**Weekly report**

날짜 : **2024-03-15**

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 학습하고 이를 기반으로 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
    - 교수님이 보내주신 논문 읽는 중
      * Observatory: Characterizing Embeddings of Relational Tables
* **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
* **기타사항**