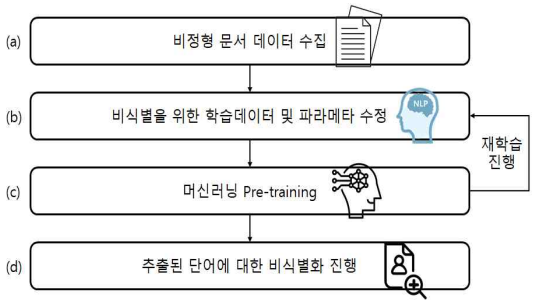

그림 1. 비식별화 알고리즘 흐름도
Fig. 1. Flowchart of non-discrimination algorithm

제안하는 비식별화 알고리즘의 상세 동작구조는 아래와 같다.

가장 먼저 그림 1 (a)에서는 비정형 데이터를 수집한다. 비정형 문서 데이터 수집은 기존의 데이터 수집에 용이한 R-DBMS 또는 NoSQL과 같은 시스템을 활용하여 대용량 데이터 수집이 용이한 형태로 시스템을 구축한 이후 수집을 진행한다. 이후 제안하는 알고리즘에서는 비식별화 알고리즘에 연계를 수행하기 위해 R, python 등 서비스를 활용하여 연계 작업을 진행한다.

그림 1 (b)에서는 비식별(식별자, 속성자)에 대한 데이터 정의 및 분류를 진행한다. 물류 데이터의 경우에는 불특정한 정보를 기반으로 의미를 도출해야한다. 따라서 이러한 정보를 추출하기 위해서는 자연언어처리(Natural language Processing, NLP) 모델 중 토큰화 기법, 필터링 기법, 정규화, N-그램, 품사 태깅 등을 활용하여 도출할 수 있다. 특히 자연언어 처리는 한국어를 도출하는데 용이한 특성이 있어 텍스트마이닝 대비 의미 도출에서 많은 활용 및 연구가 되고 있다[5-7].

그림 1 (b)를 통해 추출한 의미 데이터를 기반으로 그림 1 (c)에서는 수집된 학습데이터의 머신러닝 학습을 진행한다. 그림 1 (c)에서는 머신러닝 학습결과에 따라 파라메타 값 수정 등을 통해 요구하는 값에 도달할 때 까지 그림 1 (b)와 그림 1 (c)를 반복적으로 재학습을 진행한다.

마지막으로 그림 1 (d)에서는 머신러닝 결과로 나온 단어들에 대한 비식별화를 진행한다. 표 2는 표 1의 비정형 물류 주문 조회 데이터에 대한 비식별화 예시를 보여주고 있다. 표 2에서는 식별 전의 이름, 주민등록번호, 핸드폰번호, 주소 등에 대해 학

습된 모델을 적용하게 되면 이름과 주소지는 가명으로 변경되며, 주민등록번호와 핸드폰번호는 별표(*)로 변경되게 된다.

표 2. 물류 데이터 개인정보 비식별화 예시
Table 2. Logistics data Example of non-identification of personal information

비식별 전	202204.회원.김광석.810412-1234567.010-1234-1234.서울특별시 금천구.레저/건강.가정용가구.5601.60..배송준비중.입금완료(자동) ... (생략)
비식별 후	202204.회원.**김A002.810412-1******.010-****-****.OO시 OO구.**레저/건강.가정용가구.5601.60..배송준비중.입금완료(자동) ... (생략)

Ⅲ. 결 론

본 논문에서는 물류 운송 시스템에서 생산된 비정형 텍스트 데이터에서 등장하는 다양한 단어들에 대한 정보를 바탕으로 개인정보에 대한 비식별화 알고리즘을 설계를 진행했다. 제안하는 비식별화 알고리즘은 크게 네 단계로 구성하였다. 첫 번째 단계는 데이터 수집을 진행하며, 이후 데이터 정의 및 분류를 두 번째 단계에서 진행한다. 세 번째 단계에서는 머신러닝 학습을 수행한다. 마지막 단계는 비식별화 수행을 진행한다. 본 논문에서 제안하는 비식별화 알고리즘을 통해 공공데이터 제공 정부의 지침에 발맞추어 방대한 데이터를 공개하면서도 개인정보를 보호할 수 있는 기법으로 활용할 수 있을 것을 기대한다.

향후 연구에서는 제안하는 알고리즘 설계를 기반으로 구현 및 검증을 진행할 것이다.

참 고 문 헌

[1] 임지훈, 윤상필, 권헌영, "포스트 코로나 시대의 디지털 정부혁신 방향과 공동체의 역할 과제", 디지털 윤리, 제 4권, 제 2호, 1-17쪽, 2020.

[2] 국무조정실, 행정자치부, 방송통신위원회, 금융위원회, 미래창조과학부, 보건복지부, "개인정보 비식별 조치 가이드라인", 2016.

[3] Shah, D., Schwartz, H. A., & Hovy, D., "Predictive biases in natural language processing models: A conceptual framework and overview", arXiv preprint arXiv:1912.11078, 2019.

[4] CJ대한통운, "빅데이터로 관찰한 일상생활리포트", 2020-2021.

[5] Young, T., Hazarika, D., Poria, S., & Cambria, E, "Recent trends in deep learning based natural language processing", ieee Computational intelligenCe magazine, Vol. 13, No. 3, pp. 55-7, 2018.

[6] 박상언, "딥러닝 중심의 자연어 처리 기술 현황 분석", 한국빅데이터학회지, 제 6권, 제 1호, 63-81쪽, 2021

[7] SKTBrain, "Korean BERT pre-trained cased", Github repogitory, Github https://github.com/SKTBrain/KoBERT, 2019.