**Weekly report**

날짜 : **2023-10-25**

연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - CNE-join의 성능을 개선할 방법 고민
      * 조인하는 테이블의 개수를 3개 이상으로 확장해보기, 최적의 테이블 개수는 무엇인가? 이를 확인할 척도에 대해서도 고민
        + CNE-join의 결과로 나온 2개의 조인 가능한 테이블쌍 중에서 A-B, B-C 같이 테이블이 겹치는 쌍을 활용해 n-join-pair를 생성하는 알고리즘 및 이에 대한 융합도 산정 방안 고민 및 소스코드 작성(이전 CNE-join과의 일관성이 중요함)
      * 시간 단축 문제는 가장 나중으로 진행
      * Measuring and Predicting the Quality of a Join for Data Discovery 논문 읽고 정리 및 연구 아이디어에 사용할 부분 찾기
    - 박찬훈 연구원과 DataMap 개발(Type A, B)

* **수행결과**
  + **관계형 테이블 임베딩 :**
* Measuring and Predicting the Quality of a Join for Data Discovery 논문 읽고 핵심 내용 정리
  + CNE-join과 마찬가지로 조인 컬럼을 찾아 조인 가능한 테이블을 찾는 방법에 관한 논문.
  + 컬럼 A, B가 있을 때, 아래 식들의 값이 둘 다 1에 가까우면 A,B를 조인 가능한 컬럼이라고 판단함.
  + **폰트, 노랑, 번호, 라인이(가) 표시된 사진

    자동 생성된 설명,텍스트, 폰트, 노랑, 포스트잇 노트이(가) 표시된 사진

    자동 생성된 설명**
  + C는 Containment, K는 Cardinality Proportion의 약자
  + 논문에서 말하길 Containment 식만 사용해서는, 두 컬럼의 레코드 수가 크게 차이 나는 경우에 조인 가능한 테이블을 찾는 성능이 좋지 못하다고 함.
  + 이를 방지하기 위해 Cardinality Proportion 식을 추가로 사용하여 두 컬럼간 레코드 수가 크게 차이나는 경우에 페널티를 주는 방식.
  + 논문의 **Cardinality Proportion식을 사용해 CNE-join의 성능은 높일 수 있다고 생각되지만 n-join-pair를 찾는데 사용하기에는 어렵다고 판단됨.**

* n-join-pair를 생성하는 알고리즘 구상 및 저번 랩미팅에서 교수님께 피드백 받은 부분 보완
  + 아래에서 밑줄 친 부분은 지난주 Weekly Report 내용과 동일함. 새로 적은 내용은 텍스트를 굵게 작성하여 지난주 내용에서 추가한 점을 쉽게 알아볼 수 있게 함.
* n-join-pair 생성 알고리즘 소개

스케치, 도표, 그림, 원이(가) 표시된 사진

자동 생성된 설명

* 1. CNE-join 결과로 나온 조인 가능한 테이블 쌍을 활용해 트리 자료구조를 생성
  + 조인 가능한 테이블쌍에 나온 모든 테이블을 한 번씩 루트 노드로 하여 위의 그림같이 여러 개의 트리를 만듦
  + 조인 가능한 테이블쌍에 나온 테이블 1개당 하나의 트리가 만들어짐. N개의 테이블이 나왔다면 N개의 트리를 만듦
* 2. 각 트리에서 나올 수 있는 모든 조인 가능한 경우의 수를 구함
  + 예시)   
    테이블 A가 루트 노드인 트리의 경우 A-B-C / A-B-D / A-C-D / A-B-C-D / A-B-E / …  
    테이블 B가 루트 노드인 트리의 경우 B-A-E / B-A-F / B-E-F / B-A-C / …
* 3. 각각의 경우에 대해 Inner Join을 수행하고 아래의 식을 만족하는지 확인한다
  + 아래의 식을 만족하면 n-join-pair에 부합하다고 판단할 것임
  + > Threshold
    - **n**-**join-pair가 잘 만들어졌는지 판단하기 위한 척도로 위의 식을 설정한 이유**
      * **루트 노드의 조인 컬럼 유일값 개수가 특정 테이블들과 조인한 이후 줄어들었다면, 조인 컬럼이 특정 테이블들과 겹치는 값이 적다는 것을 의미함**
      * **겹치는 값이 적다는 것은, 조인 결과를 통해 얻을 수 있는 정보가 적다는 것을 의미함. 이는 조인이 유의미한 결과를 내지 못한다고 생각할 수 있음**
      * **예를 들어, CNE-join 결과로 다음 두 개의 테이블 쌍이 조인 가능한 것을 알았음**

****

****

* + - * **2019\_world\_happiness.csv 테이블을 중심테이블로 하고 Country or region을 조인 컬럼으로 nobel\_latest.csv 테이블의 Death\_Country 컬럼과 조인한 결과를 직접보고 유의미한 정보가 생성되었다고 판단하였고 식의 값도 =0.86 로 높은 값을 가짐**
      * **하지만 위의 조인 결과에 추가로 Salary\_Data.csv 테이블을 Gender 조인 컬럼으로 조인한 결과를 직접보고 테이블에서 유의미한 정보가 생성되지 않았다고 판단하였고 식의 값도 =0.24로 매우 낮은 값을 가짐**
      * **위의 결과를 보면**  > Threshold **식이 n-join-pair가 잘 만들어졌는지 판단할 척도로 충분히 사용될 수 있다고 생각함**
  + **Full Outer Join과 중심테이블을 왼쪽으로 하여 Left Outer Join도 고려해봤지만, Inner Join과 큰 차이가 없다고 판단됨.**
    - **Full / Left Outer Join을 진행한 후 나온 조인 결과테이블에서 < Threshold를 만족하면 n-join-pair에 부합하다고 판단하는 방법도 생각해 보았지만,**  > Threshold **식과 큰 차이가 없다고 생각되어 폐기하기로 함.**
  + **박찬훈 연구원과 함께 식을 다른 방면으로 개량할 생각도 해보았지만, 마땅한 아이디어가 떠오르지 않았음.**
* **수행계획**
  + **관계형 테이블 임베딩 :**
    - CNE-join의 성능을 개선할 방법 고민
      * 조인하는 테이블의 개수를 3개 이상으로 확장해보기, 최적의 테이블 개수는 무엇인가? 이를 확인할 척도에 대해서도 고민
        + CNE-join의 결과로 나온 2개의 조인 가능한 테이블쌍 중에서 A-B, B-C 같이 테이블이 겹치는 쌍을 활용해 n-join-pair를 생성하는 알고리즘 및 이에 대한 융합도 산정 방안 고민 및 **소스코드 작성**(이전 CNE-join과의 일관성이 중요함)
      * 시간 단축 문제는 가장 나중으로 진행
    - 박찬훈 연구원과 DataMap 개발(Type A, B)
* **기타사항**