**Weekly report**

날짜 : **2024-02-28**

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
    - 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
* **수행결과**
  + **관계형 테이블 임베딩 :**
    - Containment 식이 합리적인지 확인하고 개선할 방법 고민
    - CNE-Join에서 사용된 Containment식
      * Containment(, ) =
    - n-Join-pair 성능평가에서 사용된 Containment식
      * Containment (a, b, c) =
    - Containment식 분모에 조인 컬럼에 포함된 유니크한 값의 개수 중 가장 적은 값이 들어가게 되는데 이 부분이 불합리한 부분이 있다고 생각됨

텍스트, 스크린샷, 폰트, 디자인이(가) 표시된 사진

자동 생성된 설명

* + - 개선한 Containment 식
      * Containment(, ) =
      * Containment(, ) =
    - Co-occurrence(외부(최근) 지식)를 기반으로 두 테이블의 유사도를 판단하는 방법 고민
      * Co-occurrence를 기반으로 두 테이블의 유사도를 판단하는 방식이 Word2Vec를 기반으로 두 테이블의 유사도를 판단하는 방식과 너무 유사하다고 판단하여 다른 지표를 찾아보는 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
    - 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
* **기타사항**