ISO/IEC 5259 데이터품질특성에 기반한 반정형 및 비정형 데이터 품질검증 적용 사례

박정하, 이진용, 박경은, 허준호, 이상복

parkjh516@tta.or.kr, wlsdyd620@tta.or.kr, kepark@tta.or.kr, jhher@tta.or.kr, jangpo@tta.or.kr

Application examples of semi-structured and unstructured data quality verification based on ISO/IEC 5259 data quality characteristics

Park Jeong Ha, Lee Jin Yong, Park Kyung Eun, Her Jun Ho, Lee Sang Bok

Telecommunications Technology Association

요 약

ISO/IEC 5259시리즈는 ISO/IEC 25012와 25024에 기반한 데이터 품질특성과 품질측정 방법에 따라 데이터 품질모델을 설명하고 있으며, 데이터 분석(Data Analytics) 및 기계학습(Machine Learning) 기반의 데이터 품질특성을 부가적으로 정의하고 있다. 현재 빅데이터 및 인공지능 분야의 기반이 되는 데이터는 구조화된 데이터 외에도 반정형, 비정형 데이터와 같이 다양한 유형의 데이터가 활용되고 있으므로, 정보시스템 기반의 데이터 품질특성 외에 데이터 분석 및 기계학습을 위한 데이터 품질특성을 동시에 고려할 필요가 있다. 본 연구는 ISO/IEC 5259-2 데이터품질특성 중 본 데이터 사례에 적용 가능한 품질특성 7가지를 선정하고, AIHub의 인공지능 학습용 데이터 품질검증 항목과의 비교 및 시사점을 제안하고자 한다.

Ⅰ. 서 론

최근 빅데이터 및 인공지능 분야의 기술 발전 및 인공지능 기반의 비즈니스 가치 창출의 범위와 속도는 더욱 확대되고 빨라지고 있으며, 이러한 데이터 기반의 지능디지털 서비스는 더욱 정교화 및 세분화되고 있다. 이러한 서비스의 기반이 되는 데이터의 품질을 평가하기 위하여 국제표준기반의 데이터품질특성의 적용[1] 및 검증의 필요성은 꾸준히 증가하고 있다. 본 연구에서는 데이터품질과 관련한 ISO/IEC 5259-2에서 정의한 품질특성을 기반으로 반정형 및 비정형 데이터의 품질검증을 수행하였다. 이를 통해 데이터 분석(Data Analytics) 및 인공지능 학습용 데이터의 품질 검증을 위하여 국제표준에 기반한 데이터품질특성을 적용한 사례를 제안하고 반정형 및 비정형 데이터의 품질특성 기준 마련에 참고가 될 수 있는 가이드를 제공하고자 한다.

Ⅱ. 본론

ISO/IEC 8000시리즈와 25000시리즈에서 언급하는 정보시스템 기반의 데이터 품질특성[2]에 이어 ISO/IEC 5259시리즈는 데이터 분석(Data Analytics) 및 기계학습(Machine Learning) 기반의 데이터 품질특성을 언급하고 있다. 특히 ISO/IEC 5259-2는 ISO/IEC 25012에서 정의한 15가지의 데이터 품질특성 중 11개의 데이터 품질특성을 적용하고, 데이터 분석 및 기계학습을 위한 9가지 부가적 특성(감사 가능성, 식별 가능성, 유효성, 균형성, 관련성, 다양성, 대표성, 유사성, 적시성)을 정의하고 있다[3].

본 연구에서 데이터품질검증 사례로 '홈트레이닝을 위한 IOT 디바이스 데이터'와 'SNS해시태그 및 대화톡 데이터', '음식 탐지 데이터'를 대상으로 ISO/IEC 5259-2 데이터 품질특성을 적용 및 이해를 시도하였다. 홈트레이닝을 위한 IOT 디바이스 데이터는 트레이닝을 위한 음악을 영상(mp4, jpg)으로 제공하고, 트레이닝 시간별 칼로리 소모량 정보(xlsx)와 고객별 트레이닝 정보(csv)로 구성된 데이터셋이고, SNS해시태그 및 대화톡 데이터는 텍스트 정보(csv, txt)로 구성되었으며, 음식 탐지 데이터는 음식 이미지(jpg)와 라벨링 정보(json), 메타데이터(txt)로 구성되어 있다. 반정형/비정형 데이터 품질검증을 위하여 정보시스템에 의존적인 데이터 품질특성을 제외하고 데이터의 내재적 품질특성들과 데이터 분석 및 기계학습을 위한 품질특성을 고려하여, 본 사례에 적용 가능한 품질특성 7가지를 선정하였다. 선정한 7가지 품질특성별 검증 단위는 data item과 dataset으로 구분하며, 상세 내용은 표 1과 같다.

표 1. 국제표준 기반의 반정형/비정형 데이터 데이터품질특성 7가지

표준	데이터 품질특성	설명	검증 단위	
			data item[1)]	dataset
ISO/IEC 25012	완전성	기본적으로 데이터 값의 not null조건 및 데이터 아이템의 empty여부를 검토하고, 모든 데이터가 정해진 속성값을 갖는지 여부를 점검함	☑	☑
	정확성	(구문적 정확성) 데이터 속성별로 정해진 구조와 형식을 준수하는지 여부를 점검함	☑	☑
		(의미적 정확성) 정해진 도메인에 의미적으로 적합한 데이터 값과 범위를 갖는지 여부를 점검함	☑	☑
ISO/IEC 5259-2	유효성	기계학습의 특정 임무에 필요한 요구사항들을 충족하는지 여부를 점검함	-	☑
	균형성	데이터셋의 모든 측면에 대한 샘플 분포가 고른지를 점검함	-	☑
	다양성	다양한 값 도메인을 갖는 데이터의 분포를 점검함	-	☑
	관련성	기계학습을 통한 결과(output)데이터에서 나타날 것으로 예상/예측되는 데이터인지 여부를 점검함	-	☑
	유사성	분석 및 학습에 필요한 feature측면에서 샘플간의 유사성을 점검함	-	☑

7가지 품질특성 선정 기준은 구조화된 데이터 뿐만 아니라 텍스트, 이미지와

1) data item은 item name, data value, data type로 구성됨. (반정형 데이터 예시)

(json)... "object": 736, "relation": "공간", "position": "위에" ...			
item name	object	relation	position
data value	736	공간	위에
data type	number	string	string

같은 비정형 데이터들에 대하여 정량적 산정 기준을 우선적으로 두었으며, AIHub의 인공지능 학습용 데이터 품질검증 항목과 비교를 위해 데이터의 구문론적, 의미론적 품질특성 및 데이터의 분포 다양성을 적용하였다. 7가지 품질특성 외에 다른 품질특성들은 정성적 검증을 요구하는 특성들(신뢰성, 감사가능성 등)이거나 특정 품질특성(현재성, 식별성, 적시성 등)에 대한 데이터 속성값이 존재하지 않아 검증이 어려운 품질특성들과 정보시스템 의존적 성향을 가지는 품질특성들(접근성, 규정 준수, 이식성 등)은 제외하였다. 반정형 및 비정형 데이터에 적용한 품질특성은 표 2와 같다.

표 2. 반정형 및 비정형 데이터별 품질특성 적용 대상

데이터 품질 특성		검증 대상		
	목적	for Analytics	for Analytics	for ML
	데이터명	홈트레이닝을 위한 IOT디바이스 데이터	SNS해시태그 및 대화록 데이터	음식탐지 데이터
	파일 포맷	xlsx, csv, mp4, jpg	csv, txt	jpg, json, txt
완전성		☑	☑	☑
정확성		☑	☑	☑
		☑	☑	☑
유효성		-	-	☑
균형성		-	-	☑[2]
다양성		☑	☑	
관련성		-	-	
유사성		-	☑	

레이블이 없는 반정형/비정형 데이터의 경우, 메타데이터에 작성된 도메인 또는 범주 값에 의해 분포 다양성 검사가 가능하다. 데이터별 품질특성을 적용하기 위하여 프로퍼티별로 검사조건을 마련하였다. 예를 들어 음식탐지 데이터의 label프로퍼티는 not null 및 필수값 조건과 string타입에 의한 정규식 패턴 조건(^[가-힣0-9-()&]*(?:가-힣)+[가-힣0-9-()&]*)*[^-?,$]$])을 준수해야 한다. 상세 검사 조건 예시는 그림 1과 같다.

순번	데이터 예시	컬럼명	타입	KEY /필수	Not Null	범위 Min	범위 Max	유효한 값	패턴	중복 허용	비고
1	4.1.1	version	string	Y	Y				^[0-9].[0-9].[0-9]$	N	
2		flag	object	Y	N					N	
3		shapes	array	Y	Y					N	
4		{}	object	Y	Y					N	
5	Wuc624Wub80cWuc9c0	label	string	Y	Y				^[가-힣0-9-()&]*(?:[가-힣]+[가-힣0-9-()&]*)*[^-?,$]$	N	
6		points	array	Y	Y					N	
7		[]	array	Y	Y					N	
8	[1063.2,0.0],[1480.8,0.0],[1480.6,223.2],[1063.2,223.2]	$value$	number	Y	Y	0				N	
9	null	group_id	object	Y	N					N	
10	polygon	shape_type	string	Y	Y			"polygon", ""		N	바운딩 박스, 라벨의 존재
11		flags	object	Y	N					N	
12	Wuace4Wub4dc---.jpg	imagePath	string	Y	Y					N	
13		imageData	string	Y	N					N	
14	1,007	imageHeight	number	Y	Y	0				N	
15	1,512	imageWidth	number	Y	Y	0				N	
16		saedTime	string	Y	N					N	
17		saveTime	string	Y	N					N	
18		date	string	Y	N					N	

그림 1. 품질특성 적용을 위한 검사 조건

품질검증 산식은 ISO/IEC 25024 및 5259-2의 품질측정 정량적 산식 기반(100%-오류율)에 따라 품질속성별 전체 대상 데이터 수(A) 대비 해당 품질특성별 오류데이터 수량(B)을 오류율(A/B)로 나타내고, 백분위로 산정한 결과이다. 각 데이터별 검증 결과는 표 3과 같다.

표 3. 반정형 및 비정형 데이터 품질검증 결과

데이터 품질특성		홈트레이닝을 위한 IOT디바이스 데이터	SNS해시태그 및 대화록 데이터	음식탐지 데이터
완전성		100	99.45	100.00
정확성	구문적 정확성	100	99.99	99.99
	의미적 정확성	100	93.29	94.97
유효성		-	-	99.71
균형성		-	-	분포 확인
다양성		100[3]	분포 확인	
관련성		-	-	
유사성		-	빈도, 분포 확인	

품질특성 중 다양성 ,유사성의 경우, 분류 기준이 명확하지 않아 백분위 산정결과 외에 빈도 및 분포확인으로 그 결과를 대체하였다.

검증 결과에 따른 오류 예시는 null값 오류, 검사 구문 규칙에 정의되지 않은 프로퍼티가 추가로 존재하거나, json스키마에서 허용하지 않는 프로퍼티가 존재하는 경우, 특정 프로퍼티의 정규식이 입력문자열과 일치하지 않은 경우 등이다.

Ⅲ. 결론

본 논문에서는 국제표준인 ISO/IEC 5259기반의 데이터 품질특성을 기반으로 품질검증을 수행하였다. AIHub의 인공지능 학습용 데이터 품질검증 항목과 비교를 위해 반정형/비정형 데이터의 특징을 고려하여 완전성, 정확성, 유효성, 균형성, 다양성, 관련성, 유사성 등 총 7가지의 품질특성을 적용하였다.

AIHub의 인공지능 학습용 데이터는 구조화된 데이터셋을 포함한 반정형/비정형 데이터이며, 통계적 다양성, 정확성, 유효성의 3가지 품질특성이 적용되었다[4]. 본 논문에서 적용한 국제표준 데이터 품질특성과 비교하였을 때, AIHub의 통계적 다양성 및 정확성(구문정확성 및 의미정확성)의 검증 방법과 내용은 ISO/IEC 5259에서 정의한 다양성 및 정확성 품질특성과 일치함을 알 수 있었다. AIHub의 품질특성 중 다양성은 ISO/IEC 5259의 균형성, 다양성, 유사성을 통합한 범용적인 품질특성으로 간주할 수 있다. 특히 AIHub의 품질특성 중 AI학습을 통한 평가데이터를 기반으로 결과를 검증하는 품질특성인 유효성은 ISO/IEC 5259의 유효성과 관련성을 통합한 품질특성과 내용적 측면에서 동일하나, 그 검증 방식은 AI모델 활용 여부에 의해 상이함을 알 수 있었다.

본 논문의 적용 사례를 통해 AIHub의 반정형/비정형 데이터의 품질특성과 국제 표준 ISO/IEC 5259에 기반한 품질특성과의 통합 및 적용이 가능함을 확인하였다. 반정형 및 비정형 데이터의 체계적인 품질관리를 위해 국제표준 관점의 품질 기준과 가이드라인이 필요하며[5] 본 사례를 기반으로 더 통합적인 품질관리 연구가 이루어지길 기대한다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 초거대AI 확산 생태계 조성 사업(2100-2131-305, 2024년도 초거대AI 확산 생태계 조성 사업)에 의해서 수행되었습니다.

참 고 문 헌

[1] 현욱진 외 2인, 데이터 품질인증 심사체계에 대한 연구, 2023

[2] ISO/IEC 25024:2015(en), Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality, (https://www.iso.org).

[3] ISO/IEC DIS 5259-2(en), Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 2: Data quality measures, (https://www.iso.org).

[4] 한국지능정보사회진흥원, 인공지능 학습용 데이터 품질관리 가이드라인 및 구축안내서 v3.0, 2023, (https://www.aihub.or.kr/).

[5] 한국데이터산업진흥원, 데이터산업 백서 vol 26, 2023

2) 반정형 및 비정형 데이터의 균형성, 다양성, 관련성, 유사성은 객체의 레이블을 대상으로 품질특성을 검증하는 내용이므로 서로 연관된 품질특성으로 간주하여 검증함

3) 데이터의 업무규칙에 제시된 음악 장르 분류표를 기준으로 다양성 검증을 수행함.