**Weekly report**

날짜 : **2024-03-22**

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
    - 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 학습한 모델의 성능평가
    - 추가 학습한 모델로 두 테이블의 유사도를 판단하는 방법 구현
* **수행결과**
  + **관계형 테이블 임베딩 :**
    - BERT를 최신 데이터로 추가 학습
      * Inshorts라는 사이트에 개제된 1줄 정도의 짧은 2023-2024년도 영어 뉴스 기사 28만개로 BERT를 추가 학습 진행
      * 추가 학습된 BERT를 활용해 두 단어간 코사인 유사도 계산을 진행
        + chatgpt와 openai의 코사인 유사도 0.47 -> 0.48
        + covid와 virus의 코사인 유사도 0.59 -> 0.46
        + ukraine와 russia의 코사인 유사도 0.78 -> 0.70
      * 추가 학습된 BERT가 기존 pre-trained 모델보다 좋은 성능인지 판단이 안됨
      * 모델의 성능평가를 하는 방법이 필요함
    - Containment를 개선한 식으로 CNE-Join과 n-join-pair 코드를 교체
      * CNE-Join : Containment(a, b) = -> Containment(a, b) = . 두 컬럼의 Containment 값이 0.5 이상이면 조인 가능한 컬럼이라고 판단
      * n-join-pair : Containment (a, b, c) = -> Containment (a, b, c) = . 3개 테이블의 조인 컬럼에 대해 Containment 값을 계산하고 이 값을 활용해 성능평가를 진행
      * 교체한 코드를 활용해 3개의 테이블을 조인하고 이를 n-join-pair를 이용해 조인이 잘 된 순으로 랭킹화. 이후 containment를 활용해 n-join-pair에 대한 성능평가를 진행
      * 50여개의 테이블에 대해 CNE-Join을 사용해 조인 가능한 두 테이블쌍을 찾고 테이블쌍을 조합해 3개의 테이블을 조인한 테이블을 만들었더니 해당 개수가 8개밖에 나오지 않음.
        + 조인 테이블의 개수를 늘려야 유의미한 결과가 나올 것 같음
        + 학부연구생분들이 주신 캐글에서 테이블 데이터를 크롤링 하는 코드를 활용해 테이블 데이터 개수를 1000개 정도까지 늘릴 예정
        + 테이블 데이터가 늘어나면 현재 램 용량인 16GB로는 너무 오래 코드가 돌아갈 것 같아 추가 램 구매가 필요해 보임
* **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
    - 실험에 사용할 테이블 데이터 추가로 수집
* **기타사항**