Original Article

https://doi.org/10.12985/ksaa.2021.29.4.160
ISSN 1225-9705(print) ISSN 2466-1791(online)

항공안전 보고 데이터 텍스트 분석 기반 조성을 위한 비식별 처리 기술 적용 연구

황도빈*, 김영곤**, 심영민***

A Study on De-Identification Methods to Create a Basis for Safety Report Text Mining Analysis

Do-bin Hwang*, Young-gon Kim**, Yeong-min Sim***

ABSTRACT

In order to identify and analyze potential aviation safety hazards, analysis of aviation safety report data must be preceded. Therefore, in consideration of the provisions of the Aviation Safety Act and the recommendations of ICAO Doc 9859 SMM Edition 4th, personal information in the reporting data and sensitive information of the reporter, etc. It identifies the scope of de-identification targets and suggests a method for applying de-identification processing technology to personal and sensitive information including unstructured text data.

Key Words : De-Identification(비식별화), Aviation Safety Mandatory Report(항공안전 의무보고), Aviation Safety(항공안전), Unstructured Data(비정형 데이터), Text-Mining(텍스트마이닝)

Ⅰ. 서 론

항공안전 데이터 분석에 있어 선제적으로 이행되어야 하는 절차는 잠재적인 위해요인을 식별하여 이를 바탕으로 분석과제를 정의하는 단계로, 이를 위해서는 항공안전 보고 데이터를 마이닝 기법으로 분석된 결과를 활용하여야 한다.

항공안전 보고 중 가장 보편적이며 공식적인 데이터는 항공안전 의무보고로 통합 항공안전관리 시스템(NARMI)에서 관리되고 있으며, 데이터 항목을 사전 검토한 결과 이벤트의 발생경과와 원인·위해요인을 추정할 수 있는 핵심 데이터는 텍스트 형태로 기록된 발생내용, 조치사항, 조사결과 등으로 파악된다.

텍스트 형태의 발생내용, 조치사항, 조사결과 등의 데이터 현황을 샘플링하여 검토한 결과 기록된 데이터에서 항공안전법 제59조 제2항, 제61조의 3에 따른 정보공유 제약 사항 및 정보보호 범위에 포함되는 정보가 다수 존재하는 바, 텍스트로 기록된 핵심 보고 데이터의 분석을 위해서는 동 정보의 비식별 조치가 필요하다.

본 논문에서는 항공안전 보고 데이터 가운데 텍스트 형태의 발생내용, 조치사항, 조사결과 등의 데이터에서 개인정보 및 데이터 제공기관 측면의 민감정보에 대한 자동화된 비식별화 처리를 위한 기술적용 방안을 제시하고자 한다.

Received: 18. Aug. 2021, Revised: 09. Dec. 2021, Accepted: 13. Dec. 2021
* 항공우주산학융합원 항공우주정보센터 전임연구원
** 네이버시스템(주) I&G 사업본부 이사
*** 항공우주산학융합원 항공우주정보센터 전문위원
연락저자 E-mail : sierra7@iiaci.or.kr
연락저자 주소 : 인천 연수구 갯벌로 36, 351호

II. 본 론

2.1 국내 데이터 비식별화 적용 동향

2.1.1 정부의 개인정보 비식별 조치 절차

국내에서는 데이터 활용의 활성화를 위해 데이터 3법의 제개정과 더불어 데이터에서 존재하는 개인정보의 보호를 위해 개인정보보호법에 근거하는 개인정보 비식별 조치 가이드라인을 '18년도 제정하여 정부 및 공공기관 차원에서 분석을 위해 데이터 활용 시 보호되어야 하는 개인정보에 대한 비식별 조치사항을 규정하고 비식별 처리 기술에 대해 정의하고 있다.

비식별 조치 절차는 크게 사전검토, 비식별 조치, 적정성 평가, 사후 관리 등의 4단계를 거쳐 조치하도록 하며, 주로 비식별 처리에 대한 기술은 2단계인 비식별 조치 단계에서 적용하는 구조로 개인정보 보호를 위한 비식별화 조치에 대해 Fig. 1의 절차로 이행하도록 권고하고 있다[1].

개인정보 비식별 처리는 사전검토 단계에서 식별된 비식별 대상 개인정보에 대해 검토 후 비식별 조치 단계에서 비식별 대상 데이터에 대해 일부 삭제 또는 대체하는 등의 비식별 처리 기술을 활용하여 데이터에 대한 비식별화 조치를 이행한다[1].

적정성 평가 단계에서는 비식별화 조치된 개인정보 데이터를 타 정보와 결합하여 식별 가능한지 여부 검증하는 등의 비식별 조치 결과를 평가하여 확정하고, 사후 관리 단계에서는 비식별 조치된 데이터 모니터링을 토대로 재식별 방지를 위한 조치를 지속하는 절차를 이행하도록 권고하고 있다[1].

2.1.2 비식별 처리 기술 정의

개인정보 비식별 조치 가이드라인에서 정의하는 비식별 조치 절차 가운데 비식별 처리 기술이 적용되는

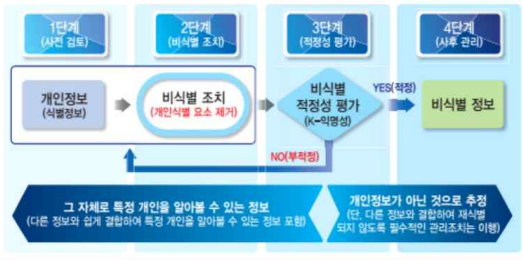

Fig. 1. Personal information de-identification action procedure[1]

2단계 비식별 조치 절차에서 적용되는 기술은 가명처리, 총계처리, 데이터 삭제, 데이터 범주화, 데이터 마스킹 5개 분야에 17개 기술을 정의하고 있으며, 개인정보의 특성과 활용성을 고려하여 선별하여 비식별 처리 기술을 적용할 수 있다[2](Fig. 2).

현재 가이드라인에서 정의된 17개 비식별 조치기술 가운데 국제표준기구에서 ISO/IEC 20889 권고 표준으로 정의된 기술과의 연계되는 범위에 포함되는 암호화, 총계처리, 부분총계, 라운딩, 재배열, 감추기, 랜덤라운딩, 범위 방법, 임의잡음 추가, 공백 처리 등의 기술은 데이터 비식별 처리에 대한 표준화 범위 내에서 해외 상호 적용이 가능한 것으로 파악된다[3].

2.2 국외 항공 데이터 비식별화 동향 및 적용 사례

국제민간항공기구는 회원국에게 항공안전 데이터 등을 수집, 저장, 분석, 활용하여 사전적이고 사후적 관리를 통한 안전담보 기반의 중요성을 강조하며, ICAO Doc 9859 SMM Edition 4th를 통해 사고예방 및 안전증진을 위한 안전 데이터수집 처리 체계(SDCPS)에서 공유되는 데이터는 개인정보 및 데이터 제공자의 민감정보를 보호할 수 있는 비식별 조치를 강구하도록 권고하고 있다[5].

국제 항공운송협회(IATA)에서는 세계 92개 회원사의 항공기로부터 생성되는 QAR(quick access recoder) 데이터를 수집하여 다양한 비행자료 분석 서비스를 제공하는 FDX(Flight Data eXchange) 프로그램을 운

Fig. 2. Relationship diagram between ISO/IEC 20889 standard and guidelines for measures against personal information de-identification[2]

영하는 과정에서 비행 데이터 분석 활용 시 비식별화를 위한 규칙을 정하고 이를 토대로 데이터 비식별 조치 후 분석에 활용하고 있다[6].

FDX(Flight Data eXchange) 프로그램에서 수집된 비행자료 가운데 비식별 조치되는 데이터 항목은 항공사, 항공기 등록부호, 항공편 등 3개 항목은 데이터 삭제 분류 가운데 레코드 삭제 기술을 적용하고 있으며, 분석 활용에 필요한 데이터 Set을 구성하는 경우 동일기종을 운항하는 3개 항공사 이상의 데이터 집합체를 구성하는 형태인 데이터 범주화의 범위 방법 비식별 처리 기술을 적용하여 항공안전 트렌드 분석에 활용하고 있다[6](Table 1).

2.3 항공안전 의무보고 데이터 비식별화 기술 적용 방안

2.3.1 항공안전 의무보고 데이터 구조

항공안전 의무보고는 항공안전 관련 사고 및 준사고 등의 이벤트 발생 시 해당하는 이벤트 발생에 대한 주요 내용을 기록하여 보고하는 체계로 통합 항공안전정보 시스템(NARMI)을 통해 운영관리되고 있다.

항공안전 이벤트 발생 시 등록되는 항공안전 의무보고 데이터는 일자 형태로 기록되는 발생일자 및 시간 정보와 코드 또는 단일항목 내 단일정보만을 기록하는 구조의 항공사, 기종, 항공기 등록부호, 운항편명, 출발공항, 도착공항, 활주로, 발생위치, 발생구분, 원인분류, 비행단계, 이벤트유형, 보고종류, 보고자, 보고자 소속 등의 항목과 단일 데이터 항목에 다수의 정보를 기록하는 비정형 텍스트 구조의 발생내용, 조치사항, 조사결과 등으로 구성된다(Table 2 및 Fig. 3).

2.3.2 항공안전 의무보고 데이터 비식별 대상 범위 및 적용 기술 정의

국내 항공안전 데이터는 항공안전법 제59조 제2항,

Table 1. Examples of de-identification processing in the IATA FDX program[6]

데이터	비식별 처리 기술	비식별화 범위
FDX	데이터 삭제 (레코드 삭제)	항공사, 등록부호, 항공편
	데이터 범주화 (범위 방법)	동일 기종 운항 3개 항공사 이상 데이터 집합체 구성을 통한 분석용 데이터 set 구성

Table 2. Aviation safety mandatory reporting data structure

데이터	특징	항목
항공안전 의무보고	코드, 단일항목 내 단일정보(정형)	발생일자, 시간, 항공사, 기종, 등록부호, 운항편명, 출발공항, 도착공항, 활주로, 발생위치, 발생구분, 원인구분, 비행단계, 이벤트 유형, 보고종류, 보고자, 보고자 소속 등
	단일항목 내 다수 정보(비정형)	발생내용, 조치사항, 조사결과 등

Fig. 3. Aviation safety mandatory report registration system

제61조의 3에 따라 항공안전보고 데이터의 외부공개 금지 및 개인정보보호 조치를 규정하고 있는 근거에 따라 항공안전 의무보고 데이터 분석 활용 시에는 개인정보 및 보고 데이터 제공자의 민감정보에 대한 비식별화 조치가 필요하다.

특히, 개인정보보호법 제2조 제1호에서의 개인정보 범위는 개인에 관한 성명, 주민등록번호 및 영상 등으로 개인을 식별할 수 있는 정보로서 해당 정보만으로는 특정 개인을 식별할 수 없더라도 타 정보와 결합하여 식별 가능한 경우에도 개인정보 보호의 범위에 포함하고 있으므로 이를 고려한 비식별 조치 대상 범위가 정의되어야 하므로, 항공안전 의무보고 상에서 정보 제공자 측면의 민감정보를 포함한 비식별 조치가 요구된다[7].

항공안전 의무보고 데이터 항목 가운데 개인정보는 보고자 및 보고자의 소속에 해당하는 데이터 항목으로 파악되며, 데이터 제공자 측면의 민감정보는 IATA FDX 프로그램 적용 사례와 보고 당사자의 신분 확인을 간접적으로 확인 가능한 수준에 대한 특성을 고려한 결과 항공사, 항공기 등록부호, 운항편명, 출발공항,

도착공항, 발생내용, 조치사항, 조사결과 등으로 파악된다(Table 3).

비식별 조치 대상 데이터 항목 가운데 항공사, 등록부호, 운항편명, 출발공항, 도착공항, 보고자, 보고자 소속 등의 정보는 단일항목 내 단일항목 속성 및 코드 형태 속성의 정형화된 데이터로 저장되어 있어 IATA FDX 프로그램에서 적용된 레코드 삭제와 같은 데이터 삭제 기술 또는 임의잡음 추가 및 공백과 대체와 같은 데이터 마스킹 등의 기술을 적용하여 일괄적인 비식별화 처리가 가능하다(Table 4).

특히, ISO/IEC 20889 표준과의 연계성을 고려할 경우 데이터 삭제 기술보다는 임의잡음 추가 및 공백과 대체와 같은 데이터 마스킹 기술 적용이 바람직할 것으로 판단된다.

Table 3. Data columns required for de-identification during aviation safety mandatory report

데이터	특징	비식별 대상 항목
항공안전 의무보고	코드, 단일항목 내 단일정보(정형)	항공사, 등록부호, 운항편명, 출발공항, 도착 공항, 보고자, 보고자 소속 등
	단일항목 내 다수 정보(비정형)	발생내용, 조치사항, 조사결과 등

Table 4. Data de-identification technology[4]

분야	기술	설명	ISO/IEC 20889
가명처리	휴리스틱 가명화	식별자에 해당하는 값을 정해진 규칙으로 대체 또는 인위적 판단에 의해 가공하여 개인정보를 숨기는 기술	-
	암호화	정보 가공 시 일정한 규칙의 알고리즘을 적용하여 암호화하여 개인정보를 대체하는 기술	Cryptographic tool
데이터 삭제	레코드 삭제	다른 정보와 뚜렷하게 구별되는 레코드 전체를 삭제 처리	-
데이터 범주화	범위 방법	임의의 특정 범위 구간을 설정하여 해당 범위 또는 구간 내 식별정보를 포함하여 감추는 기술	Top & bottom coding
데이터 마스킹	임의잡음 추가	개인 식별 가능 정보에 임의 잡음 추가 기술	Noise addition
	공백과 대체	특정 항목 일부 또는 전부에 대체문자 치환	Masking

비식별 조치 대상 데이터 항목 가운데 발생내용, 조치사항, 조사결과 등의 정보는 단일항목 내 불규칙적인 텍스트 형태 속성에 다수의 비식별 대상 정보를 포함하는 비정형 데이터로 저장되어 있어 일괄적인 비식별 조치 기술을 적용하는 데 한계가 있다.

2.3.3 텍스트 형태 보고 데이터 비식별화 처리 알고리즘 모델 정의

비식별 조치 대상 데이터 항목 가운데 발생내용, 조치사항, 조사결과 등의 정보에 대한 sampling 검토 결과, 불규칙적으로 기록된 항공사, 등록부호, 운항편명, 출발공항, 도착공항 등의 보고자 측면의 민감정보를 추적하는 rule을 생성하고 이를 토대로 식별된 대상 정보를 비식별 코드로 치환하는 비식별화 처리 알고리즘 기술 적용이 요구된다(Table 5).

특히, 발생내용, 조치사항, 조사결과 등에 불규칙적으로 기록된 항공사, 운항편명, 출발공항, 도착공항 등의 비정형 텍스트 데이터에 대해서는 식별하는 Rule을 정의하여 식별 추적하고, 동일한 의미로 비식별하는 치환값으로 대체할 수 있는 데이터 마스킹 처리 알고리즘을 적용한다.

식별 대상 문자열 가운데 비식별 처리 과정에서 오류를 야기할 수 있는 문자열에 대해서는 문자열 시작값과 종료 값에 오류 발생을 야기하는 문자열 패턴을 식별처리하는 rule을 적용하여 예외 처리하는 알고리즘을 적용한다.

텍스트 형태의 보고 데이터에 대한 비식별 처리 알고리즘 구조는 Fig. 4에 정의된 프로세스 모델에 따라 구현하여 '20년도 항공안전 의무보고의 발생내용 항목에 적용한 결과 불규칙화된 텍스트 구조로 기록된 항공사, 등록부호, 운항편, 공항 등의 정보가 데이터 마스킹 형태의 비식별 처리 기술을 적용하여 대체문자로 치환 처리된다.

텍스트 보고 데이터 비식별화 처리 알고리즘을 '20년도 1,094건의 항공안전 의무보고에 반영한 결과, 예외처리 미반영시 4,930건의 문자가 비식별 처리되었으며, 예외처리 반영 시 4,303건의 문자가 비식별 처리된다(Table 6).

항공안전 의무보고에서 예외처리를 반영하지 않는 경우 비식별 처리 정확도가 87.3%(오인식률 12.7%) 수준으로 파악되었으나, 비식별 오류를 보정한 예외처

Table 5. Define for de-identification target tracking rule

비식별 대상 텍스트 정보	비식별 대상 추적 rule 정의	조치사항
등록부호	HL + 4자리 숫자 구조 식별	4자리 숫자 일괄 #### 대체문자 처리
운항편	IATA 2 letter code + 3 또는 4자리 숫자 구조 (예, KE858, BX8808 등)	IATA 2 letter code를 일괄 '운항편'으로 대체문자 처리
	ICAO 3 letter code + 3 또는 4자리 숫자 구조 (예, KAL858, ABL8808 등)	ICAO 3 letter code를 일괄 '운항편'으로 대체문자 처리
	예외 처리 : TAKE, SMOKE 등 앞자리 및 뒷자리 문자열 동반 구조 문자	
항공사	국문 항공사명	항공사 국문명을 일괄 '항공사'로 대체문자 처리
	IATA 2letter code (예, KE, BX 등)	IATA 2 letter code를 일괄 '항공사'로 대체문자 처리
	ICAO 3letter code (예, KAL, ABL 등)	
	예외 처리 : TAKE, SMOKE 등 앞자리와 뒷자리 단일 문자가 아닌 문자열	ICAO 3 letter code를 일괄 '항공사'로 대체문자 처리
공항명	국문 공항명	공항 국문명을 특정 가명 변환 처리
	IATA 3letter code (예, GMP, PUS 등)	IATA 3 letter code를 특정 가명 변환 처리
	ICAO 4letter code (예, RKSS, RKPK 등)	ICAO 4 letter code를 특정 가명 변환 처리

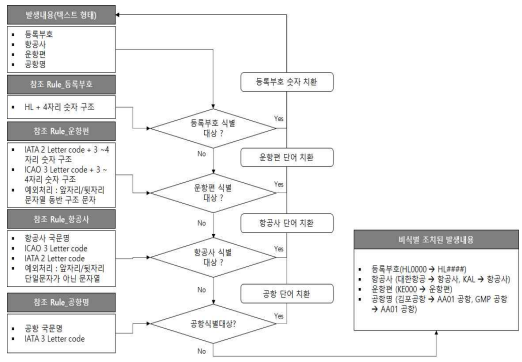

Fig. 4. Aviation safety mandatory report text data de-identification process algorithm model

리 rule을 반영한 결과치에서는 100%의 정확도를 나타내고 있다(Fig. 5).

'20년도 1,094건의 항공안전 의무보고 데이터를 대상으로 비식별 알고리즘을 활용하여 처리하는 데 소요된 시간은 3시간 수준으로, 기존 수작업으로 1개년도 발생내용 데이터의 비식별 조치에 14시간 이상의 시간이 소요되는 점을 고려할 경우, 70% 수준의 비식별 처리 시간 단축 효과를 기대할 수 있는 것으로 판단된다.

Table 6. Results of de-identification of aviation safety mandatory report

구분	의무보고 비식별 처리 실적	
	예외처리 미반영	예외처리 반영
'20년 의무 보고 건수	1,094	
전체 비식별 처리 문자 수	4,930	4,303
보고 1건당 평균 비식별 처리 문자 수	4.5	3.9

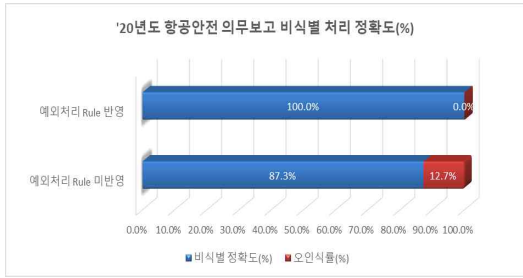

Fig. 5. Aviation safety mandatory report de-identification algorithm accuracy (Misrecognition rate due to application of exception handling)

Ⅲ. 결 론

본 논문에서 제시된 기술은 잠재적인 항공안전 위해요인을 식별하기 위해 선제적으로 이행되어야 하는 항공안전 보고 데이터 분석에 필요한 개인정보 및 민감정보의 신속한 비식별 처리를 가능하게 하는 모델이다.

수작업으로 정형화되지 않은 텍스트를 보고 비식별화 대상 데이터를 식별하여 처리하던 기존 방법에서 자동화 처리의 기반을 마련하여 항공안전 분석 과정의 효율성을 한층 증대시킬 수 있다.

본 연구에서 제시된 기술은 빅데이터 기반 항공안전 분석 기술 가운데 비정형화된 텍스트 데이터를 분석 가능하도록 처리하는 과정에서 적용되어야 하는 비식별화 기술로 향후 항공안전 빅데이터 분석 플랫폼의 데이터 처리에 대한 기능으로 설계되어 구현이 필요하다.

비정형 텍스트 데이터에 포함된 개인정보 및 민감정보를 추적·식별하여 비식별화 처리하는 알고리즘은 항공안전 빅데이터 분석 플랫폼 상에서 항공안전 보고 데이터를 분석하는 데 필요한 비정형 데이터 처리 기술적 기반을 제공함으로써 항공안전 이해관계자가 개인정보 및 민감정보의 노출로부터 안전하게 데이터를 공유할 수 있는 체계적 기반 조성 효과가 기대된다.

후 기

본 연구는 빅데이터 기반 항공안전관리 기술 개발 및 플랫폼 구축사업(20BDAS-B158275-01)을 통해 수행되었다.

References

1. OGPC, MOIS, KCC, FSC, MSIFP, and MOHW, “Guidelines for actions against personal information de-identification”, 2016, KISA, p.3.
2. OGPC, MOIS, KCC, FSC, MSIFP, and MOHW, “Guidelines for actions against personal information de-identification”, 2016, KISA, p.30.
3. Cha, Y. C., “Data economy and personal information de-identification technology trends”, Weekly Technology Trend, IITP, 2019, p.22.
4. OGPC, MOIS, KCC, FSC, MSIFP, and MOHW, “Guidelines for actions against personal information de-identification”, 2016, KISA, pp.31-35.
5. IIACI, and INNOSKY, “Big data-based aviation safety management technology development and platform development 1st R&D technical research report - ICAO aviation safety data system”, KAIA, 2020, pp.16-17.
6. IIACI, “Big data-based aviation safety management technology development and platform development 1st R&D Report – IC100 Establishment of aviation data classification system, definition of technical standard and establishment of format”, KAIA, 2020, p.17.
7. OGPC, MOIS, KCC, FSC, MSIFP, and MOHW, “Guidelines for actions against personal information de-identification”, 2016, KISA, p.52.