ISO/IEC 25024 품질특성을 활용한 구조화된 형식(csv)의 파일데이터 품질검증 방법 및 적용 사례

박경은, 박정하, 이진용, 이상복

한국정보통신기술협회

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

Method and Application Case of Quality Verification of File Data in Structured Format(csv) Using ISO/IEC 25024 Quality Characteristics

Park Kyung Eun, Park Jeong Ha, Lee Jin Yong, Lee Sang Bok

Telecommunications Technology Association

요 약

본 논문은 구조화된 형식(csv)의 파일데이터의 품질을 정량적으로 평가하기 위해 국제표준 ISO/IEC 25024의 품질특성 중 일부를 활용한 품질검증 방법을 제안한다. 구조화된 형식(csv)의 파일데이터에 적용할 수 있는 국제표준의 품질특성을 구분하고, 하위 세부 품질특성 별 검증 방법을 정의한다. 그리고 해당 세부 품질특성을 활용하여 IoT 센서 기반 도심 모니터링 데이터의 품질검증을 수행한 사례를 소개한다.

Ⅰ. 서 론

최근 데이터는 국내외 산업에서 수집, 가공 및 거래 등 다양한 형태로 활용되고 있으며, 데이터 산업 시장 규모는 지속적인 성장세를 이어가고 있다.[1] 이와 더불어, 데이터의 품질관리 필요성 또한 중요한 요소가 되고 있으며, 데이터 품질을 평가하기 위한 국제표준(ISO/IEC 2502n 시리즈, ISO 8000 시리즈 등)이 활용되고 있다. 이때 데이터 유형에 따라 품질특성(Quality Characteristic)을 적용할 수 있는 범주(Category)가 달라지며, 유형별로 적용 가능 특성을 구분 및 적용하는 연구가 필요하다. 국내 사례에는 에너지 관리 정보시스템에 ISO/IEC 25024의 품질특성을 적용해 데이터 품질 테스트를 진행한 연구가 있다[2]. 본 논문에서는 구조화된 형식(csv)의 파일데이터 품질검증에 적용할 수 있는 ISO/IEC 25024의 품질특성을 구분 및 정의하고, 실제 데이터에 적용하여 품질검증을 수행한 사례를 소개한다.

Ⅱ. 본 론

1. ISO/IEC 25024 개요 및 품질특성

1.1 개요

ISO/IEC 25024는 데이터 품질 측정 요구에 따라 만들어진 국제표준으로, 데이터 품질을 정량적으로 측정하기 위한 방법을 정의한다. 이는 정보시스템(Legacy Information System), 데이터 웨어하우스(Data Warehouse) 및 웹(Web) 등 다양한 규격에 존재하는 데이터에 범용적으로 적용할 수 있는 표준이다. 측정 대상은 데이터 생명주기에 따라 생성 및 관리되는 객체로, 데이터 파일, 데이터 모델, 아키텍처 및 DBMS 등이 해당된다.[3] 본 논문에서는 이러한 측정 대상 중 구조화된 형식(csv)에 해당하는 데이터 파일로 정의한다.

1.2 품질특성

데이터 품질에 영향을 미치는 데이터 품질 속성(Attributes)의 범주를 품질특성이란 용어로 정의하고, 15개의 대분류 및 63개의 소분류로 구분한다. 대분류는 정확성, 완전성, 일관성, 신뢰성, 현재성, 접근성, 준수성, 기밀성, 효율성, 정밀성, 추적성, 이해성, 가용성, 이식성 및 복구성으로 구분한다. 각 대분류 품질특성 하위의 세부 특성을 정의하고, 이에 대한 설명 및 측정함수 등을 제공한다. 특히, 모든 데이터에 대해 세부 특성을 제공할 수 없으므로, 데이터에 적용되는 기술에 맞도록 새롭게 정의 및 개선하여 활용하도록 권장하고 있다.[3]

2. 센서 데이터 품질검증 방법

2.1 품질검증 대상 정의

본 논문에서 품질검증 대상은 ISO/IEC 25024 품질특성을 적용하기 위해 구조화된 형식(csv)으로 수집할 수 있는 센서 데이터로 정의한다. 센서 데이터는 센서 장치들로부터 수집되는 수치들의 집합으로[4], 센서의 종류에 따라 수집된 데이터의 형태가 달라진다. 특히 IoT 센서로부터 원시(Raw) 데이터를 수집할 때 구조화된 형태로 실시간 수집한다는 특성이 존재한다. 따라서 구조화된 형식(csv)에 가장 적합한 센서 데이터를 품질검증 대상으로 정의한다.

2.2 품질검증 방법

센서 데이터 품질검증에 적용할 수 있는 세부 품질특성을 6가지로 구분하였고, 이는 표 1과 같다. 정확성(Accuracy)은 데이터가 속성별로 정의된 규칙에 대해 올바른 값(Value)을 나타내고 있는지를 확인하는 특성이며, 완전성(Completeness)은 데이터 파일 관점과 관련된 예상값을 갖는 정도를 확인하는 특성이다. 즉, 데이터가 파일 단위로 얼마나 완전하게 수집되었는지, 그리고 속성 단위로 얼마나 정확하게 수집되었는지를 정량적으로 평가할 수 있다.

표 1. 센서 데이터 품질검증에 적용한 세부 품질특성

No.	품질특성	세부 품질특성	설명
1	정확성	구문 데이터 정확성	도메인에 정의된 규칙에 대한 데이터 값의 근접성 비율
2		의미 데이터 정확성	의미론적 측면(유효한 값)에서 데이터 값의 정확성 비율
3		데이터 범위 정확성	요구되는 지정 범위에 맞게 포함되어있는 데이터 값의 비율
4	완전성	레코드 완전성	데이터 파일 내 레코드의 완전성 비율
5		데이터 값 완전성	데이터 파일 내 데이터 값이 예상되는 값을 만족하는 비율
6		데이터 파일 완전성	데이터 파일 내 예상 레코드들의 완전성 비율

세부 품질특성 중 '구문 데이터 정확성'은 속성별 구문 규칙(패턴 등) 및 타입에 어긋나는 값이 있는지 검증한다. 두 번째 '의미 데이터 정확성'은 속성별로 수집되어야 하는 유효한 값이 아닌 값이 있는지 검증한다. 세 번째 '데이터 범위 정확성'은 속성별로 최대/최소 범위를 벗어나는 값이 있는지 검증한다. 네 번째 '레코드 완전성'은 파일 내 비어있는 레코드가 있는지 검증한다. 다섯 번째 '데이터 값 완전성'은 파일 내 null을 허용하지 않는 속성에서 null 값이 있는지 검증한다. 마지막 '데이터 파일 완전성'은 수집되어야 하는 예상 레코드 수와 실제 레코드 수를 비교한다. 이렇게 6가지 세부 품질특성을 얼마나 만족하는지 측정하는 함수를 공통으로 정의하자면 다음과 같다.

$$X(\%) = \frac{A}{B} \times 100$$

A: 해당 특성(구문 규칙, 유효값, 범위, not null 등)을 만족하는 값의 수
B: 해당 특성의 전체 값(또는 레코드)의 수

3. 적용 사례

품질검증 대상 데이터는 이동 중인 차량에 IoT 센서를 장착하여 실시간으로 도심 내 교통/도로/주변 상황 등을 수집한 도심 모니터링 데이터로, 수집 시간, 위/경도, 온/습도 및 운행속도 등의 정보를 제공하여 인공지능 기반 시공간 데이터 분석/추정/예측 기술에 활용할 수 있는 데이터다. 이러한 센서 데이터를 '2절'에서 언급한 세부 품질특성 6개를 적용하여 데이터의 품질을 정량적으로 검증해 보았다.

먼저, 품질특성 중 '정확성'은 데이터 속성별로 준수해야 할 규칙을 정의한 후 해당 규칙을 얼마나 만족하고 있는지를 검증 도구를 활용하여 확인하였다. 즉, 속성별 데이터 타입(string, number 등), 준수해야 할 규칙(패턴, 유효값, 범위)을 정의하고[5] 이에 따라 검증 대상이 되는 세부 품질특성을 구분하였다. 예를 들면, '위도'란 속성은 대한민국 위도 범위인 33도에서 43도 사이의 값만 정상 수집된 값으로 정의하고, 해당 범위를 벗어난 값이 있는지를 검사하여 품질을 정량적으로 측정하였다.

그리고 품질특성 중 '완전성'은 파일 단위로 비어있는 레코드 유무, null 값 유무 및 실제 레코드 수를 검증 도구를 활용하여 확인하였다. 예를 들면, 전체 속성이 null을 허용하지 않는다고 정의하고, 해당 파일에 null 값이 존재하는지를 검사하여 데이터가 얼마나 완전하게 수집되었는지 정량적으로 측정하였다. 이렇게 데이터 속성별, 파일별 준수해야 할 특성을 정의하는 분석 과정을 거친 결과는 표 2와 같고, 실제 준수 여부는 도구로 검사하여 품질을 평가하였다.

표 2. 품질특성 정확성 관련 속성 분석 예시

속성 명	타입	준수해야 할 규칙		세부 품질특성
시간	string	패턴	YYYY-MM-DD HH:MM:SS	구문 데이터 정확성
지오해시	number	패턴	8자리 문자열	구문 데이터 정확성
위도	number	범위	33 ~ 43	데이터 범위 정확성
경도	number	범위	124 ~ 132	데이터 범위 정확성
시	string	유효값	서울특별시	의미 데이터 정확성
운행속도	number	범위	0 ~ 260	데이터 범위 정확성
온도	number	범위	0 ~ 100	데이터 범위 정확성
습도	number	범위	0 ~ 100	데이터 범위 정확성
유동인구	number	범위	0 ~	데이터 범위 정확성
조도	number	범위	0 ~ 70,000	데이터 범위 정확성

Ⅲ. 결 론

본 논문은 구조화된 형식(csv)의 파일데이터를 대상으로 국제표준 ISO/IEC 25024의 품질특성 일부를 활용한 품질검증 방법을 제안하였다. 또한, IoT 센서로 수집된 도심 모니터링 데이터에 해당 품질검증 방법을 적용하여 품질을 정량적으로 평가하였다. 이때 파일데이터는 관계형 데이터베이스의 스키마(Schema) 등과 같은 논리적 구조가 없어 비교적 복잡도가 적었고[6], 품질검증 시 적용할 수 있는 국제표준의 품질특성 범위가 적다는 한계점이 존재하였다. 따라서 향후 파일데이터의 또 다른 품질검증 방법과 더불어 다양한 구조의 파일데이터 품질검증 방법에 관해 연구할 계획이다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] 한국데이터산업진흥원. "2023 데이터산업 백서." pp. 76-81

[2] 주승환, 서희석. "산업단지 에너지 관리 시스템의 데이터 품질 테스트 방법." 한국콘텐츠학회, 2023.04

[3] ISO/IEC 25024:2015. "Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality." 2015.

[4] 정보통신단체표준(TTAS). "외부 환경 센서를 연결한 증강 현실 글래스의 데이터 형식." TTAK.KO-10.1476, 2023.12

[5] 과학기술정보통신부, 한국지능정보사회진흥원, 한국정보통신기술협회. "인공지능 학습용 데이터 품질관리 가이드라인 및 구축 안내서 v3.0 - 제1권 품질관리 가이드라인 v3.0." 2023.

[6] 정보통신단체표준(TTAS). "데이터 품질관리 프레임워크." TTAK.KO-10.0378, 2009.12