**Weekly report**

날짜 : **2024-03-08**

연구원 : **최윤석**

* **이전 수행계획**
  + - 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
    - Containment를 개선한 식으로 CNE-Join과 n-join-pair 코드를 교체
    - 외부(최근) 지식을 기반으로 두 테이블의 유사도를 판단하는 방법 고민
      * 언어 모델을 활용하여 진행하는 방식
      * 임베딩 벡터간 거리를 계산하는 방식과 구별하여 진행
      * 최신 데이터를 incremental하게 반영하는 방법
        + Incremental Word2Vec을 이용하는 방법
        + 기존 Word2Vec + 최신 외부지식을 반영하는 방법
* **수행결과**
  + **관계형 테이블 임베딩 :**
    - 외부(최근) 지식을 기반으로 두 테이블의 유사도를 판단하는 방법 고민
      * Incremental Word2Vec을 이용하는 방법
        + 현재 두 테이블간 유사도를 판단할 때 Google News 데이터를 활용한 Pre-trained Word2Vec 모델을 사용하는 중
        + Pre-trained Word2Vec 모델이 제공하는 것은 각 단어의 임베딩 벡터로 모델의 전체 파라미터를 제공해주지 않기 때문에 이를 추가로 학습하는 것은 불가능함
      * 최신 외부지식을 반영하는 방법
        + BERT를 최신 데이터로 추가학습하고 학습된 BERT를 활용해 두 테이블의 유사도를 판단하는 방법
        + 대부분의 테이블은 가장 좌측의 컬럼이 가장 중요한 정보를 담고 있음. 문장으로 비유하면 문장의 주어가 테이블의 가장 좌측 컬럼이라고 볼 수 있음 -> 가장 좌측 컬럼 값이 비슷한 테이블들은 유사하다고 판단할 수 있다

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 번호, 평행이(가) 표시된 사진

자동 생성된 설명

* + - * + 가장 좌측의 컬럼값들을 BERT로 임베딩하고 이들의 평균값을 해당 테이블을 대표하는 임베딩으로 정함
        + 각 테이블의 임베딩 벡터끼리 코사인 유사도 계산을 통해 유사도 값이 높으면 두 테이블이 유사하다고 판단함
      * BERT를 추가 학습하는 방법에 대해 공부 중
* **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
    - Containment를 개선한 식으로 CNE-Join과 n-join-pair 코드를 교체
    - 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 학습하고 이를 기반으로 두 테이블의 유사도를 판단하는 방법 구현
* **기타사항**