논문 2023-4-9 http://dx.doi.org/10.29056/jsav.2023.12.09

빅데이터를 위한 데이터 품질 평가 방법에 대한 연구

최옥주*, 김유경**†

A Survey of Data Quality Assessment Methods for Big Data

Okjoo Choi*, Yukyong Kim**†

요 약

빅데이터 분석이나 머신러닝과 같은 데이터 기반의 정보기술 분야에서는 데이터 품질이 매우 중요하다. 고성능의 데이터 분석 알고리즘이나 머신러닝 학습 모델을 사용하더라도 입력 데이터의 품질이 보장되지 않는다면, 그 결과물은 신뢰할 수 없기 때문이다. 따라서 빅데이터를 활용하고 분석하기 위해서는 방대하고 복잡한 데이터들로부터 고품질의 데이터를 추출할 수 있어야 한다. 본 논문에서는 고품질을 보장하는 빅데이터를 위한 품질 평가 방법에 대해 고찰해본다. 빅데이터에 대한 데이터 라이프사이클을 살펴보고, 데이터 품질 요소에 대한 국제표준화 동향과 함께 라이프 사이클에 따라 고려되어야 하는 데이터 품질 특성을 정의한다. 또한 빅데이터 관련 데이터 품질 평가에 대한 기존의 주요 연구들을 비교 분석하고, 이를 바탕으로 빅데이터 품질 평가에 필요한 요소들을 살펴본다. 정의된 요소들을 반영하여 목적기반 데이터 품질지표를 이용한 빅데이터 품질 평가 프로세스를 제안한다. 제안된 프로세스는 향후 빅데이터 품질을 평가할 수 있는 평가 지표 개발과 통합된 평가 프레임워크 개발의 기초자료로 활용될 것으로 기대한다.

Abstract

Data quality is very important in data-based information technologies such as big data analysis or machine learning. Even if high-performance data analysis algorithms or machine learning models are used, if the quality of the input data is not guaranteed, the results cannot be trusted. Therefore, in order to utilize and analyze big data, it is necessary to be able to extract high-quality data from massive and complex data. In this paper, we consider quality evaluation methods for big data that guarantee high quality. We examine international standardization trends for data life cycle and data quality factors for big data and define data quality characteristics that should be considered according to the big data life cycle. In addition, we compare and analyze existing major studies on data quality evaluation related to big data, and based on the result, we examine the elements necessary for big data quality evaluation. Based on the defined elements, we propose a big data quality evaluation process using goal-driven data quality metric. In the future, we expect that the proposed process will be used to develop evaluation indicators that can evaluate big data quality and to develop an integrated evaluation framework.

한글키워드 : 빅데이터, 데이터 품질, 데이터 품질 평가, 데이터 품질 메트릭, 데이터 품질 평가 프레임워크

keywords : Big data, Data quality, Data quality assessment, Data quality metric, Data quality assessment framework

* 강원대학교 융합보안대학원 빅데이터-융합보안사업단
** 숙명여자대학교 기초공학부
† 교신저자: 김유경(ykim.be@sookmyung.ac.kr)
접수일자: 2023.12.02. 심사완료: 2023.12.12.
게재확정: 2023.12.20.

1. 서 론

최근 클라우드, IoT, SNS 등 정보기술 환경의 발전으로 급격하게 데이터가 증가하여 제타바이트(Zetta Byte, ZB) 단위의 데이터가 생성되고 있다. 이러한 현상은 학계 뿐 아니라 정부, 다양한 산업 도메인에서 빅데이터에 대한 중요성을 인식하고 있고 다양한 용도로 빅데이터를 수집하고 분석, 활용하고 있다. 빅데이터 활용이 증가하면서 빅데이터 분석이나 머신러닝과 같은 데이터 기반의 정보기술 분야에서는 데이터 품질이 중요한 화두가 되고 있다. 고 성능의 데이터 분석 알고리즘이나 머신러닝 학습 모델을 사용하더라도 저품질의 입력 데이터를 사용하는 경우 해당 결과물은 신뢰할 수 없기 때문이다[1]. 최근 MIT의 연구에서는 머신러닝의 데이터 셋의 대부분은 라벨링 오류로 인한 데이터 셋 품질 문제가 발생한다고 보고하였다[2]. 이러한 관점에서 빅데이터의 원본 데이터의 품질 확보가 필수적이라고 할 수 있다. 이 외에도 많은 연구에서 데이터 분석하는 전체 과정에서 데이터 품질 문제를 제기하고 있다.

따라서 데이터 수집 단계부터 데이터를 분석하고, 학습하거나 시각화하기까지의 데이터 생애주기에 따라 데이터 품질에 대한 관리가 이루어져야 하고, 데이터 품질 개선에 대한 노력이 수반되어야만 한다. 즉, 빅데이터의 수집 저장부터 처리, 분석, 예측 활용까지 전체 주기에 걸쳐 부가가치를 창출하기 위해서는 고품질의 데이터가 중요한 전체 조건이 된다. 현재 고품질을 보장하는 빅데이터를 위한 품질 평가 방법이 연구 분야나 연구자마다 차이가 있어서 포괄적인 방법론이 여전히 부족하나 점차 표준화된 품질 평가가 일반화되고 있는 상황이다.

본 논문에서는 빅데이터와 머신러닝과 같은 데이터 기반 정보기술 분야에서 다루는 데이터에 대한 데이터 라이프사이클을 정의해 보고, 데이터 품질 요소에 대한 국제표준화 동향과 함께 라이프사이클에 따라 고려되어야 하는 데이터 품질 특성을 살펴본다. 또한 빅데이터 관련 데이터 품질 평가에 대한 기존의 주요 연구들을 비교분석하고, 목적기반 데이터 품질지표를 이용한 빅데이터 품질 평가 프로세스를 제안한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 빅데이터의 특징과 라이프사이클에 대해 정의하고, 3장에서는 빅데이터 관련 국제표준화 동향에 대해 살펴본다. 4장에서는 빅데이터 품질평가 방법에 대한 연구들을 비교 분석해 본다. 5장에서는 빅데이터 품질 평가 프로세스를 제안하고, 마지막으로 6장에서 빅데이터 품질 평가에 대한 제언과 결론 및 향후 연구과제를 기술한다.

2. 관련 연구

2.1 빅데이터 특징

데이터 수집 및 저장, 관리 관점에서 빅데이터 기술의 등장은 '데이터 범용성'의 기반을 제공했다[1]. 서로 다른 목적을 가지고 생성된 이질적인 데이터들이 다양한 분야에서 광범위하게 사용되기 시작했고, 빅데이터는 기존의 데이터와 비교해 너무 방대하여 기존의 방법이나 도구로 수집 및 저장, 분석이 어려운 정형 및 비정형 데이터라고 정의할 수 있는데, 빅데이터는 단순히 양이 많은 것뿐만 아니라 4V라고 불리는 기존 데이터와 구분되는 특징이 있다: Volumne(볼륨), Velocity(속도), Variety(다양성), Value(가치) [3]. 이렇게 4V의 특징을 갖고 있는 빅데이터를 활용하고 분석하기 위해서는 방대하고 가변적이며 복잡한 데이터 셋에서 활용가능한 고품질의 데이터를 추출하는 것이 중요하다.

2.2 데이터 품질

품질이라는 개념은 다양한 정의가 있지만 ISO에서는 사용자의 명시적이고 암시적인 요구 사항을 충족시키는 모든 특성으로 정의한다. 데이터 품질이라는 용어가 보편적이고 정확한 정의는 없지만 가장 널리 인정되는 정의는 MIT Rechard Wang 교수의 “Fitness for use(사용 적합성)”이다[4]. 다시 말해 데이터의 품질은 사용자가 기대할 수 있는 용도에 적합한 데이터이다.

2.3 빅데이터 라이프사이클과 데이터 품질

빅데이터 분석의 절차는 일반적으로 수집, 저장, 추출 및 전처리, 분석, 시각화 단계로 구분된다. 데이터의 흐름과 라이프사이클 관점에서 정의해 보면, 그림 1과 같이 빅데이터 분석 절차에 따라 데이터의 생성, 저장, 준비, 은퇴/삭제, 공유로 구분할 수 있다[5].

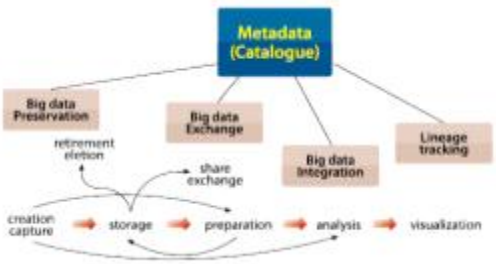

그림 1. 빅데이터 라이프사이클과 메타데이터[5]
Fig. 1. Big data lifecycle and metadata

또한 [6]에서는 빅데이터 라이프사이클은 다양한 소스로부터 데이터를 수집하고, 저장 및 분석, 결과를 시각화하는 과정을 구성된다고 기술하였다. 기존의 연구들을 비교 분석하고, 그 결과로 빅데이터의 데이터 셋의 라이프사이클은 수집(collection), 표현(annotation), 저장(storage), 테스트(testing), 소멸(destruction) 등으로 구분하였다. 이 중 데이터 수집 및 표현 단계는 데이터 셋의 본질적인 품질에 큰 영향을 받게되며, 데이터 테스트 단계에서는 적용 가능한 작업 관점에서 데이터 셋의 품질을 평가해야 한다고 기술하고 있다.

따라서, 데이터 수집 단계에서는 원본 데이터의 품질 확보가 필수적임을 알 수 있다. 수집 단계에서 데이터 소스 품질 제약은 데이터 소스의 표준화와 데이터 소스의 보안 및 안정성과 같은 다양한 측면에서 정의될 수 있다. 데이터 품질 보장을 위해서는 빅데이터 분석 프로세스의 각 단계에서 데이터 품질요소를 평가하고 각 지표를 분석하는 것이 필요하다.

3. 빅데이터를 위한 데이터 품질 국제표준화 동향

데이터 품질요소를 정의하는 국제표준인 ISO/IEC 25012는 데이터 품질을 크게 ‘내재된 데이터 품질’, ‘내재적이며 시스템에 의존적인 데이터 품질’, 그리고 ‘시스템에 의존적인 품질’로 구분하고, 이를 다시 15개 품질로 세분화하고 있다[7]. 이 품질 요소를 기초로, 빅데이터와 머신러닝 관련 데이터 품질 요구사항을 정의하고 데이터 품질을 평가하기 위한 표준으로 ISO/IEC 5259가 개발중이다. ISO/IEC 5259 Artificial Intelligence - Data quality for analytics and machine learning (ML)는 표 1과 같이 6개 항목에 대해 개발이 진행되고 있다[8].

ISO/IEC 5259-1은 빅데이터 분석 및 머신러닝 분야에서의 데이터 품질에 대한 개요를 제공하고, 데이터 품질 프레임워크 및 공통 용어에 대한 정의를 포함하고 있으며, 전체 시리즈 표준의 구성과 연관성에 대한 전반적인 내용을 포함하고 있다. ISO/IEC 5259-2은 빅데이터 분석 및 머신러닝 모델에 대한 데이터 품질을 측정하고 보고하는 절차에 대한 지침을 제공한다. ISO/IEC 5259-3은

빅데이터 분석 및 머신러닝에 사용되는 데이터의 품질을 설정, 구현 및 유지하고 지속적으로 개선하기 위한 요구사항과 가이드라인을 제공한다.

표 1. 데이터 품질 표준
Table 1. Data quality standardization

표준명	제목
ISO/IEC 5259-1	Overview, terminology, and examples
ISO/IEC 5259-2	Data quality measures
ISO/IEC 5259-3	Data quality management requirements and guidelines
ISO/IEC 5259-4	Data quality process framework
ISO/IEC 5259-5	Data quality governance framework
ISO/IEC 5259-6	Visualization framework for data quality

ISO/IEC 5259-4는 빅데이터 분석과 머신러닝에서 사용되는 데이터 품질을 보장하기 위해 적용 가능한 일반적인 지침을 제공하기 위한 것으로 데이터 라이프사이클에 따라 수집된 데이터에 대한 학습과 평가에 전반적으로 활용할 수 있도록 한다. ISO/IEC 5259-5는 빅데이터 분석 및 머신러닝을 위해, 데이터 라이프사이클 전반에 걸쳐 적절한 제어를 통해 데이터 품질 측정, 관리 및 관련 프로세스의 구현과 운영을 지시하고 감독할 수 있는 거버넌스 프레임워크를 제공한다. ISO/IEC 5259-6은 2023년 2월 Committee Draft로 제출되어 있으며, 빅데이터 분석 및 머신러닝을 위한 데이터 라이프 사이클 각 단계에서 다양한 이해관계자가 사용할 수 있는 시각화 사례를 제공하기 위한 것이다[9].

4. 빅데이터 품질 평가 방법

빅데이터를 활용하고 분석하기 위해서는 방대하고 복잡한 데이터들로부터 고품질의 데이터를 추출할 수 있어야 한다. 빅데이터 품질 문제는 데이터 가치에 대한 품질 요구 사항이 충족되지 않을 경우 주로 발생하게 되며, 이는 다양한 수준에서 여러 요인으로 인해 발생한다.

데이터 품질 문제로 빅데이터 분석 알고리즘이나 머신러닝 모델의 성능 저하, 부정확하고 신뢰할 수 없는 결과라는 위험이 발생한다. 이러한 위험을 최소화하고 방지하기 위해서는 빅데이터 분석 초기 단계부터 전체 단계에서 데이터 평가가 이루어져야 한다.

본 논문에서는 빅데이터 품질 평가 방법에 대한 선행 연구들을 다음 세 가지 측면에서 검토한다: (1) 데이터 품질 평가 지표 (2) 데이터 품질 평가, (3) 품질 평가 도구 비교.

4.1 데이터 품질 평가 지표

데이터 품질 평가 지표는 정성적 측면과 정량적 측면에서 평가할 수 있다. 많은 문헌에서 다음 그림 2와 같은 계층적 단계를 통해 데이터 품질 평가 지표를 설정하고 있다.

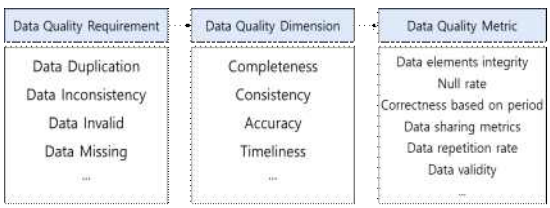

그림 2. 데이터 품질 평가 프레임워크
Fig. 2. Data quality assessment framework

데이터 품질 요구(Data Quality Requirement)는 데이터 품질과 관련된 이슈나 관련 위험으로 정의한다. 예를 들어 데이터 중복, 데이터 불일치, 부정확한 데이터, 데이터 누락 등으로 정의할 수 있다. 데이터 품질 요구사항은 빅데이터 분석을 수행하는 조직이나 분석 목적에 따라 정의한다.

데이터 품질 차원(Data Quality Dimension)은 데이터의 특성을 나타내는 일련의 데이터 품질 속성으로 데이터 측정 기준이 된다. 각 데이터 차원은 성능을 측정할 수 있는 특정 데이터 품질 메트릭으로 구성된다.

데이터 품질 메트릭(Data Quality Metric)은 데이터 차원을 기반으로 데이터 품질을 평가할 수 있는 정성적 또는 정량적 지표로 정의한다.

표 1과 같이 데이터 차원과 데이터 지표는 연구마다 다르게 매핑하고 있으며 각 조직에서는 데이터 품질 요구사항과 목적에 적합하게 선정하여 활용하는 것이 중요하다.

표 1. 데이터의 차원과 지표
Table 1. Data dimension and metrics

참고문헌	데이터차원	데이터지표
[6]	8개	32개
[10]	4개	Data Profiling
[11]	5개	14개
[12]	3개	49개
[13]	6개	9개

데이터 차원은 그림 3과 같이 활용 빈도에 차이가 있으나 일반적으로 많이 활용되고 있는 지표는 Accuracy, Completeness, Consistency, Timeliness이다.

데이터 평가 지표는 통계 기반 지표와 머신러닝 기반 지표가 있으며 대표적인 통계 기반 지표는 Completeness, Timeliness, Imbalance가 있고 머신러닝 기반 지표에는 Validity, Accuracy 등으로 분류하기도 한다[6].

데이터 평가 지표는 빅데이터 라이프 사이클의 각 단계별 매핑을 통해 데이터 품질 평가를 수행하고 가능한 초기 단계에서 고품질의 데이터를 확보해야 빅데이터 분석의 좋은 결과를 기대할 수 있다. 그림 4는 빅데이터 단계와 데이터 차원에 대한 예이다.

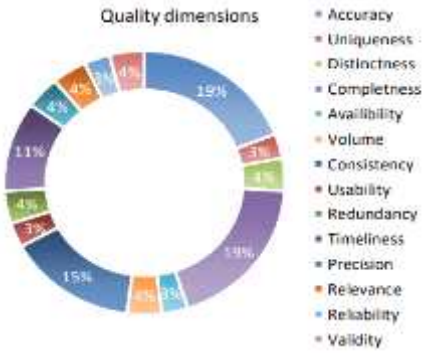

그림 3. 데이터 차원 활용 비율[3]
Fig. 3. Quality dimensions utilization ratio

Stages of big data	Data quality dimensions
Data collection	Availability
	Relevance
Data preprocessing	Usability
	Reliability
Data storage	Usability
	Availability
Data analysis	Reliability
	Usability

그림 4. 빅데이터 단계와 데이터 차원 예
Fig. 4. Examples of big data stages and data dimensions

4.2 데이터 품질 평가

데이터 품질 평가는 평가 지표의 측정 결과와 데이터 사용자의 품질 요구와 비교하여 사용자의 데이터 목표에 이루게 하는 것이다,

빅데이터에서는 데이터 품질에 직접적인 영향을 미치는 4가지 특성 4V을 고려하여 평가되어야 한다. 최근 빅데이터 특성이 10V로까지 확대되면서 품질 목표나 평가가 더더욱 복잡해지고 어려워지고 있다. 빅데이터 속도(Velocity)는 새로운 데이터가 생성되고 처리해야 하는 빠른 방식을 고려할 때 데이터 분석이 실시간으로 수행되므로 데이터 품질 평가도 실시간으로 수행되면서 수시로 재평가되어야 한다는 의미이다.

기존 데이터 품질은 주로 정형 데이터에 초점을 맞추고 있지만 빅데이터는 비정형 데이터의 비중이 매우 높다. 이는 이기종 데이터 소스로부터 다양한 형태의 데이터가 생성되는 빅데이터 다양성(Variety)의 특성으로 데이터 형태나 데이터 출처 유형에 따라 데이터 차원 및 평가 알고리즘을 다르게 적용해야 한다.

또한 데이터의 급격한 변화로 인해 데이터의 적시성이 매우 짧아지기 때문에 이러한 데이터를 기반으로 한 분석 및 처리 결과의 신뢰성이 떨어지거나 불필요한 결과가 나올 수 있다[14].

따라서 빅데이터 품질평가는 그림 5에 나타난 것과 같이 빅데이터 특성을 고려해야 하므로, 기존의 품질 모델이나 방법을 사용하기 어렵고, 빠르게 많은 데이터가 생성되므로 수시로 평가 방법 및 성능을 을 점검해야 한다. 특히 항공, 철도, 국방, 의료 등의 Safety-Critical SW에서 사용하는 빅데이터는 도메인 별로 다르게 적용할 수 있는 품질평가에 대한 지속적인 연구가 요구된다.

Data quality dimensions	Big Data V's			
	Volume	Velocity	Variety	Veracity
Accuracy	X		X	X
Completeness	X	X		X
Consistency		X	X	X
Currency		X		X

그림 5. 데이터 차원과 빅데이터 4V[3]
Fig. 5. Data dimensions and big data 4V

4.3 데이터 품질 평가 도구

빅데이터 분석을 위한 고품질의 데이터 품질을 확보하기 위해서는 데이터 전처리 단계에서부터 품질확보가 되어야 한다. 따라서 데이터 분석가는 좋은 데이터(Good data)를 확보하기 위해 전처리 단계에서 대량의 데이터를 처리하는데 많은 시간과 에너지를 소비한다.

이 문제를 해결하기 위해 MetricDoc도구는 누락된 값, 유효하지 않은 데이터, 잘못된 데이터 생성 및 데이터 중복에 대한 지표를 높이고 탐지 규칙을 사용하여 데이터 품질 문제를 식별하였다[15]. Song et al.는 철자 오류, 중복 레코드, 필드 충돌, 데이터 불일치 등의 데이터 품질 문제를 식별하고 전처리 효율성을 높이기 위한 N-gram 기반 중복 레코드 감지와 같은 데이터 정리 방법을 사용하는 방법을 제안하였다[16]. Guo et al.은 얼굴인식 데이터의 경우 잘못된 ID 라벨이 있는 이미지가 일반적이기 때문에 낮은 품질의 얼굴 이미지를 유지하면서 더 높은 얼굴 모델을 훈련시키기 위한 ID 태그를 청소하는 도구를 제안하였다[17].

이와같이 데이터 품질 평가 도구는 빅데이터 라이프 사이클의 전처리 단계에서 다양한 데이터 형태에 따라 좋은 데이터를 확보하기 위한 많은 연구가 진행하고 있다.

5. 빅데이터 품질 평가 프로세스

빅데이터 품질평가는 그림 6과 같이 데이터 분석 프로세스 전 과정에서 동시에 각 단계별 이루어져야 데이터 품질 보증이 된다. 또한 빅데이터 특성을 고려한 동적인 평가가 수행되어야 한다. 이러한 점을 고려하여 본 논문에서는 빅데이터를 위한 목적 기반 데이터 품질 지표(G-DQM)기반의 데이터 품질 평가 프로세스를 제안한다.

그림 6. 데이터 품질보증
Fig. 6. Data quality assurance

5.1 목적 기반 데이터 품질 지표

기존 문헌과 표준에서 데이터 품질 평가를 위해 평가 지표를 제시하고 있다. 사용자는 많은 문헌에서 제공하는 지표를 빅데이터 분석 목적과 용도에 맞게 선정하여야 한다. 본 논문에서는 Victor Basili가 제안한 GQM(Goal Question Metric) 접근방법을 활용하여 목적 기반 데이터 품질 지표(G-DQM: Goal-driven Data Quality Metric)를 선정하고자 한다. GQM은 목표를 정하고 목표에 맞는 여러 개의 질문을 통해 목표를 명확화하고 질문에 대한 대답에 해당하는 지표를 로 선정하는 방법이다.

5.2 데이터 품질 평가 프로세스

본 논문에서는 목적 기반 데이터 품질 지표(G-DQM)기반의 데이터 품질 평가 프로세스를 제안한다. 제안된 G-DQM 기반 빅데이터 품질평가 프로세스는 그림 7과 같다. 데이터 품질 평가 지표를 선정하기 위한 첫 번째 단계로 데이터 품질 요구사항을 도출하기 위한 데이터 품질 목표를 수립한다. 데이터 품질이 데이터의 유형에 따라, 의료, 금융, 부동산 등 산업 분야별로 다르다. 따라서 빅데이터 특성을 고려한 품질 목표, 빅데이터 분석을 위한 비즈니스 목표, 빅데이터 프로젝트의 목표, 빅데이터의 도메인 특성을 고려한 모든 목표를 수립한다. 다음 단계는 각 목표를 충족할 수는 있는 질문을 도출한다. 하나의 목표를 만족하기 위해 질문은 하나가 될 수도 있고 다수의 질문이 도출될 수 있다. 다음 단계는 각 질문을 만족시키는 품질 지표를 결정한다. 목표-질문-지표를 도출하는 과정은 빅데이터 분석 목적과 용도에 적합한 지표를 도출할 때까지 반복한다.

품질평가 지표가 선정되면 지표 Raw Data와 빅데이터를 수집하고 데이터 전처리를 수행한다. 충분한 전처리를 통해 Good Data를 확보하였으면 이후에 G-DQM을 이용하여 품질평가를 수행한다. 이후에는 품질평가 결과와 베이스라인을 비교한다. 베이스라인은 데이터 품질 요구사항 단계에서 목표 도출 시 결정한다. 데이터 품질평가 결과 베이스라인을 만족하지 못하면 데이터 수집을 다시 수행한다. 베이스라인을 만족한다는 것은 데이터 품질이 확보된 상태이므로 이후에는 빅데이터 라이프 사이클의 단계의 데이터 분석을 수행한다. 빅데이터 분석이후에는 피드백을 받아 필요시 G-DQM 검토를 하고 선정단계를 반복할 수 있다. 빅데이터는 생성 주기도 빠르고 다양한 형태가 데이터 생성되므로 품질 평가는 주기적으로 또는 데이터 수집 주기에 따라 반복한다.

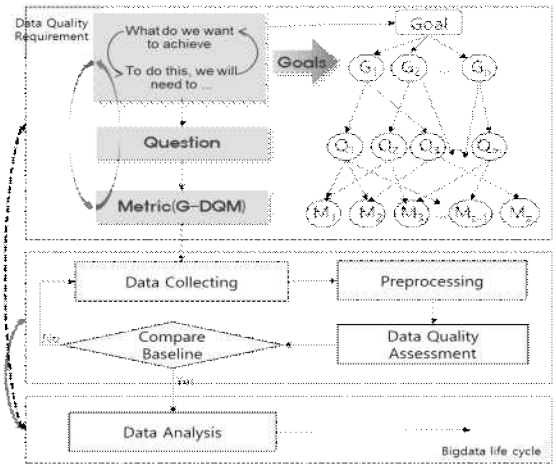

그림 7. G-DQM기반 데이터 품질평가 프로세스
Fig. 7. Data quality assessment process based on G-DQM

6. 결론 및 향후연구과제

빅데이터를 활용하고 분석하기 위해서는 방대하고 복잡한 데이터들로부터 고품질의 데이터를 추출할 수 있어야 한다. 빅데이터를 위한 고품질의 데이터 확보가 어려운 이유는 다음과 같이 세 가지 요인으로 생각해 볼 수 있다.

첫째, 데이터 소스의 다양성으로 인해 데이터 유형이 다양해지고 데이터 구조가 복잡해지면서 데이터 통합이 어려워진다.

둘째, 데이터의 양이 방대하고 합리적인 시간 내에 데이터 품질을 판단하기 어렵다.

셋째, 데이터 변경 속도가 매우 빠르고, 데이터의 타임라인이 매우 짧기 때문에 처리 기술에 대한 더 높은 품질 요구사항이 요구된다.

본 논문에서는 빅데이터나 머신러닝과 같은 데이터 기반 정보기술 분야에서 다루는 데이터에 대한 데이터 라이프사이클을 살펴보고, 데이터 품질 요소에 대한 국제표준화 동향과 함께 라이프 사이클에 따라 고려되어야 하는 데이터 품질 특성에 대해 기술하였다. 또한 빅데이터 관련 데이터 품질 평가에 대한 기존의 주요 연구들을 비교 분석하고, 이를 바탕으로 빅데이터 품질 평가에 필요한 요소들을 정의하였다. 정의된 요소들을 반영하여 목적기반 데이터 품질지표를 이용한 빅데이터 품질 평가 프로세스를 제안하였다.

빅데이터 품질 평가는 기존의 소프트웨어 품질 모델이나 평가 방법을 직접적으로 적용하기 어렵고, 비정형 데이터의 비중이 높은 대량의 데이터가 빠르게 생성된다는 특징을 고려해야 한다. 따라서 제안된 프로세스를 기반으로 향후 빅데이터의 품질을 효과적으로 평가 할 수 있는 평가 지표 개발과 통합된 평가 프레임워크 개발로 확장된 연구가 요구된다.

참 고 문 헌

[1] Suwook Ha, Eui-hyeon Jeong, “Data quality standardization for data analysis and machine learning”, TTA Journal, Vol. 192, pp.22-27, 2020. Available at https://www.tta.or.kr/tta/preportNewsN

[2] C. Northcutt, L. Jiang, I. Chuang, “Confident Learning: Estimating Uncertainty in Dataset Labels”, Journal of Artificial Intelligence Resrarch, Vol. 70, pp.1371-1411, 2021. DOI: https://doi.org/10.1613/jair.1.12125

[3] O. Reda, I. Sassi, A. Zellou, S. Anter, “Towards a Data Quality Assessment in Big Data”, Proceedings of the International Conference on Intelligent Systems: Theories and Applications, pp.1-6, 2020. DOI: https://doi.org/10.1145/3419604.3419803

[4] D. M. Strong, Y. W. Lee, and R. Y. Wang, “Data quality in context”, Communications of the ACM, 40(5), 103-110, 1997. DOI : https://doi.org/10.1145/253769.253804

[5] ITU-T Y.3604, Big data - Requirements and conceptual model of metadata for data catalogue, 2019. https://www.itu.int/rec/T-REC-Y.3603-201912-1/en.

[6] Youdi Gong, Guangzhen Liu, Yunzhi Xue, Rui Li, Lingzhong Meng, “A Survey on dataset quality in machine learning”, Infromation and software technology, 162(107268), 2023. DOI: https://doi.org/10.1016/j.infsof.2023.107268.

[7] ISO/IEC 25012, Software engineering - Software product quality requirements and evaluation (SQuaRE) - Data quality model, 2008.

[8] ISO/IEC DIS 5259. Available at https://www.iso.org/standard/81088.html, 2023.

[9] Wo Chang, “ISO/IEC JTC1/SC 42(AI)/WG2(Data) Data Qulaity for Analytics and Machine Learning (ML)”, May 24, 2022. Available at https://jtc1info.org/wp-content.

[10] I. Taleb, M.A. Serhani, R. Dssouli, “Big data quality: A survey”, Proceeding of IEEE International congress on Big Data, pp.166-173, 2018. DOI: https://doi.org/10.1109/BigDataCongress.2018.00029

[11] Y. Li, H. Song, Y. Xu, “Studies on data quality evaluation index system for internet plus government services in big data era”, Journal of Physices: Conference Series, 1584(1), 2020. DOI: https://doi.org/10.1088/1742-6596/1584/1/012014

[12] H. Chen, D. Hailey, N. Wang and P. Yu, “A Review of Data Quality Assessment Methods for Public Health Information Systems”, International Journal of Environmental Research and Public Health, 11(5), pp.5170-5207, 2014. DOI: https://doi.org/10.3390/ijerph110505170

[13] S. Chug, P. Kaushal, P. Kumaraguru, et al., “Statistical learning to operationalize a domain agnostic data quality scoring”, 2021. DOI: https://doi.org/10.48550/arXiv.2108.08905

[14] L. Cai, and Y. Zhu, “The challenges of data quality and data quality assessment in the big data era”, Data science journal, 14(2), 2015. DOI: https://doi.org/10.5334/dsj-2015-002

[15] B. Christian, G. Theresia, K. Simone, et al., “Visual interactive creation, customization, and analysis of data quality metrics”, Journal of Data Information Quality, 10(1), pp.1-26, 2018. DOI: https://doi.org/10.1145/3190578

[16] F. Ridzuan, W. M. Nazmee Wan Z., “A Review on Data Cleansing methods for Big data”, Proceeia Computer Science, 161, pp.731-738, 2019. DOI: https://doi.org/10.1016/j.procs.2019.11.177

[17] G. Guo, M. Jazaery, “Automated cleaning of identity label noise in a large face dataset with quality control”, IET Biometrics, 9(1), pp.25-30, 2019. DOI: https://doi.org/10.1049/iet-bmt.2019.0081

저 자 소 개

최옥주(Okjoo Choi)

2008.2 숙명여자대학교 컴퓨터과학과 박사
1990.8-1996.3 LG전자 생산기술원 주임원구원
1996.7-2009.8 한국오라클 수석컨설턴트
2009.9-2022.12 한국과학기술원 연구교수
2023.10-현재 강원국립대학교 산학협력 중점교수
<주관심분야> 데이터사이언스, 데이터분석, 소프트웨어품질, 프로젝트 관리, 정보보호

김유경(Yukyong Kim)

2001.8 숙명여자대학교 컴퓨터과학과 박사
2005.9-2006.8 UC Davis, Post-doc.
2006.9-2013.9 한양대학교 컴퓨터공학과 연구교수
2018.3-현재 숙명여자대학교 기초공학부 교수
<주관심분야> 웹서비스 QoS 평가, SOA 기반 IoT 신뢰 평가, 소프트웨어 품질평가