**Weekly report**

날짜 : **2024-08-02**

연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - DataMap 개발(Type A, B)
    - 자동 융합 시나리오에 대한 명세를 작성해보기(그림으로) – 논문 서론에 사용할 용도로 간략화 추상화된 그림을 작성
      * 쿼리 바이 테이블도 설명은 하되, 나는 어디까지 구현을 했는지를 표현
      * 서비스 사용자 입장에서 어떻게 융합된 테이블을 얻게 되는지
      * 서비스 개발자 입장에서 무엇을 하는지
      * PPT로 작성 Type A B 구분하지 말고
      * 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
      * 나만의 Feature selection 아이디어가 있으면 연구의 Sub결과로 활용 가능하니 아이디어의 변형이 있으면 고민해보기 - CNE-Join에서 Feature selection으로 시간 줄이기 가능 / N-Join에서 Feature selection은 성능 향상으로 가능
    - TF-IDF 방식의 평가 척도 개선
      * 예시를 통해 논리를 보강할 필요가 있음, 예시를 우선적으로 제시하고 TF-IDF의 일반적인 방식을 소개하는 것으로
      * 아이디어가 너무 심플해 보임. 살짝은 더 로직을 추가
        + 재료 테이블 TF-IDF 1위부터 10위 중 1위에 가까운 단어들이 조인 테이블 TF-IDF 30위 안에 들어 있을 시 더 높은 가중치를 두는 방식 -> 해당 방식으로 실험하기
      * 튜닝 아이디어 : 재료 테이블과 융합 테이블에 가중치를 부여하는 방법?
        + 재료 테이블의 전체 레코드 수 와 융합 테이블의 전체 레코드 수가 조인 품질 척도의 인자로 들어가는 것도 생각해보기
    - 영문 저널논문 작성을 위한 실험 진행
    - 영문 저널논문 작성
    - 연구과제 보고서 작성
* **수행결과**
  + **관계형 테이블 임베딩** 
    - 영문 저널논문 준비
      * 학위논문을 영어로 번역하여 작성 완료
        + 최소 참고문헌 개수가 25개여서 참고문헌 추가 필요
        + 추가실험 내용을 반영하여 추가 작성 예정
      * 추가실험 아이디어
        + CNE-Join 실험시에 컬럼의 일부를 샘플링하고 샘플링한 일부의 컬럼을 활용해 조인 가능한 테이블쌍을 찾는 실험

기존 CNE-Join의 Named Entity Recognition(NER) 단계에서 모든 컬럼값에 대해 NER을 하지 않고 컬럼값 중 20%만 샘플링하여 NER을 진행함

실험 결과로 얻은 F1 score는 아래의 표와 같음

|  |  |  |  |
| --- | --- | --- | --- |
| CNE-Join Hyperparameter | | CNE-Join | CNE-Join sampling |
| Column name Embedding Threshold 0.3 | Containment Threshold 0.6 | 0.656 | 0.629 |

실행시간의 경우 기존 CNE-Join이 1000개의 테이블 데이터에 대해 100시간이 걸린 것에 반해 샘플링 기법을 사용한 CNE-Join은 25시간으로 4배 줄어들었음

기존 CNE-Join이 찾은 조인 가능한 테이블 쌍 개수 14916  
 실제로 조인 가능한 테이블 쌍 개수 7580

샘플링CNE-Join이 찾은 조인 가능한 테이블 쌍 개수 14351  
 실제로 조인 가능한 테이블 쌍 개수 7058

CNE-Join Hyperparameter를 변경해가며 추가 실험 예정

* **수행계획**
  + **관계형 테이블 임베딩 :**
    - DataMap 개발(Type A, B)
    - 자동 융합 시나리오에 대한 명세를 작성해보기(그림으로) – 논문 서론에 사용할 용도로 간략화 추상화된 그림을 작성
      * 쿼리 바이 테이블도 설명은 하되, 나는 어디까지 구현을 했는지를 표현
      * 서비스 사용자 입장에서 어떻게 융합된 테이블을 얻게 되는지
      * 서비스 개발자 입장에서 무엇을 하는지
      * PPT로 작성 Type A B 구분하지 말고
      * 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
      * 나만의 Feature selection 아이디어가 있으면 연구의 Sub결과로 활용 가능하니 아이디어의 변형이 있으면 고민해보기 - CNE-Join에서 Feature selection으로 시간 줄이기 가능 / N-Join에서 Feature selection은 성능 향상으로 가능
    - TF-IDF 방식의 평가 척도 개선
      * 예시를 통해 논리를 보강할 필요가 있음, 예시를 우선적으로 제시하고 TF-IDF의 일반적인 방식을 소개하는 것으로
      * 아이디어가 너무 심플해 보임. 살짝은 더 로직을 추가
        + 재료 테이블 TF-IDF 1위부터 10위 중 1위에 가까운 단어들이 조인 테이블 TF-IDF 30위 안에 들어 있을 시 더 높은 가중치를 두는 방식 -> 해당 방식으로 실험하기
      * 튜닝 아이디어 : 재료 테이블과 융합 테이블에 가중치를 부여하는 방법?
        + 재료 테이블의 전체 레코드 수 와 융합 테이블의 전체 레코드 수가 조인 품질 척도의 인자로 들어가는 것도 생각해보기
    - 영문 저널논문 작성을 위한 실험 진행
    - 연구과제 보고서 작성
* **기타사항**
  + **한국전자거래학회 저널 논문 투고 (8월 개제)**