[Data Science] 데이터 사이언스 개념 - 8.토픽 모델 / 네트워크 분석 — 나무늘보의 개발 블로그

노트북: 첫 번째 노트북
만든 날짜: 2021-01-22 오전 7:02
URL: https://continuous-development.tistory.com/218?category=833358

Data Science

[Data Science] 데이터 사이언스 개념 - 8.토픽 모델 / 네트워크 분석

2021. 1. 14. 17:26 수정 삭제 공개

토픽모델

토픽 모델(Topic model)이란 문서 집합의 추상적인 "주제"를 발견하기 위한 통계적 모델 중 하나로, 텍스트 본문의 숨겨진 의미구조를 발견하기 위해 사용되는 텍스트 마이닝 기법

1.백오즈워즈와 음수 미포함 행렬 분해

백오브워즈 - 각 문서에서 단어의 출현 빈도를 행렬형식으로 정리한 것
대량의 문서중에 어떤 화제의 문장이 있는지 요약 정보를 얻고 싶을 때 사용한다.

희소행렬 - 값이 거의 0 인 행렬
문서 군의 요약을 구할 때는 이 희소행렬을 분해하는 방법을 사용하는 경우도 있다.

이것을 **잠재의미 분석(Latent Semantic Analysis, LSA)** 라고 부른다.
행렬의 각 요소가 양수인 성질에 주목하면 음수 미포함 행렬 분해라는 방법도 적용할 수 있다.

음수 미포함 행렬 분해 - 어떤 행렬 X를 X~WH로 분해하는 것이다.

이런 행렬 분석 기법을 이용할 때 한 가지 문제가 되느 것이 확률적으로 다루기 곤란하다는 점이다.
이 문제를 개선한 것이 **잠재 디리클레 할당(Latent Dirichlet allocation, LDA)**으로 토픽 모델이라고 불리는 모델의 일종이다.

2.동전 던지기 모델

앞면이 나올 확률이 @인 동전이 있다. 이 @의 값에 따라 결과가 나오는 방식이 달라진다.
@가 0.5이면 앞면이 나올 확률과 뒷면이 나올 확률이 같다.

3.확률 모델

사전정보를 포함하지 않는 추정값 -> 최대 우도 추정값
사전정보를 포함하는 추정값 -> 최대 사후확률 추정값
사전정보를 포함함으로써 데이터가 적을 때도 극단적인 추정값이 되는 것을 방지할 수 있다.

네트워크 분석

1.네트워크란?

네트워크 데이터란 노드(점)와 엣지(변)으로 표현되는 데이터 형식을 말한다.
관계형 데이터로 부르는 경우도 있다.
웹페이지의 경우 각 페이자가 노드에 해당하고, 하이퍼 링크가 엣지에 해당 한다.

노드에서 나온 엣지의 총수를 **출차수** 라고 부르고 , 어떤 노드에 들어오는 엣지의 총수를 **입차수** 라고 부른다.
엣지의 방향 정보를 무시하는 것은 **무향 네트워크**, 명시적으로 엣지의 방향을 다루는 것을 **유향 네트워크**라고 부른다.

노드와 엣지의 종류가 여러 개인 네트워크도 있다. 이런 네트워크는 멀티플렉스 네트워크, 헤테로지니어스 네트워크라고 부른다.
ex)기업간의 거래 관계와 기업과 사람을 견결하는 고용관계, 사람간의 네트워크가 기록된 데이터가 있다고 했을때 이때 노드는 두종류이고 엣지는 3종류가 된다.

2.페이지랭크와 검색 엔진

페이지 랭크 - 입차수가 많은 페이지가 좋은 페이지다 라는 아이디어를 기반으로 도입한 개념
이것은 구글이 검색엔진 기술로서 개발한 기법이다.
웹 페이지의 도달하는 확률을 정하고 출차수를 나눈다. 이러한 방법으로 페이지 랭크를 계산한다.

3.커뮤니티 추출

커뮤니티 추출 - 네트워크 구조에 기반해 노드를 몇 개 그룹으로 클러스터링 하는 기법
같은 그룹이면 엣지를 연결할 확률이 높고, 다른 그룹이면 엣지를 연결할 확률이 낮다라는 상황을 수학적으로 정의하고 이를통해
라벨이 일치하면 1 일치하지 않으면 0 으로 되는 함수를 만들어 사용한다.

이러한 함수가 최대회되도록 노드를 클러스터링함으로써 커뮤니티 추출을 실행할 수 있다. 이 식의 값을 모듈러리티라고 한다.
이 모듈러리티가 가장 높아지는 순으로 바텀업 방식으로 클러스터링해 감으로써 극대화 한다.

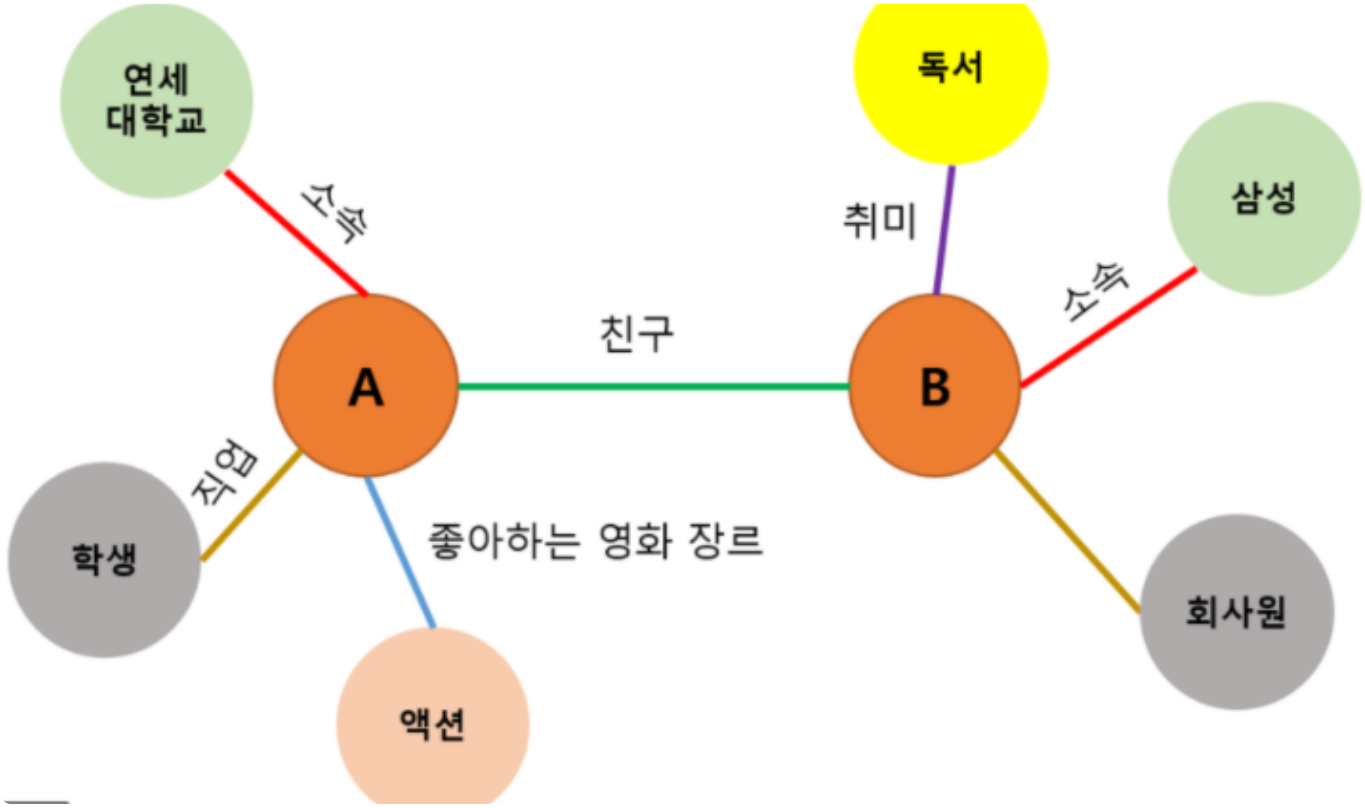
4.링크 예측

링크예측 - 주어진 네트워크에서 어느 링크가 빠져있는지, 어느 링크가 미래에 새로 생길 가능성이 높은지 예측하는 문제
ex) 페이스북 친구 추천, 과거 구매기록을 통해 미래에 소비자가 그 밖에 어느 상품을 구매할지 예측하는 문제
링크 예측을 하는 방법은 다양하지만 여기서는 **소속 통계**를 이용하는 방법과 **행렬분해**를 이용하는 방법을 보겠다.

소속 통계 - 링크가 없는 노드 간에 얼마만큼 공통된 인접 노드가 있는지 세어보는 것
친구 네트워크의 경우, 공통된 친구가 있으면 있을수록 그 두 사람이 친구이 가능성이 높을 거라는 예측을 나타낸다.

5.추천 시스템과 지식 그래프 보완

지식그래프의 보완이란 데이터베이스에서 빠진 정보를 보완하는데 자주 사용 된다.
중간에 연결다리의 역할을 하는 것을 보완해준다고 생각하면 된다.



본 내용은 그림으로 배우는 DataScience 데이터 과학을 참고한 내용입니다

'Data Science' 카테고리의 다른 글□

[Data Science] 데이터 사이언스 개념 - 10.딥러닝□

[Data Science] 데이터 사이언스 개념 - 9.신경망이 기초□

[Data Science] 데이터 사이언스 개념 - 8.토픽 모델 / 네트워크 분석□

[Data Science] 데이터 사이언스 개념 - 7.비지도 학습□

[Data Science] 데이터 사이언스 개념 - 6.분류문제□

[Data Science] 데이터 사이언스 개념 - 5.앙상블 학습□

네트워크분석 추천시스템 토픽모델

나아무늘보

혼자 끄적끄적하는 블로그 입니다.