**Weekly report**

날짜 : **2023-03-23**

연구원 : **최윤석**

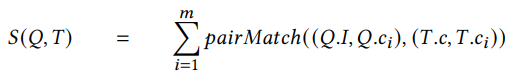
* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - Value overlap + Semantic overlap + Embedding(EmbDI) 을 사용하여 테이블 융합이 가능한지 확인하는 나만의 아키텍처 연구
    - 테이블 간 조인 / 유니온 하기위한 컬럼을 찾는 문제에 관해 생각 (Elmo / Bert / Word2Vec) 등도 생각 / 내가 손으로 하는 것을 자동화하는 방법으로 생각
    - SemProp(Attribute overlap + Value overlap + Embedding) 문제점을 찾고 문제점을 타개할 해결방안 찾기, 큰 부분은 모방해도 되지만 독창성 있는 아이디어로 나의 연구를 할 것
    - Contextual word embeddings for tabular data search and integration 저자가 올린 소스코드 설치 및 실행
    - SANTOS: Relationship-based Semantic Table Union Search 읽고 정리
    - Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach 읽고 정리
    - DeepJoin: Joinable Table Discovery with Pre-trained Language Models 읽고 정리

* **수행결과**
* **관계형 테이블 임베딩** : (임베딩 / 융합 가능성 파트 분류해서 작성)
  + SANTOS: Relationship-based Semantic Table Union Search 읽고 핵심 아이디어와 연구에 적용할 점 정리
    - 논문의 아이디어
      * Union이 가능한 테이블을 찾기 위해 기존의 연구는 두 테이블의 각 컬럼들이 의미적으로 유사한지만을 보았지만, 논문에서는 그것과 더불어 테이블에서 다른 컬럼과의 관계도 유사한지를 봐야된다고 함

도표이(가) 표시된 사진

자동 생성된 설명

* + - SANTOS 에서 Union이 가능한 테이블을 검색하기 위해서는 Query 테이블과 Intent column을 제시해야 함.
      * Intent column은 검색자가 가장 관심있는 컬럼이며, 이 컬럼을 트리의 루트로 하여 다른 컬럼과의 관계를 가지는 트리를 만듦. 이 트리를 Semantic Graph라고 부름
      * 이후 Intent column과 의미적으로 유사한 것을 루트로 하고 다른 컬럼들과도 관계도 비슷한 Sematic Graph를 가진 테이블들을 data lake에서 찾아 제시함.
    - Semantic Graph는 외부 Knowledge Base를 활용해 만듦.
    - Semantic Graph끼리의 유사도 계산식, 즉 Union이 얼마나 잘 될지 나타내는 식은 다음과 같음





* + - 논문내용 중 내 연구에 적용할 부분
      * Union이 가능한 테이블을 찾을 때 각 컬럼들이 의미적으로 유사한지와 더불어 테이블에서 다른 컬럼과의 관계도 유사한지를 고려하는 모델을 구상
      * 논문에서는 각 컬럼간의 의미적으로 유사한지와 테이블에서 다른 컬럼과의 관계가 유사한지를 구할 때, Knowledge Base를 이용하는데, 이를 Embedding(EmbDI)을 사용하여 나의 연구에 적용하는 아이디어 구상
* **수행계획**
  + **관계형 테이블 임베딩 :**
    - Value overlap + Semantic overlap + Embedding(EmbDI) 을 사용하여 테이블 융합이 가능한지 확인하는 나만의 아키텍처 연구
    - 테이블 간 조인 / 유니온 하기위한 컬럼을 찾는 문제에 관해 생각 (Elmo / Bert / Word2Vec) 등도 생각 / 내가 손으로 하는 것을 자동화하는 방법으로 생각
    - SemProp(Attribute overlap + Value overlap + Embedding) 문제점을 찾고 문제점을 타개할 해결방안 찾기, 큰 부분은 모방해도 되지만 독창성 있는 아이디어로 나의 연구를 할 것
    - Contextual word embeddings for tabular data search and integration 저자가 올린 소스코드 설치 및 실행
    - Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach 읽고 정리
    - DeepJoin: Joinable Table Discovery with Pre-trained Language Models 읽고 정리
    - 테이블 융합 관련 논문 추가 검색
    - 연구 논문에 사용할 만한 재료(논문 제목과 아이디어)
* **기타사항**
  + **리포트 작성 시, 논문의 내용정리 같이 남이 한 부분에 대한 설명은 자세하게 하지 말고 그 중에서 내가 얻은 아이디어나 내 연구에 관한 부분을 자세하게 설명할 것**
  + **데이터 융합 프로세스**



* **융합 가능도 척도**
  + 임베딩 거리 기준 값을 정하고 기준값보다 임베딩 거리가 가까우면 융합 가능한 데이터셋이라 판단함
  + 융합 가능한 데이터셋을 사람이 하나씩 확인하여 유의미한 융합이 되는 데이터셋을 찾음
  + 을 융합 가능도 척도라 정의함



혼합형 데이터셋 예시

등급의 semantic

융합 가능도 딥러닝 아키텍처 : 임베딩 벡터 + 메타데이터(융합 가치를 정량화할 수 있는 무언가)   
조인할 수 있는 컬럼이 필요함(A : 주민번호, B : 주민등록 번호로 조인이 가능한 테이블이면?) 융합을 하는 건 가치가 있기 때문, 가치가 있다? -> 임베딩이 가깝게 되어있다.

조인할 수 있다, 없다 -> 사람이 판단

이걸 조인해서 가치있는 정보가 나오나? -> 모델이 판단

Id를 가지는 컬럼이 의미는 같지만 겉모양이 다를때

그래프 임베딩

정형테이블을 그래프로 바꾸는 논문읽기

그래프 정보를 임베딩한 정보와 기존 윤종찬 연구원과의 차이점 발견

나만의 새로운 패러다임 찾기

데이터셋은 영문 데이터만 사용

융합 가능성 평가는 임베딩 성능을 확보한 후에

테이블 융합의 연산은 Union / Join이 있음 -> Union / Join 을 구분해줄 수 있는 모델이 있으면 좋겠음.

Ex. Output 값이 1이면 Union 0이면 Join

조인키만 있어도 조인이 되는게 아닐 수 있음. 나중에 코드를 짤 때 조인(유니온)이 가능한 예제를 sql로 조인해보고 조인이 가능한 예제를 모델에 넣어서 조인이 가능하다고 판독이 되는 것을 확인해보기

6월 학술대회 준비 (완성된 논문이 아니더라도 결과가 의미가 있는 것이라면 논문 작성 가능)

유니온 / 조인 중 하나만 따로라도 혹은 기존의 실험 내용 중 소논문으로 만들거리를 찾아보기 (정보과학회 투고 예정)