**Weekly report**

날짜 : **2024-04-05**

연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - CNE-join의 성능을 개선할 방법 고민
      * 시간 단축 문제는 가장 나중으로 진행
    - Measuring and Predicting the Quality of a Join for Data Discovery 논문이 참고한 문헌을 읽고 n-join-pair 알고리즘 아이디어 얻기
    - DataMap 개발(Type A, B)
    - 자동 융합 시나리오에 대한 명세를 작성해보기(use case diagram)
      * PPT로 작성
      * 융합을 할 때 Join을 할 지 Union을 할 지 확인하는 방법 – 나는 Join 파트
    - 테이블 임베딩을 위한 Feature selection 방법 탐색
      * 테이블에서 임베딩 해야 하는 중요 컬럼을 자동 / 휴리스틱을 이용해 알아내는 방법 탐색
      * 아래의 논문 읽고 아이디어 얻기
        + <https://arxiv.org/pdf/2310.07736.pdf>
        + <https://arxiv.org/pdf/2207.02773.pdf>
    - BERT 최신 데이터로 추가 학습한 모델의 성능평가

추가 학습한 모델로 두 테이블의 유사도를 판단하는 방법 구현

* **수행결과**
  + **관계형 테이블 임베딩 :**
    - 300개의 테이블 데이터를 입력으로 CNE-Join 코드를 돌리는 중
      * 코드가 거의 다 돌아간 상태, 이후 CNE-Join의 결과를 n-join-pair 코드의 입력으로 넣어 실험을 진행할 예정
    - Observatory: Characterizing Embeddings of Relational Tables 논문 정리
      * 특정 Task(Table interpretation/augmentation, Table Question Answering…)를 수행하려는 연구자에게 적합한 테이블 임베딩 모델, 언어 모델을 추천해주는 프레임워크인 OBSERVATORY를 소개하는 논문

텍스트, 도표, 라인, 폰트이(가) 표시된 사진

자동 생성된 설명

* + - * Observatory는 기존 테이블 임베딩 모델이나 언어모델들의 임베딩 특성을 정량적으로 특성화하기 위해 8 가지 기본 속성과 해당 측정 방법을 정의함
        + Row Order Insignificance
        + Column Order Insignificance
        + Join Relationship
        + Functional Dependencies
        + Sample Fidelity
        + Entity Stability
        + Perturbation Robustness
        + Heterogeneous Context
      * 위의 8 가지 기본 속성과 측정 방법을 통해 얻은 기존의 테이블 임베딩 모델과 언어 모델 9 개에 대한 insight를 알려줌
        + 조인 가능한 테이블들을 찾을 때, Containment 방식은 Embedding 방식을 보완해줄 수 있다
        + BERT를 두 컬럼간 의미적 유사성을 찾는 데 사용하였으면 두 컬럼간 구문론적 유사성을 찾을 때에는 Containment를 사용하는 것이 좋다

텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진

자동 생성된 설명

* + - * + BERT는 컬럼의 일부만 샘플링하여 임베딩한 벡터와 컬럼 전체를 임베딩한 벡터간 코사인 유사도가 매우 높다 -> 샘플링을 해도 컬럼 전체를 임베딩한 것과 큰 차이가 생기지 않는다

텍스트, 도표, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

* + - * 논문에서 주는 언어 모델에 대한 insight는 BERT를 활용한 테이블 융합 연구에서 유용하게 사용할 수 있을 것으로 보임
    - DIWIFT: Discovering Instance-wise Influential Features for Tabular Data 읽는 중
      * 테이블 데이터의 Instance별로 Feature selection 하는 방법에 관한 논문
* **수행계획**
  + **관계형 테이블 임베딩 :**
    - CNE-join 코드를 개선할 방법 고민
      * 시간 단축(코드 리팩토링을 통한)
    - Measuring and Predicting the Quality of a Join for Data Discovery 논문이 참고한 문헌을 읽고 n-join-pair 알고리즘 아이디어 얻기
    - DataMap 개발(Type A, B)
    - 자동 융합 시나리오에 대한 명세를 작성해보기(그림으로) – 논문 서론에 사용할 용도로 간략화 추상화된 그림을 작성
      * 쿼리 바이 테이블도 설명은 하되, 나는 어디까지 구현을 했는지를 표현
      * 서비스 사용자 입장에서 어떻게 융합된 테이블을 얻게 되는지
      * 서비스 개발자 입장에서 무엇을 하는지
      * PPT로 작성 Type A B 구분하지 말고
      * 융합을 할 때 Join을 할 지 Union을 할 지 확인하는 방법 – 나는 Join 파트
    - 테이블 임베딩을 위한 Feature selection 방법 탐색
      * 테이블에서 임베딩 해야 하는 중요 컬럼을 자동 / 휴리스틱을 이용해 알아내는 방법 탐색
      * 아래의 논문 읽고 아이디어 얻기
        + <https://arxiv.org/pdf/2207.02773.pdf>
    - BERT 최신 데이터로 추가 학습한 모델의 성능평가
    - 추가 학습한 모델로 두 테이블의 유사도를 판단하는 방법 구현
* **기타사항**