**Weekly report**

날짜 : **2023-02-03**

연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - 수치형 데이터를 이산화한 것을 범주형 데이터로 바꾸는 임베딩 방식을 보완하는 방법을 찾고 실험하여, 기존의 윤종찬 연구원의 연구와 **융합 가능도 척도**를 사용하여 성능 비교
    - 각 컬럼마다의 중요도를 고려해서 임베딩하는 방식과 수치형 컬럼에서 숫자의 크기가 크고 작음에 부여된 의미를 고려하는 임베딩 방식 연구
    - 나만의 임베딩 아키텍처 / 융합 가능성 아키텍처(임베딩 아키텍처에 dependant한) 그려보기, 기존의 것과 나만의 것을 분리하여 정리 (PPT로)
    - 모델의 성능이 안 좋은 이유 찾고 해결. (데이터의 문제 / 임베딩 방식의 문제) – 데이터의 문제 : 융합 가능한 라벨들을 찾아서 추가
    - Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks을 인용한 논문 중 연구와 관련된 것 읽어보기, Valentine: Evaluating Matching Techniques for Dataset Discovery 읽고 공부
    - EmbDi 관련 코드 설치 및 실행

* **수행결과**
* **관계형 테이블 임베딩** : (임베딩 / 융합 가능성 파트 분류해서 작성)
  + Valentine: Evaluating Matching Techniques for Dataset Discovery 논문 읽고 공부
    - 논문의 내용은 기존의 7가지 테이블 매칭 기법을 다양한 테이블 데이터셋에 적용하여 성능을 비교해보는 것임. 위의 6개는 state-of-art 매칭 기술이고, Jaccard-Levenshtein은 베이스라인 기법으로 모든 컬럼쌍마다 자카드 유사도를 계산해 출력으로 유사도 순으로 랭킹을 출력해줌

테이블이(가) 표시된 사진

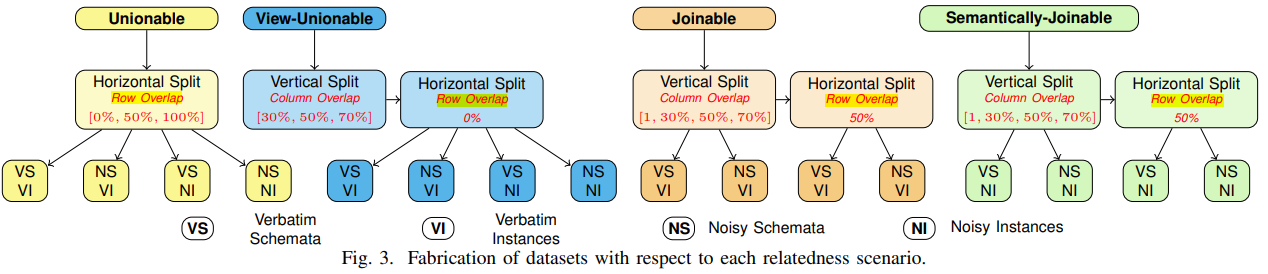
자동 생성된 설명

* + - * Attribute Overlap : 컬럼명의 철자가 겹치는지를 기준으로 매칭하는 기법
      * Value Overlap : 컬럼에 포함된 값들의 철자가 겹치는지를 기준으로 매칭하는 기법
      * Semantic Overlap : knowledge base에 의거해 컬럼이나 컬럼에 속한 값이 의미적으로 겹치는지를 기준으로 매칭하는 기법
      * Data Type : 범주형 / 수치형 같은 데이터 타입을 기준으로 매칭하는 기법
      * Distribution : 컬럼에 포함된 값의 분포가 비슷한지를 기준으로 매칭하는 기법
      * Embeddings : 임베딩의 유사도를 기준으로 매칭하는 기법
    - 4가지 케이스의 두 데이터셋이 관계 있는 경우를 정의하고 각 케이스에 대해서 7가지 테이블 매칭 기법을 각각 적용하여 실험 진행

테이블이(가) 표시된 사진

자동 생성된 설명

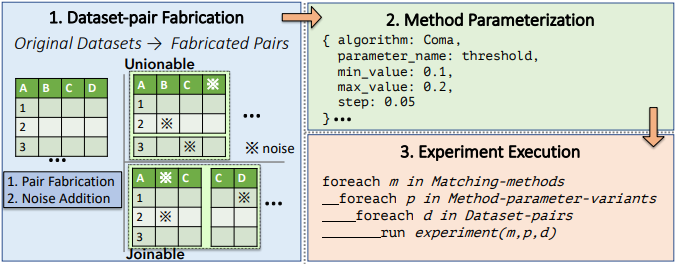
* + - 4가지 케이스에 사용될 데이터셋도 생성함. 이 때 생성하는 방법은 2가지를 사용함. 하나는 1개의 테이블을 쪼개서 2개의 테이블로 만들어, 만들어진 2개의 테이블을 입력으로 학습하는 것이고(540개), 다른 하나는 사람이 선별해서 융합이 가능해 보이는 두 테이블을 입력으로 사용하는 것임(13쌍).



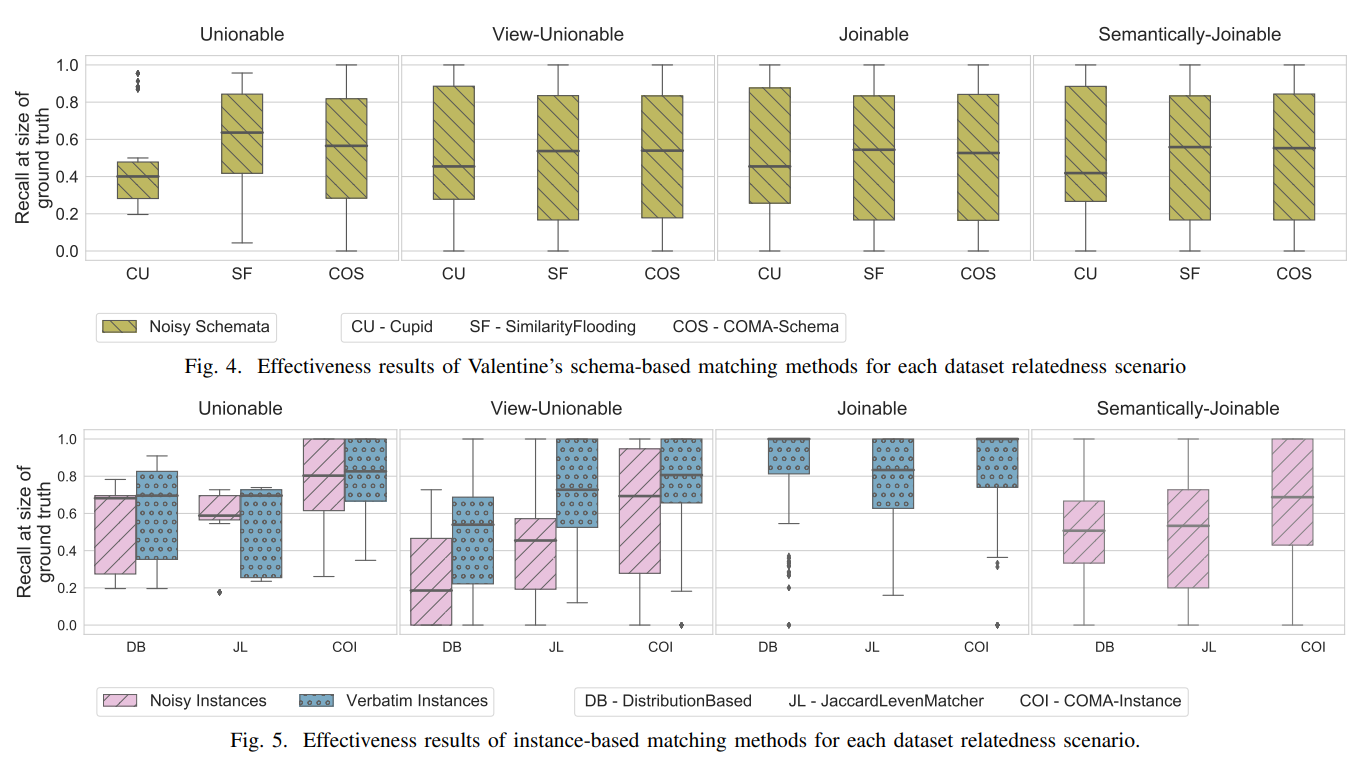
* + - 첫번째 방법인 한 데이터셋을 둘로 나누는 것을 각 케이스에 대해 방법을 제시한 그림
    - Noisy Instances는 데이터에 노이즈를 더하는 것. 범주형 데이터면 랜덤하게 철자에 오타를 내고 수치형 데이터면 각 컬럼내에서 수치형 데이터에 분포에 따라 값을 바꿔주는 것
    - Noisy Schemata는 컬럼명에 노이즈를 더하는 것. 기존 테이블 이름 + 기존 컬럼명으로 컬럼명을 변경하거나 컬럼명을 축약하거나, 컬럼명에서 모음을 없애는 방식을 사용함
    - 성능 지표로 Recall@ground truth를 사용

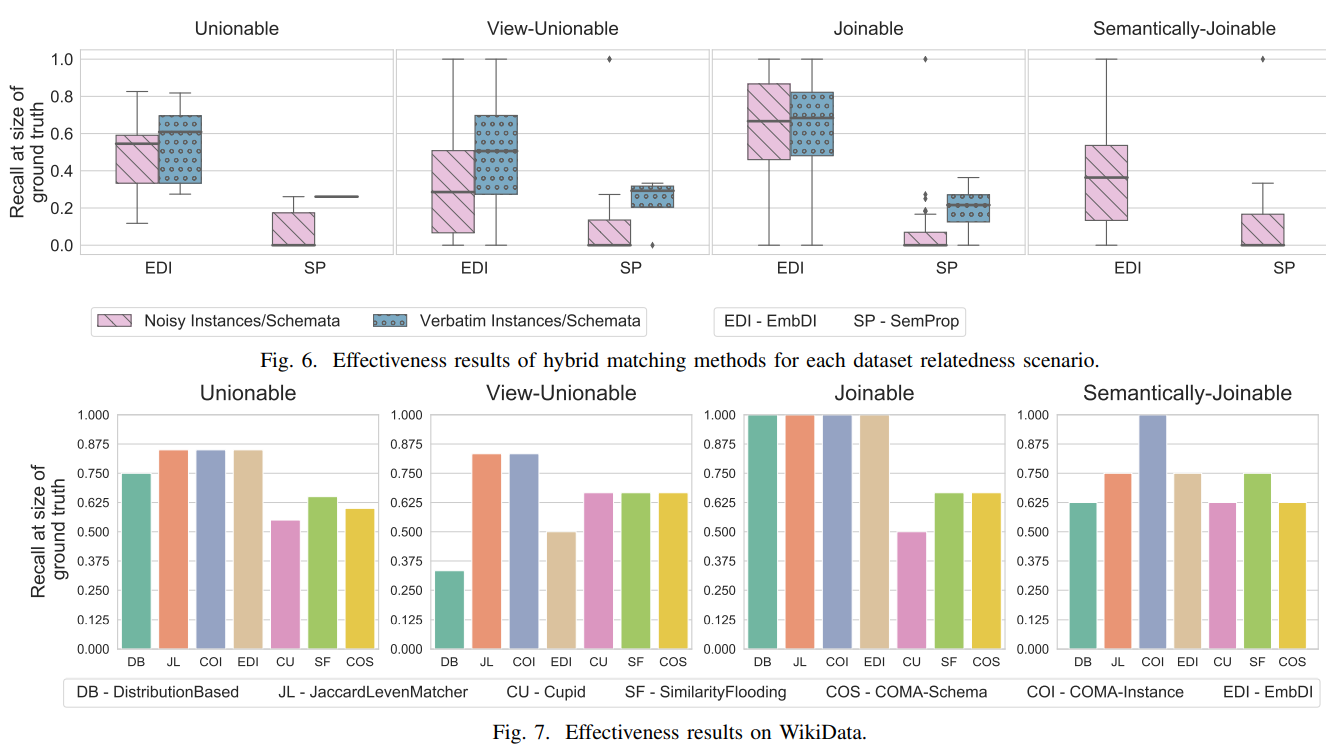
텍스트이(가) 표시된 사진

자동 생성된 설명



* + - 실험 과정의 3단계
    - 실험결과 – 스키마 기반 / 인스턴스 기반 / 하이브리드 의 세가지 카테고리로 정리함
      * 스키마 기반 : 컬럼명 / 데이터 타입 / 맥락 정보
      * 인스턴스 기반 : 각 컬럼에서 데이터 값의 분포
      * 하이브리드 : 스키마 기반과 인스턴스 기반 방식 둘 다 사용





* + - 결론 – 모든 케이스에서 다른 모델보다 우수한 모델은 없었음. 임베딩 기반 모델이 스키마 기반 모델이나 인스턴스 기반 모델보다 안 좋은 성능을 보임. 임베딩 기반 모델의 성능 향상을 위해서 스키마 기반 모델이나 인스턴스 기반 모델과의 통합 등을 생각해봐야 함.
  + (임베딩) : EmbDI 코드 설치 및 예시 데이터셋을 활용하여 EmbDI 실행. EmbDI 코드 및 학습에 필요한 파일구조 분석 진행

텍스트이(가) 표시된 사진

자동 생성된 설명

* **GAN 논문 세미나 :**
  + **Pix2Pix | Image-to-Image Translation with Conditional Adversarial Networks** 논문 읽는 중
* **수행계획**
  + **관계형 테이블 임베딩 :**
    - 수치형 데이터를 이산화한 것을 범주형 데이터로 바꾸는 임베딩 방식을 보완하는 방법을 찾고 실험하여, 기존의 윤종찬 연구원의 연구와 **융합 가능도 척도**를 사용하여 성능 비교
    - 각 컬럼마다의 중요도를 고려해서 임베딩하는 방식과 수치형 컬럼에서 숫자의 크기가 크고 작음에 부여된 의미를 고려하는 임베딩 방식 연구
    - 나만의 임베딩 아키텍처 / 융합 가능성 아키텍처(임베딩 아키텍처에 dependant한) 그려보기, 기존의 것과 나만의 것을 분리하여 정리 (PPT로)
    - 모델의 성능이 안 좋은 이유 찾고 해결. (데이터의 문제 / 임베딩 방식의 문제) – 데이터의 문제 : 융합 가능한 라벨들을 찾아서 추가
    - Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks을 인용한 논문 중 연구와 관련된 것 읽어보기 (Embedding + Semantic Overlap(외부 knowledge base)을 섞는 방식 관련해서)
    - 테이블 간 조인 / 유니온 하기위한 컬럼을 찾는 문제에 관해 생각 (Elmo / Bert / Word2Vec) 등도 생각 / 내가 손으로 하는 것을 자동화하는 방법으로 생각해보기
    - EmbDI 코드에 커스텀 데이터를 입력하여 임베딩 및 성능 확인
  + **GAN 논문 세미나 :**
    - **Pix2Pix | Image-to-Image Translation with Conditional Adversarial Networks** 논문 읽고 세미나 준비 (PPT 30분 이내 발표) , 각 논문의 핵심사항을 알아내는게 목적. 연구 동기, 배경, 관련연구 등은 생략하고 이해한 범위 내에서 핵심사항 중심으로 발표 진행. 성능 평가는 한 페이지(메인 그래프만)로 충분함. 그림과 키워드 위주로 발표. 예제 위주 이해 위주 친절한 자신만의 이해한 언어로 진행
* **기타사항**
  + **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