**Weekly report**

날짜 : **2023-12-20**

연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - CNE-join의 성능을 개선할 방법 고민
      * 조인하는 테이블의 개수를 3개 이상으로 확장해보기, 최적의 테이블 개수는 무엇인가? 이를 확인할 척도에 대해서도 고민
        + CNE-join의 결과로 나온 2개의 조인 가능한 테이블쌍 중에서 A-B, B-C 같이 테이블이 겹치는 쌍을 활용해 n-join-pair를 생성하는 알고리즘 및 이에 대한 융합도 산정 방안 고민 및 소스코드 작성(이전 CNE-join과의 일관성이 중요함)
      * 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
    - CNE-join 결과로 나온 조인 가능한 테이블쌍 정보를 활용해 최대 4개의 테이블을 조인한 테이블을 벡터로 임베딩
      * 조인한 테이블의 임베딩 벡터와 조인에 사용된 테이블의 임베딩 벡터간의 거리를 활용해 n-join-pair에 관한 insight(티어제 융합결과 랭킹법)를 얻기
    - 박찬훈 연구원과 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
* **수행결과**
  + **관계형 테이블 임베딩 :**
    - 조인한 테이블의 임베딩 벡터와 조인에 사용된 테이블의 임베딩 벡터간의 거리를 활용해 n-join-pair에 관한 insight(티어제 융합결과 랭킹법)를 얻기
      * 내가 가지고 있는 50여개의 데이터셋 중에서 3개의 테이블을 조인한 테이블의 임베딩 벡터와 조인에 사용된 테이블의 임베딩 벡터간의 거리의 평균을 구하고, 구한 값들을 0~1 사이로 정규화를 진행함
      * 정규화를 마친 후의 거리 값이 0~0.25 사이면 1티어 조인, 0.25 ~ 0.5 사이면 2티어 조인, 0.5 ~ 0.75 사이면 3티어 조인, 0.75 ~ 1 사이면 4티어 조인으로 판단하기로 함
      * 아래의 그림은 내가 가지고 있는 데이터셋에 대해 티어를 나누어본 것
        + 해당 티어가 유의미한 결과를 가지고 있는지는 추후에 정성적으로 평가하여 확인해볼 예정임

텍스트, 흑백, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 흑백, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷, 흑백이(가) 표시된 사진

자동 생성된 설명

* + **빅데이터 자료구조 조교 :**
    - 빅데이터 자료구조 기말고사 채점 및 엑셀에 기말 성적 기입 완료
  + **대학원 수업**
    - 고급정보검색론 기말고사 대비
    - 인간컴퓨터상호작용 최종 프로젝트 제작 및 발표 준비
* **수행계획**
  + **관계형 테이블 임베딩 :**
    - CNE-join의 성능을 개선할 방법 고민
      * 조인하는 테이블의 개수를 3개 이상으로 확장해보기, 최적의 테이블 개수는 무엇인가? 이를 확인할 척도에 대해서도 고민
        + CNE-join의 결과로 나온 2개의 조인 가능한 테이블쌍 중에서 A-B, B-C 같이 테이블이 겹치는 쌍을 활용해 n-join-pair를 생성하는 알고리즘 및 이에 대한 융합도 산정 방안 고민 및 소스코드 작성(이전 CNE-join과의 일관성이 중요함)
      * 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
    - CNE-join 결과로 나온 조인 가능한 테이블쌍 정보를 활용해 최대 4개의 테이블을 조인한 테이블을 벡터로 임베딩
      * 조인한 테이블의 임베딩 벡터와 조인에 사용된 테이블의 임베딩 벡터간의 거리를 활용해 n-join-pair에 관한 insight(티어제 융합결과 랭킹법)를 얻기
        + 3개의 테이블을 조인한 테이블의 임베딩 벡터와 조인에 사용된 테이블의 임베딩 벡터간의 거리의 평균을 티어를 나누는 기준으로 사용하고 있는데, 이것이 유효한지 성능평가를 진행
        + 추가로 3개의 테이블을 조인한 테이블의 임베딩 벡터와 조인에 사용된 테이블의 임베딩 벡터간의 거리 중 가장 먼 거리를 티어를 나누는 기준으로 사용하여 평균과의 성능 비교를 진행
    - 박찬훈 연구원과 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
* **기타사항**