**Weekly report**

날짜 : **2024-02-21**

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
    - Table2Vec를 활용해 n개의 테이블을 조인한 테이블의 임베딩 벡터와 원천 테이블의 임베딩 벡터간의 거리의 평균을 조인 결과의 순위를 나누는 기준으로 하여 n-join-pair의 성능평가
    - 저널에 제출할 Full Paper 수정
* **수행결과**
  + **관계형 테이블 임베딩 :**
    - 저널에 제출할 Full paper 수정하여 교수님께 제출
    - Table2Vec를 활용해 n개의 테이블을 조인한 테이블의 임베딩 벡터와 원천 테이블의 임베딩 벡터간의 거리의 평균을 조인 결과의 순위를 나누는 기준으로 하여 n-join-pair의 성능평가

텍스트, 도표, 라인, 평면도이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

* + - * Table2vec에서 쿼리와 테이블간 유사도를 계산하는 방법은 총 4가지가 있음
      * 이중 Late-sum과 Late-max는 큰 의미가 없다고 판단함
      * Early fusion 방식은 테이블 개수가 부족하여 TF-IDF를 사용해도 효과가 좋지 않을 것이라고 판단함
      * Late-avg 방식으로 코드를 작성하여 성능평가 진행
        + 기존 윤종찬 연구원의 TabBERT 방식으로 성능평가를 했을 때 보다 좋은 성능평가 결과를 얻음

Figure 1. Table2Vec를 활용하여 만든 n-join-pair 모델에 대한 성능평가 결과

Figure 2. TabBERT를 활용하여 만든 n-join-pair 모델에 대한 성능평가 결과

* **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
    - Containment 식이 합리적인지 확인하고 개선할 방법 고민
    - Co-occurrence(외부 지식)를 기반으로 두 테이블의 유사도를 판단하는 방법 고민
      * 언어 모델을 활용하여 진행하는 방식
      * 임베딩 벡터간 거리를 계산하는 방식과 구별하여 진행
      * Co-occurrence보다 좋은 지표가 있으면 그것을 활용
* **기타사항**