한국컴퓨터정보학회 하계학술대회 논문집 제29권 제2호 (2021. 7)

서로 다른 특성의 파편화된 데이터 결합 방법

문재원O
OO한국전자기술연구원
e-mail: jwmoon@keti.re.krO

The way to combine heterogeneous time series data

Jaewon MoonO
OInformation Media Research Center, Korea Electronics Technology Institute

● 요 약 ●

본 논문에서는 다양한 환경에서 수집된 서로 다른 시계열 데이터를 통합하여 분석 활용하기 위해 추가로 생성해야 할 시계열 데이터의 메타 정보를 정의하고 이를 기반하여 새로운 통합 데이터를 생성하는 방법을 소개한다. 시계열 데이터는 표준화된 기술 방법이 없고 다양한 소스에서 생성되기 때문에 이를 통합하고 활용할 경우 그 기준이 없기 때문에 전문적 지식이 없다면 처리에 어려움을 겪는다. 그러므로 서로 다른 특성의 데이터를 새로운 기준에 의거하여 통합하는 것을 목적으로 필요한 메타 정보를 정의하고 이를 기준으로 데이터를 재가공할 수 있도록 하였다.

키워드: 시계열 데이터 (Timeseries data), 이종 데이터 통합 (Heterogeneous Data integration)

I. 서론

IoT 기술의 발전 및 보급으로 인해 지금도 각종 센서에서 수많은 시계열 데이터가 생성되어 저장되고 있다. 시계열 데이터는 시간 흐름 내 특정 순간인 타임스탬프를 기준으로 필요한 정보가 추가로 저장되는 구조를 따른다[1]. 관련 연구를 위해서 보통 미리 수집할 데이터의 특성을 설계하고 이에 기반하여 데이터를 수집하며 분석에 활용하고 있다. 그러나 실제 센서에서 수집된 데이터들은 수집 주기/값의 범위/포맷 등 성격이 모두 다르므로 적절한 전처리 없이는 쉬운 통합 활용이 어려운 실정이다. 그러므로 본 눈문에서는 여러 곳에서 수집되어 수집 주기 및 특성이 서로 다른 시계열 데이터를 함께 통합적으로 활용하기 위해 고려해야 할 요소를 정의하고 이를 기반한 시계열 데이터 통합 방법을 제안한다.

II. 개별 메타 데이터 기반 통합

본 논문에서는 서로 다른 이질적인 데이터를 새로운 시간 스탬프에 의거하여 통합하는 것을 목적으로 필요한 추가 메타 정보를 정의하였다. 이를 위해서는 데이터의 통합 범위, 개별 데이터 타입 특성, 데이터의 재가공 가능 여부, 데이터 재가공 방법, 데이터 재가공 주기에 대한 추가 정보가 필요하다.

1. 데이터 통합 범위

통합될 데이터의 범위를 설정해야 한다. 일반적으로 별다른 조건이 없다면 통합하려는 데이터의 시작 시점 중 가장 늦은 시간과 데이터의 끝나는 시점 중 가장 빠른 시간을 통합 범위로 설정할 수 있을 것이다. 그러나 통합의 범위는 자유롭게 설정될 수 있다.

2. 개별 데이터 특성

각각의 파편적 데이터의 독립적인 컬럼 데이터에 대해 아래와 같은 정보를 정의해야 한다.

- 데이터 타입: Numeric/Category/String
- 측정된 시간스탬프간 데이터 의존성: Yes/No
- 데이터 수집 주기 (ex> 1min, 1hour, 주기 없음)
- 데이터의 발생 시점 (Continuous, Event)

데이터 타입은 크게 연산이 가능한 Numeric, 한정적인 데이터 값을 갖는 Category, 무한한 개수의 문자열 값을 갖을 수 있는 String으로 구분하였다. 또한 특정 타임스탬프에 샘플링되어 측정된 데이터인지 혹은 사건 발생시점에만 발생한 데이터인지에 따라 Continuous, Event로 구분할 수 있다. 보통 데이터의 주기가 없다면 Event 데이터일 확률이 높다.

3. 데이터 재가공 가능성

데이터가 통합되어 원 데이터와 다른 타임스탬프로의 재배열이 가능한지, 특정 데이터만 재배열 할 것인지에 대한 정의가 필요하다. 이는 데이터의 특성에 의거하여 유추할 수 있지만 최종적으로는 데이터를 다루는 목적과 특성에 의해 결정되어야 할 것이다.

대부분의 주기적이고 연산이 가능한 데이터 타입들은 비교적 데이터 재생성이 쉽다. 그러나 Category 데이터의 경우 산술형 데이터로 변환하여 연산 후 다시 변환 하는 등의 보간 방법을 활용할 수 있다. String 데이터의 경우에는 비정형 특성으로 인해 일반적인 방법으로는 다른 데이터와의 결합 재가공이 어렵다.

4. 데이터 재가공 방법

서로 다른 상황에서 모여져서 각각의 특성을 갖는 데이터는 서로 다른 재가공 방법을 선택해야 하며 통합되기 이전에는 각각의 데이터 주기에 의해 수집되었으나 통합을 하는 과정에서 데이터를 나타내는 시간축의 값이 변하기 때문에 원 데이터보다 주기가 늘어나는 업샘플링의 경우와 원 데이터보다 주기가 줄어드는 다운샘플링의 경우를 구분해야 한다.

보통 Numeric 데이터는 시간 구간 평균 및 중간값 사용 같은 간단한 통계적 기법부터 기계학습 및 회귀식에 의한 보간 등 과거 데이터에 의한 복구 방법까지 다양한 재생성 방법을 정의할 수 있다. 보통 업샘플링의 경우 결측치가 더 발생하게 되며 이를 보완할 수 있는 방법에 대해 데이터별 정의가 필요하다.

5. 데이터 재가공 주기

시계열 데이터의 응용, 분석 학습을 위해서는 통합된 데이터도 일정한 주기를 갖는 것이 좋기 때문에 통합 데이터에 대한 재생성 주기를 설정한다.

6. 메타 정보에 의거한 통합

통합을 위한 메타데이터에 의해 생성된 파라미터의 예제는 아래와 같다.

```
{
    duration":{
        "start_time":"2021:03:01",
        "end_Time":"2021:03:10"
    },
    "frequency": "Timedelta('0 days 01:00:00')",
    "column_characteristics":[
        {
            "column_name":"col_1",
            "column_type":"dtpe('int64')",
            "column_frequency" : "Timedelta('0 days 00:10:00')",
            "pointDependency":"yes",
            "occuranceTime":"Continuous",
            "upsampling_method":"linear interpolation",
            "downsampling_method":"max"
        },
 ................ (생략)
}
```

먼저 데이터를 원 데이터의 타임 스탬프들을 모두 사용하여 통합하고 전체 통합 범위에 의거하여 데이터를 선택한다. 통합 선택 데이터는 타임 인덱스가 늘어나고 주기가 같지 않은 구간은 결측값으로 채워지게 된 형상이기 때문에 개별 데이터들이 재가공 가능한 데이터라면 재가공 주기에 의거하여 한번더 재생성 한다. 재가공시에는 기준 재가공 주기에 의거하여 각 데이터를 업샘플링 할지 다운샘플링 할지를 결정하고 각각의 경우에 맞는 결측값 보간 방법을 사용하여 데이터를 재생성 한다. 재구성 주기가 원 데이터들의 주기보다 작을 경우에는 데이터가 강제로 늘어나기 다수의 결측치가 발생하게 되는데, 이는 다시 추가적인 데이터 보간 방법을 이용하여 보간한다.

III. 결론

본 논문에서는 서로 다른 특성의 데이터를 통합하기 위한 정보 및 기준을 제시하였다. 향후 파편화된 데이터를 통합하는 플랫폼에 기반 기술을 적용할 예정이다.

ACKNOWLEDGEMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00034, 파편화된 데이터의 적극 활용을 위한 시계열 기반 통합 플랫폼 기술 개발)

REFERENCES

[1] Wei, William WS. "Time series analysis." The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2. 2006.