데이터 가공절차 표준화

박천웅 한국데이터산업진흥원 데이터유통지원단 팀장
이창수 국가기술표준원 기술규제대응국 국장

1. 머리말

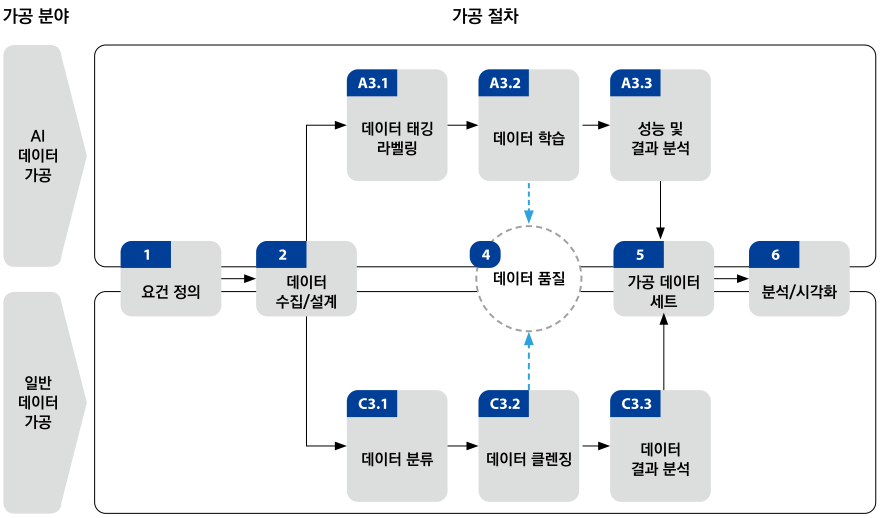
데이터 수명주기는 간단하게는 수집-처리-소멸의 3단계[1]로 이뤄지며, 좀 더 상세하게는 생성-수집-처리-저장-관리-분석-시각화-해석의 8단계[2]로 구분하기도 한다. 몇 단계로 구분하든 데이터 수명주기에서 가장 중요한 단계 중 하나는 처리(processing) 단계이다.

TTA 용어사전[3]에 따르면 데이터 처리는 ① 정보를 받아 특정 결과를 만드는 과정, ②순수 자료를 다시 사용 가능하도록 적당한 형태로 나열하거나 정리하는 것, ③자료나 정보의 기본 요소를 포함하는 주 매체를 다루거나, 그러한 자료를 분류, 연산, 요약, 기록과 같은 동작을 하기 위한 과정의 정확한 규칙에 따라 다루는 것으로 정의한다.

ISO 8000-61[5]에 따르면 데이터 처리의 목적은 적용 가능한 작업지시서에 따라, 해당되는 데이터 명세에 있는 요구사항을 만족시키는 데이터를 제공하는 것이다. ISO 8000-150[6]에서는 데이터 품질관리의 책임(구현, 진단, 개선)과 역할(관리, 운영, 기술)을 구분하고 있는데, 데이터 처리는 데이터 구현 수준에서 기술적 역할에 해당된다.

데이터베이스 관점에서 데이터를 검색, 추가, 삭제, 갱신하는 기능을 데이터 조작(data manipulation)이라고 하며[4], TTA 용어사전에서는 데이터 조작을 데이터의 분류, 갱신, 변경, 첨가, 제거, 입출력 동작, 보고서 작성, 정렬 등과 같은 데이터 처리 작업이라고 정의한다. 데이터 조작이라는 개념은 구조화된 데이터에 적용되는 것이 일반적이었으나 데이터의 유형과 사용 목적이 다양해짐에 따라 데이터 조작이라는 표현과 함께 데이터 가공이라는 표현이 등장하기 시작하였다. 여기에서는 데이터 조작과 데이터 가공이라는 개념을 동일하게 사용하였다.

최근 AI 모델에 적용되는 학습 데이터의 사용이 증가함에 따라 데이터의 신뢰성에 대한 요구가 높아지고 있다. 본 원고는 일반 데이터뿐만 아니라 AI용 데이터에 대한 가공 절차를 제시하고 있다. 본 데이터 가공 절차 표준화는 2020년 AI데이터 표준화 포럼(현재는 지능정보기술



[그림 1] 데이터 가공 분야별 가공 절차 개념도

<표 1> 요건 정의 가공 직무

가공 직무	가공 직무 설명
1.1 업무 분석	수요처가 추진하려고 하는 서비스에 대한 이해와 데이터 가공 필요성 등을 파악하는 과정으로, 핵심 관계자를 대상으로 인터뷰 실시 및 내부 업무 매뉴얼 등을 통해 업무를 분석하는 과정이다.
1.2 기획/설계	업무 분석을 통해 파악된 내용을 기준으로 데이터 가공을 위한 추진 방향, 추진 일정, 사업수행 계획 등 가공서비스 추진 전반의 계획을 수립하는 과정이다.
1.3 데이터 식별	수요처에서 서비스 분석 및 활용 목적에 필요한 데이터 식별을 위해 자체 보유 중인 데이터 및 추가 수집이 필요한 데이터를 식별하는 과정이다.

<표 2> 데이터 수집/설계 가공 직무

가공 직무	가공 직무 설명
2.1 데이터 수집	서비스 실현을 위해 필요데이터를 수집하는 과정이다. 데이터 수집은 데이터 유형(정형, 비정형)에 따라 수집 방식이 다양하며, 목적에 맞게 수집하는 것이 중요하다.
2.2 데이터 저장	수집된 데이터를 활용 목적에 맞게 특정 저장 공간에 저장하는 과정이다. 데이터 저장은 모델설계 과정과 연관성이 높으며, 동시에 수행될 수 있다.
2.3 모니터링	데이터 수집 및 저장이 원활하게 진행되는지 모니터링하는 과정으로, 일정한 주기에 따라 수집되는 데이터인 경우 모니터링이 필수적이다.
2.4 모델 설계	수집된 데이터가 분석 및 활용 목적에 맞게 특정 저장 공간에 구조화된 형태로 저장될 수 있도록 저장 구조를 설계하는 과정이다.

<표 3> 데이터 태깅/라벨링 가공 직무

가공 직무	가공 직무 설명
A3.1.1 데이터 추출	이미지, 동영상, 텍스트 등 비정형데이터를 대상으로 객체를 정의하고 해당 객체에 레이블을 지정하여 필요한 데이터를 추출하는 과정이다.
A3.1.2 데이터 분류	반복학습을 통해 동일 속성의 객체를 인식하고, 해당 객체의 특징에 따라 데이터가 분류될 수 있도록 처리하는 과정이다.
A3.1.3 코딩/개발	데이터 추출 및 분류 등의 과정에서 필요한 데이터를 목적에 맞게 가공하기 위해 코딩 또는 개발과정은 필수이며, 데이터가 디지털화로 자동 변환될 수 있도록 하는 과정이다.
A3.1.4 메타데이터추출/정의	디지털화로 변환된 데이터를 대상으로 정보를 효율적으로 관리하고 활용할 수 있도록 해당 데이터의 의미를 부여하고, 정의하는 과정이다.
A3.1.5 데이터비식별화	누군가의 정체성이 공개되지 않도록 예방하기 위해 사용되는 데이터 가공 과정으로, 데이터 내 개인을 식별할 수 있는 정보를 익명화하거나 가명화하는 과정이다.

포럼 AI데이터 위원회로 명칭 변경)에서 포럼 표준으로 개발되어 일부 수정을 거쳐 TTA 표준(TTAK.KO-10.1250)으로 제정되었다. 이 표준을 발췌하여 소개한다.

2. 데이터 가공 절차 표준화

2.1 데이터 가공 절차

산업현장에서 AI 및 일반데이터의 분석 및 활용 목적에 따라 데이터를 가공하기 위한 일반적인 절차를 기술하고 있다. 또한 가공 절차별 필요 직무 정의를 통해 가공 업무 수행을 위한 기본 방향성을 제시한다. 데이터 가공 분야는 AI 데이터 가공과 일반 데이터 가공으로 구분한다. AI 데이터 가공은 비구조화 형태의 동영상, 이미지 등의 데이터를 분석 가능한 학습 데이터로 가공하는 작업이다. 반면 일반 데이터 가공은 일반적으로 구조화 형태의 데이터베이스, 로그 등의 데이터를 분석 가능한 학습 데이터로 가공하는 작업이다. 데이터 가공 분야에 따른 데이터 가공절차는 [그림 1]과 같다.

2.2 요건 정의

요건 정의 단계는 수요처의 요구 사항을 정의

<표 4> 데이터 학습 가공 직무

가공 직무	가공 직무 설명
A3.2.1 데이터 학습	머신러닝, 딥러닝 등을 통해 데이터 학습을 수행하는 과정이다. 데이터 학습 과정은 매우 복잡하며, 알고리즘에 따라 수행된다.
A3.2.2 데이터 모델링	데이터 학습의 정확도를 높이기 위해 데이터 모델링은 반복적으로 수행되며, 데이터 모델링은 데이터 성능, 튜닝 과정과 연관성이 높다.

<표 5> 성능 및 결과 분석 가공 직무

가공 직무	가공 직무 설명
A3.3.1 데이터 성능	데이터모델에 대한 성능을 측정하는 과정으로, 문제 발생 시 측정결과에 따라 반복적으로 성능을 체크해야 된다.
A3.3.2 데이터 튜닝	데이터 모델의 정확도가 낮거나, 모델로써 활용하기 어려운 경우 모델에 대한 튜닝 과정을 수행한다.
A3.3.3 데이터 검증/평가	데이터 처리과정을 통해 가공된 데이터를 검증하고 평가하는 과정이다. 검증 및 평가 과정에서 가장 중요한 부분은 분석 및 활용 목적에 따라 데이터 세트가 구성되었는지 확인하는 것이다.

<표 6> 데이터 분류 가공 직무

가공 직무	가공 직무 설명
C3.1.1 데이터 추출	RDBMS, 스프레드시트, 로그 등 정형데이터를 대상으로 수집 및 저장된 데이터를 목적에 맞게 추출하는 과정이다.
C3.1.2 데이터 분류	추출된 데이터를 대상으로 데이터 성격 및 유형, 유사성 등을 기준으로 데이터를 분류하는 과정이다.
C3.1.3 코딩/개발	데이터 추출 및 분류 등의 과정에서 필요한 데이터를 목적에 맞게 가공하기 위해 코딩 또는 개발 과정은 필수이며, 데이터가 디지털화로 자동 변환될 수 있도록 하는 과정이다.

<표 7> 데이터 클렌징 가공 직무

가공 직무	가공 직무 설명
C3.2.1 데이터 정제	데이터 처리과정에서 발견된 불필요한 데이터 또는 활용 범위에서 벗어난 데이터를 정제하는 과정이다.
C3.2.2 데이터 비식별화	누군가의 정체성이 공개되지 않도록 예방하기 위해 사용되는 데이터 가공 과정으로, 데이터 내 개인을 식별할 수 있는 정보를 익명화하거나 가명화하는 과정이다.

<표 8> 데이터 결과 분석 가공 직무

가공 직무	가공 직무 설명
C3.3.1 데이터 검증/평가	데이터 처리과정을 통해 가공된 데이터를 검증하고 평가하는 과정이다. 검증 및 평가 과정에서 가장 중요한 부분은 분석 및 활용 목적에 따라 데이터 세트가 구성되었는지 확인하는 것이다.

<표 9> 데이터 품질 가공 직무

가공 직무	가공 직무 설명
4.1 데이터 품질진단/개선	정해진 기준에 따라 데이터의 품질을 진단하고 진단결과, 발견된 오류데이터를 개선하는 과정이다.
4.2 데이터 표준화	데이터 표준관리 요소인 단어, 용어, 도메인을 업무 및 비즈니스 영역에서 활용될 수 있도록 표준사전을 구축하는 과정이다.

<표 10> 가공 데이터 세트 가공 직무

가공 직무	가공 직무 설명
5.1 데이터 세트 정의	가공절차에 따라 구축된 데이터를 정의하는 과정으로, 수요처의 서비스 목적에 맞게 활용할 수 있는 형태로 데이터 세트를 제공하는 것이 중요하다.

<표 11> 분석/시각화 가공 직무

가공 직무	가공 직무 설명
6.1 데이터 분석	목적에 따라 구축된 데이터 세트를 활용하여 데이터 분석을 수행하는 과정이다. 데이터 분석은 다양한 분석 기법이 존재하며, 분석 기획 및 시나리오에 따라 수행된다.
6.2 데이터 시각화	분석된 결과를 시각화하는 과정으로, 데이터 시각화는 데이터 분석 결과를 쉽게 이해할 수 있도록 도표라는 시각적 수단을 통해 정보를 효율적으로 전달하는 과정이다.

<표 12> 데이터 가공 절차별 투입물 및 산출물

가공 절차		투입물	산출물
1. 요건 정의		사업수행계획서 인터뷰 및 요구사항	업무 분석서 요구사항 정의서
2. 데이터 수집/설계		수집 원본 데이터	분석/제공 데이터 모델 설계서 환경 구축서
3.데이터 처리	A3.1 데이터 태깅/라벨링	분석 데이터 세트	메타 데이터 정의서 데이터 비식별 정의서
	A3.2 데이터 학습	분석 데이터 세트 데이터 알고리즘	학습 데이터 알고리즘
	A3.3 성능 및 결과 분석	분석 데이터 세트	성능 및 결과 분석서
	C3.1 데이터 분류	분석 데이터 세트	데이터 분류 결과서
	C3.2 데이터 클렌징	분석 데이터 세트	분석 데이터 클렌징 결과서 (비식별 포함)
	C3.3 데이터 결과 분석	분석 데이터 세트	데이터 검증 결과서
4. 데이터 품질		분석 데이터 세트	데이터 품질 진단/개선 보고서
5. 가공 데이터 세트		분석 데이터 세트	제공 데이터 세트
6. 분석/시각화		분석 데이터 세트	분석 결과서, 시각화 결과물(대시보드, 그래프 등)

하는 단계로 업무 분석, 기획/설계, 데이터 식별의 세부 직무로 <표 1>과 같이 구성된다.

2.3 데이터 수집/설계

데이터 수집/설계 단계는 원본 데이터를 수집하고 가공 데이터 세트를 설계하며, 가공환경을 구축하는 단계이다. 본 절차에서 수행되는 가공 직무는 <표2>와 같다.

2.4 데이터 태깅/라벨링

데이터 태깅/라벨링 단계는 작업 전 데이터 추출 및 데이터의 분류 작업을 통하여 태깅/레이블링 작업을 진행하고 이후 관련 메타데이터 추출/정의 및 개인정보에 대한 비식별화 처리를 하는 단계이다. 본 절차에서 수행되는 가공 직무는 <표 3>과 같다.

2.5 데이터 학습

데이터 학습 단계는 알고리즘의 반복 학습을 통하여 모델을 수립하는 단계이다. 본 절차에서 수행되는 가공 직무는 <표 4>와 같다.

2.6 성능 및 결과 분석

성능 및 결과 분석 단계는 처리 관련 성능 모니터링 및 관련 튜닝을 하는 단계이다. 본 절차에서 수행되는 가공 직무는 <표 5>와 같다.

2.7 데이터 분류

데이터 분류 단계는 관련 데이터 추출 및 데이터 분류, 관련 데이터를 처리하기 위해 개발하는 단계이다. 본 절차에서 수행되는 가공 직무는 <표 6>과 같다.

2.8 데이터 클렌징

데이터 클렌징 단계는 오류 데이터에 대한 정제 및 개인정보에 대한 비식별화 처리를 하는 단계이다. 본 절차에서 수행되는 가공 직무는 <표 7>과 같다.

2.9 데이터 결과 분석

데이터 결과 분석 단계는 처리된 데이터에 대한 검증 및 평가를 하는 단계이다. 본 절차에서 수행되는 가공 직무는 <표 8> 같다.

2.10 데이터 품질

데이터 품질 단계는 각 공정의 공통 공정으로 각 단계별 데이터 품질을 진단하고 정제하는 단계이다. 본 절차에서 수행되는 가공 직무는 <표 9>와 같다.

2.11 가공 데이터 세트

가공 데이터 세트 단계는 수요처에 전달되는 최종 데이터 세트로 처리하는 단계이다. 본 절차에서 수행되는 가공 직무는 <표 10>과 같다.

2.12 분석/시각화

분석/시각화 단계는 다양한 관점으로 볼 수 있는 형태로 분석 및 시각화를 개발하는 단계이다. 본 절차에서 수행되는 가공 직무는 <표 11>과 같다.

3. AI 및 일반 데이터 가공 절차별 투입물 및 산출물

데이터 가공 절차 전 단계에 걸친 투입물 및 산출물은 <표 12>와 같다.

4. 맺음말

데이터 수집 단계에서부터 목적에 맞는 데이터를 구조화하여 수집한다면 이후 과정에서 데이터를 처리하는 것이 비교적 용이하게 진행될 것이다. 그러나 이러저러한 이유로 다양한 형식의 데이터가 다양한 방법으로 수집될 경우, 데이터 처리의 필요성이 절대적이다. 일반 데이터뿐만 아니라 학습용 데이터의 경우에도 원하는 목적에 부합하는 결과를 얻기 위해서는 체계적인 데이터 가공 절차가 필요하다.

본 표준은 산업현장에서 AI 및 일반데이터를 분석하고 활용 목적에 따라 데이터를 가공하기 위한 일반적인 절차를 기술하고 있다. 또한 가공 절차별 필요 직무 정의를 통해 가공 업무 수행을 위한 기본 방향성을 제시한다.

참고문헌

[1] Big data, Preliminary Report 2014, ISO/IEC JTC 1
[2] Tim Stobierski, 8 steps in the data life cycle, Business Insights, 2021.2.2.
[3] terms.tta.or.kr
[4] Michael Widenius, David Axmark, Kaj Arno, MySQL Reference Manual, O'Reilly, 2002
[5] ISO 8000-61:2016 Data quality — Part 61: Data quality management: Process reference model
[6] ISO 8000-150:2022 Data quality — Part 150: Data quality management: Roles and responsibilities