Data of Geology, Ecology, Oceanography, Space Science and Polar Science [3.3] : 35-40 (2021)
DOI https://dx.doi.org/10.22761/DJ2021.3.3.004

[Space Science] Opinion

인공지능 데이터의 품질 관리 및 검증 현황

김명언 · 안효정*
한국항공우주연구원 미래혁신연구센터, 대전 34133, 대한민국

Quality Control and Verification of Artificial Intelligence Data

Myeung Un Kim · Hyojung Ahn*
Future Innovation Research Center, Korea Aerospace Research Institute, Daejeon 34133, Republic of Korea

Received: 9 September 2021, Revised: 29 September 2021, Accepted: 30 September 2021

요약문 높은 품질의 인공지능 데이터는 인공지능 모델을 개발하는데 있어 정확한 정보를 제공함으로써 모델의 효율성을 높이는 데 일조한다. 반면 품질이 낮을 경우 상호 데이터 간의 정보 불일치로 인하여 연구의 방향성을 해칠 수 있다. 이와 같이 인공지능 기반 모델 개발 연구의 질을 높이기 위해서 연구에 활용되는 데이터의 높은 품질을 확보하기 위해 체계적인 관리와 인증이 필요하다. 현재 우리나라의 데이터 품질 인증제도 뿐 아니라 미국의 데이터 품질 법, 국제 표준화 기구 ISO 8000 시리즈, 유엔의 빅데이터 품질 검증 기준 등 데이터 품질 관리에 대한 지침을 가지고 있다. 본 연구에서는 데이터 품질 관리 현황을 파악하고, 이에 대한 시사점을 고찰한다.

주요어: 데이터 품질 관리, 데이터 품질 검증, 인공지능 데이터

Abstract High-quality artificial intelligence (AI) data provides accurate information for developing AI models. These results in increasing the efficiency of the model. On the other hand, if low-quality data is used, it may adversely affect the development of AI models. To improve the quality of our research, we need to increase the quality of AI data. This is possible through systematic quality control and verification of the data. Currently, there are various guidelines such as the data quality act of the US, the ISO 8000 series of the International Organization for Standardization, and the Big Data quality verification standard of the United Nations, as well as Korea's database quality certification. In this study, the current status of data quality management is identified and its implications are considered.

Keywords: Data quality control, data quality verification, artificial intelligence data

*Corresponding author: hjahn@kari.re.kr

1. 서론

높은 품질의 인공지능 데이터는 인공지능 모델을 개발하는데 있어 정확한 정보를 제공함으로써 모델의 효율성을 높이는 데 일조한다. 반면 품질이 낮을 경우 상호 데이터 간의 정보 불일치로 인하여 연구의 방향성을 해칠 수 있다. 연구의 질을 높이기 위해서 연구에 활용되는 데이터의 높은 품질을 확보하기 위해 체계적인 관리와 인증이 필요하다. 현재 우리나라뿐 아니라 다양한 국가 및 기관에서 데이터 품질 관리에 대한 지침을 가지고 있다. 본 연구에서는 데이터 품질 관리에 대한 현황을 파악하고 시사점을 고찰하고자 한다.

2. 국가 및 기관의 데이터 품질 관리 현황

2-1. 한국의 데이터 품질 인증제도

한국데이터산업진흥원 부설 데이터 품질관리 인증센터에서 2006년 2월부터 품질 인증제도(DQC; Database Quality Certification)를 시행하고 있다 (Jang et al., 2018). 범국가적 데이터 품질 제고 및 고도화를 위해 데이터(Data value), 데이터 관리(Data management), 데이터 보안(Data security) 등을 심사 및 인증한다. Figure 1은 각 항목별 평가 내용을 나타낸다. 데이터 인증은 데이터 자체에 대한 품질 영향 요소 전반을 심사 및 인증한다. 데이터 관리 인증은 행정 및 업무 지원, 의사 결정 및 정책 지원 등을 목적으로 운영되는 시스템에 대한 데이터 관리 수준을 심사 및 인증한다. 데이터 보안 인증은 데이터 보안에 대한 기술 요소 전반을 심사 및 인증한다.

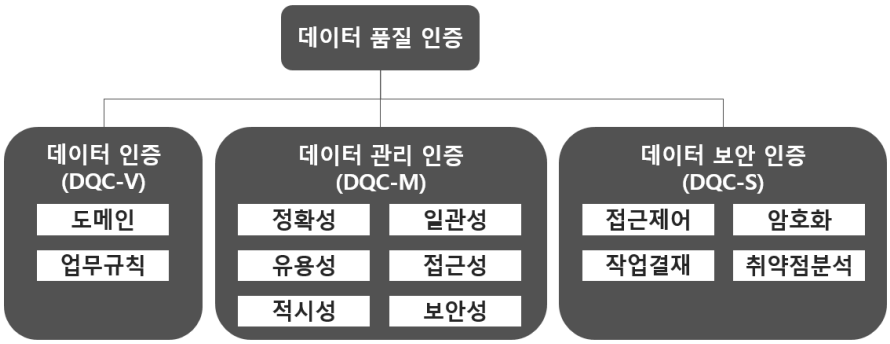

Figure 1. Data Quality Certification Category

2-2. 미국의 데이터 품질 법

미국에서는 연방 정부의 정보 품질 확보를 위해 2000년대 초반에 데이터 품질 법(Data quality act)을 제정해 각 산하기관 등에 데이터 품질 가이드라인 수립과 이행을 권고하고 있다 (FindLaw Attorney Writers, 2017). 데이터 품질 법에서는 3가지 요소를 기반으로 데이터 품질을 관리한다. 첫 번째는 이용 편의성으로, 제공되는 데이터는 데이터를 이용하는 사람이 쉽게 구할 수 있으며 이용 만족도가 높아야 한다. 두 번째는 안전성 및 신뢰성으로, 데이터에 대한 기술적, 관리적, 물리적인 사용자의 접근 권한 관리를 통하여 비 권한자로부터 데이터를 보호해야 한다. 세 번째는 공익성으로, 데이터의 근거와 출처는 분명해야 하며 편향되지 않아야 한다. 데이터 품질 법은 데이터 품질 관리 활동을 수행하며 발견되는 다양한 개선사항을 탄력적으로 반영하며 지침을 유지하고 있다.

2-3. 국제 표준화 기구 ISO 8000 시리즈

국제 표준화 기구(ISO; International Organization for Standardization)에서는 데이터 품질 기준에 관련하여 ISO 8000 시리즈를 제공하고 있다 (Schwarzenbach J, 2016). ISO 8000 시리즈는 데이터 품질 및 엔터프라이즈 마스터 데이터에 대한 글로벌 표준으로, 2011년 12월 한국데이터산업진흥원에서 개발한 데이터 품질 관리 프레임워크

(ISO 8000:150)도 포함되어 있다. Table 1은 ISO 8000:150의 데이터 품질 관리 프레임워크를 나타낸다. 여기서 Data operations은 데이터의 품질과 데이터 사용에 영향을 미치는 요소에 대한 과정을 의미하고, Data quality monitoring은 데이터 품질의 수준을 평가하는 체계적인 접근에 대한 과정을 의미하며, Data quality improvement는 데이터 오류의 근본적인 원인을 없애고 바로잡는 과정을 의미한다.

Process / Role	Data operations	Data quality monitoring	Data quality improvement
Data manager	Data architecture management	Data quality planning	Data stewardship / flow management
Data administrator	Data design	Data quality criteria setup	Data error causes analysis
Data Technician	Data processing	Data quality measurement	Data error correction

Table 1. Framework for Master Data Quality Management (ISO 8000:150)

2-4. 유엔의 빅데이터 품질 검증 기준

유엔 유럽 경제 위원회(UNECE; United Nations Economic Commission for Europe)는 국가 통계 작성 시 빅데이터 활용을 권고했으며, 2014년에 'The Role of Big Data in the Modernization of Statistical Production' 프로젝트를 통해 Figure 2와 같이 3단계로 나눠진 빅데이터 품질 기준을 제시하였다 (UNECE Big Data Quality Task Team, 2014; Jin JH et al., 2016).

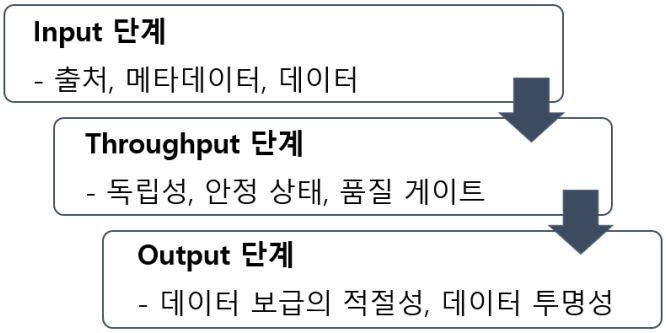

Figure 2. 3 Steps of Big Data Quality Verification

첫 번째는 Input 단계로, 데이터 수집과 관련하며 크게 출처, 메타데이터, 데이터 영역으로 나뉜다. 출처에 관한 품질 검증 요소는 제도/사업적 환경, 개인 정보 보호/보안 등이 있다. 메타데이터에 관한 품질 검증 요소는 복잡성, 완전성, 유용성, 시간 관련 요소, 연계 가능성, 일관성, 타당성 등이 있다. 데이터에 관한 품질 검증 요소는 정확성, 연계 가능성, 일관성, 타당성 등이 있다. 두 번째는 수집된 데이터를 가공하고 분석하는 Throughput 단계로, 데이터 품질에 대한 시스템의 독립성, 데이터 세트의 안정 상태, 품질 게이트 등 총 3가지 영역에 대한 검증 원칙을 제시한다. 세 번째 Output 단계에서는 데이터 소비자에게 전달되는 최종 데이터의 품질을 평가하며 데이터 보급의 적절성과 데이터 투명성에 초점을 두고 품질을 평가한다.

3. 데이터 품질 관리 및 검증 연구

국가 및 기관들의 데이터 품질 관리 지침뿐만 아니라, Table 2와 같이 데이터 품질을 관리 및 검증과 관련된 다양한 연구가 진행되었다. 본 절에서는 아래 연구들에 대한 정리를 하고자 한다.

연구 제목	연구 연도
Automating Large-Scale Data Quality Verification (Schelter et al., 2018)	2018
Visual Analytics of Flight Trajectories for Uncovering Decision Making Strategies (Andrienko G et al., 2018)	2018
Beholder Automated Data Validation in Flight Test (M. Arévalo Nogales, 2020)	2020
Aviation Abstract Quality Model Engineering Report (McMillan A, 2018)	2018

Table 2. Study on Data Quality Assessment and Evaluation

(Schelter et al., 2018) 에서는 대규모의 데이터 셋으로 확장하고, 생산 활용 사례 요건을 충족하는 규모에 맞는 데이터에 대한 품질 검증을 자동화하는 시스템을 제시한다. 또한, 데이터 품질 자동화 시스템 API(Application Programming Interface)를 제시한다. API는 사용자들이 공통 품질 제약조건을 지정하면 이 조건을 데이터의 품질 검증 지표로 변환하여, 제약조건을 사용자 정의 검증 코드와 결합함으로써 데이터에 대한 단위 테스트를 가능하게 한다. Table 3은 데이터 품질 자동화 시스템에서 지원하는 데이터 품질 측정 기준을 나타낸다.

Metric	Semantic
Dimension completeness	
Completeness	Fraction of non-missing values in a column
Dimension consistency	
Size	Number of records
Compliance	Ratio of columns matching predicate
Uniqueness	Unique value ratio in a column
Distinctness	Unique row ratio in a column
Value Range	Value range verification for a column
Data Type	Data type inference for a column
Predictability	Predictability of values in a column

Table 3. Metrics in Data Quality Automation System

(Andrienko G et al., 2018) 에서는 비행 궤적 데이터 품질 및 데이터 분석 워크플로우의 평가를 지원하기 위한 일련의 시각적 분석 기법을 제안한다. 이때의 비행 궤도 데이터는 유럽에서의 항법 충전 구역(Navigation Charging Zone)과 그 구역의 단위 속도에 대한 데이터이다. 비행 데이터는 기관마다 수집되므로 각각의 품질과 해상도가 다르기 때문에 데이터 품질 조사가 필요하다. 이를 위해 시각적인 접근 방법을 통해서 계산적 데이터 처리를 결합해 크고 정보에 입각한 의사결정을 가능하게 한다. 시각적 분석을 통해 데이터 작성 및 분석 비용을 절감하고 시간적 효율성을 가져 왔다. 궤적 데이터 품질의 문제는 누락된 데이터와 중복된 데이터이다. 일부 누

락된 궤적 데이터의 경우 육안 검사를 통하여 데이터 품질 관리가 수행될 수도 있다. 대부분의 경우는 구역마다 다른 요일과 시간 동안의 비행 밀도를 고려하여 주어진 위치 및 시간 간격에 대한 예상 값에서 실질적으로 벗어나는 경우를 통해 누락된 데이터를 식별할 수 있다. 중복된 데이터는 비행 식별, 타임스탬프, 위치 및 추가 속성 등의 데이터가 중복된 것으로 후보 오류 감지를 위한 계산 처리, 오류가 간헐적 또는 체계적인지 파악하여 오류의 논리를 밝히는 육안 검사, 오류 수정을 위한 계산 방법을 사용하여 중복된 궤적 데이터를 식별할 수 있다.

(M. Arévalo Nogales, 2020) 에서는 항공기의 복잡한 데이터 수집 시스템에 대한 데이터 검증 및 정리하기 위하여 'Beholder'라는 맞춤형 해결책을 개발하였다. 규칙 분석 및 이상 징후 탐지를 위하여 데이터 검증 이후에 데이터 품질 또한 검사한다. 이상 징후, 이상한 작업 조건, 한계나 임계값이 초과된 상황 등을 감지하여 시스템 운영에 대한 데이터 품질 수준을 규정하는 규칙 집합을 작성할 수 있다. 이 소프트웨어는 규칙의 형태로 사용자 지식을 활용하여 데이터 흐름에 따라 자동으로 로드되고 실행되는 검증 코드로 만든다. 자동화된 프로세스 실행 시스템을 지원하기 위해 현재의 데이터 서버 기술을 기반으로 한다. 최종 사용자가 사용할 수 있도록 자동화된 시스템을 만드는 것은 비행 시험 데이터 유효성 검사를 개선할 뿐만 아니라 사용자가 시스템의 획득 작업에 참여할 수 있도록 도와준다. 데이터가 자동으로 처리되기 때문에 검증에 많은 양의 테스트를 사용할 수 있으며 향후 빅데이터와 머신러닝 기반의 새로운 검증 기법에 적용할 수 있다.

(McMillan A, 2018) 에서는 A3C (Accuracy, Currency, Completeness and Consistency) 품질 프레임 워크와 여러 표준을 기반으로 항공 도메인에 대한 추상 품질 모델 (AQM; Abstract Quality Model)을 제시한다. AQM은 SDCM (Service Description Conceptual Model)과 결합하면 비행 서비스에 대해 적용할 수 있는데 이를 통해 비행장 구조 정보, 비행 정보, 날씨 정보 등에 대하여 사용자에게 최상의 결정을 내리는 데 필요한 정보를 제공할 수 있다.

4. 결론 및 토의

데이터의 품질은 인공지능 모델을 개발하는데 다방면으로 중대한 영향력을 끼친다. 품질이 낮은 데이터는 곧바로 인공지능 모델 성능의 저하로 연결되기 때문에, 높은 품질의 데이터를 생산하고 관리하는 과정의 엄밀성을 위해 최선의 노력을 다해야 한다. 앞에서 언급한 여러 국가 및 기관의 데이터 품질 관리 및 검증 사례와 관련 연구들과 같이, 인공지능 연구를 위한 데이터의 품질 관리 및 검증에도 적절한 기준을 확립할 필요가 있다.

5. 사사

본 연구는 한국항공우주연구원 주요과제 FR21L01 "인공지능 및 금속3D 프린팅 기술기반 항공우주 핵심기술 연구"의 지원으로 수행되었다.

6. 참고문헌

Andrienko G, Andrienko N, Fuchs G, Scarlatti D, Cordero Garcia J M, Vouros G A, Herranz R and Marcos R (2018) Visual analytics of flight trajectories for uncovering decision making strategies, Paper presented at the Eighth SESAR Innovation Days, Salzburg, Austria

FindLaw Attorney Writers (2017) Federal Agencies Subject to Data Quality Act, https://corporate.findlaw.com/law-library/federal-agencies-subject-to-data-quality-act.html

Jang WJ, Kim DS, Min DG (2018) An improvement plan of information system operational audit for database operational management based on data quality, Journal of Service Research and Studies, Vol.8, No.2, pp41-65

Jin JH, Ko KJ (2016) UN's big data quality criteria and their implications: focusing on national statistics, Health and

Welfare Policy Forum, No.241, pp110-121
M. Arévalo Nogales (2020) Beholder automated data validation in flight test, ettc2020 - European Test and Telemetry Conference, pp170-177
McMillan A, Meek S (2018) OGC Testbed-13: Aviation abstract quality model engineering report, https://docs.ogc.org/per/17-032r2.html
Schelter S, Lange D, Schmidt P, Celikel M, Biessmann F, Grafberger A (2018) Automating large-scale data quality verification. Proceedings of the VLDB Endowment, Vol.11, No.12, pp1781-1794
Schwarzenbach J (2016) ISO 8000 Part 150 A framework for data governance, DPA, UK
UNECE Big Data Quality Task Team (2014) A suggested Framework for the quality of big data