**논문 요약**

**날짜: 2020.12.02**

**이름: 전우진**

|  |  |
| --- | --- |
| 논문제목 | Graph-based context-aware collaborative filtering |
| 논문 요약 | 추천 시스템은 빅데이터 시대에 정보 과부하로 인한 부작용을 완화하고 사용자가 거대한 아카이브로 이동하는 데 도움을 주는 인기 있는 도구다. 추천 시스템은 사용자가 관심을 가질 만한 제품, 비디오, 노래 또는 뉴스와 같은 개별 항목의 목록을 제공한다.  이는 구매, 조회, 클릭, 코멘트 등 사용자의 과거 상호작용을 통해 습득한 모델을 구성함으로써 이루어진다.  컨텍스트 인식 추천 시스템(CARS)은 사용자가 항목을 평가하거나 소비하는 특정 컨텍스트를 고려하여 전통적인 권장 접근방식을 확장하고 그에 따라 권장사항을 조정한다.  사전 필터링 및 상황별 모델링 접근법을 적용할 때 가장 큰 어려움은 데이터 범위다.  일반 사용자들은 대개 소수의 항목만 평가해 기존 권고 방식으로도 정확한 예측을 내기 어렵다.  문맥의 도입은 사전 필터링의 경우 관련 등급의 수를 더욱 감소시키거나 문맥 모델링의 경우 데이터 치수성을 증가시켜 둘 다 데이터 범위 증가를 초래한다.  특히 문맥 모델링의 또 다른 문제는 추가 데이터 치수의 도입으로 인한 계산 복잡성의 상당한 증가다.  본 논문에서는 희박한 데이터의 부정적 영향을 완화하기 위해 고안된 상황 인식 권장사항에 대한 그래프 기반 방법을 제안한다.  컨텍스트를 모델링하기 위해, 우리의 방법은 컨텍스트 조건에 따라 오리지널 아이템 등급을 나누어 가상 아이템을 도입하는 아이템 분할 기법을 사용한다(Baltrunas & Ricci, 2009). 그러나 단일 문맥조건으로 단일 분할을 허용하는 원래의 항목 분할 기법과 달리, 여기서는 모든 문맥조건 조합을 사용하여 항목을 분할한다. 우리는 새로운 사용자 항목 매트릭스를 초당적 사용자 항목 그래프로 나타내고, 그래프 상의 연결을 바탕으로 두 사용자 또는 두 항목 사이의 유사성을 계산한다.  그래프 표현을 사용하면 항목 분할 단계 이후 매우 희박한 직접 사용자 항목 연결 외에 사용자 또는 항목 간, 그리고 그래프 상의 타전적 관계의 형태로 사용자 또는 항목 간 거리 간접 연결을 모델링할 수 있다. 제안된 방법은 그래프의 전이적 관계를 활용함으로써 데이터 스패너성의 부정적인 영향을 줄일 수 있는 반면 컨텍스트 조건의 조합에 대한 검색을 요구하지 않는다. 우리의 방법은 효율적인 그래프 제안 방법을 사용하여 그래프 기반 유사성 계산을 수행할 수 있기 때문에 계산 복잡성은 노드 수(사용자 및 항목 수)와 에지 수(래팅 수)에서 선형이다.  이를 통해 보다 문맥적인 요인을 고려할 수 있고 실용적일 수 있을 만큼 효율적이다. 우리는 세 가지 상황에 따라 태그가 지정된 데이터 세트에 대해 제안된 방법을 실험적으로 평가했고 다른 상황 인식 방법과 비교했다.  Item splitting. 이 단계는 가상의 항목을 생성하여 다차원적 차원에 걸친 원래의 다차원 사용자 항목 등급 매트릭스를 2차원 매트릭스로 변환한다.  Graph construction. 이 단계에서는 이전 단계에서 형성된 등급 매트릭스의 그래프 표현을 작성한다.  Graph-based similarity calculation. 이 단계에서는 연관성이 그래프를 통해 측정할 때 사용자 간 또는 항목 간 유사성 점수를 계산한다.  Recommendation generation. 이것은 마지막 단계로, 이전 단계에서 계산된 유사성 점수와 함께 k-NNN을 사용하여 권고사항을 생성한다.  우리는 먼저 이 다차원 매트릭스를 가공의 아이템을 도입하여 2D 사용자 아이템 등급 매트릭스로 변환하는데, 각각은 오리지널 아이템과 컨텍스트 조건의 조합이다. 이 절차는 "아이템 분할"이라고 알려져 있다. 이를 통해 가지치기 또는 선택 단계 없이 가능한 모든 컨텍스트 조건 조합을 고려할 수 있다. 기존의 사용자 기반 k-NN 방법에서 사용자 u와 사용자 v의 유사성은 두 벡터 사이의 Pearson 상관관계 또는 코사인(cosine)으로 계산되며, 그 중 요소는 u와 v가 공동 등급 항목에 대해 부여한 등급이다.  항목 기반 k-NN에서 두 항목의 유사성은 두 항목을 공동 평가한 사용자를 기준으로 유사하게 계산된다.  평점 데이터가 희박한 경우, 두 사용자(또는 항목 기반 k-NN에서 두 항목을 공동 평가한 사용자 수)가 0과 같거나 작기 때문에 유사도를 계산할 수 없거나 불가능하다.  이 문제는 아이템 분할 후 더욱 심각해지며, 이로 인해 사용자-항목 데이터 범위가 더욱 넓어진다. 본 연구에서는 그래프를 이용하여 사용자간 또는 항목간 유사성을 계산한 후 사용자 기반 또는 항목 기반 k-NN에 대해 계산된 유사성 점수를 사용할 것을 제안한다.  제안된 알고리즘의 핵심 단계는 유사성 점수 계산이다. 이 단계는 위에서 설명한 것처럼 매트릭스 곱셈으로 할 수 있지만, 계산적으로 비용이 많이 든다. 그래프 G는 가장자리 수가 희박하기 때문에 그래프 G에 퍼지는 활성화 알고리즘을 사용함으로써 매트릭스 곱셈을 피할 수 있다. 본 연구에서는 Weston, Elisseff, Zoo, Leslie, Noble(2004)의 확산 활성화 알고리즘을 채택하여 사용자 쌍과 페어 항목의 유사성을 계산한다. 활성 사용자 u 및 최대 경로 길이 L에 대한 확산 활성화 알고리즘은 그림 5와 같다. 이 알고리즘은 기존 등급에 해당하는 기존 에지만 고려한다는 점에 유의한다.  우리의 실험에서 세 가지 데이터 세트를 사용했는데, DepaulMovie, MovieLens, InCarMusic. 추천 업무에는 등급 예측과 항목 추천(또는 상위 N 권장사항)의 두 가지 유형이 있다.  등급 예측의 경우, 가장 인기 있는 평가 지표는 평균 오차(MAE), 평균 제곱 오차(RMSE), 평균 예측 오차(MPE)이다.  항목 예측은 Precision@N, Recall@N, MAP@N 와 같은 순위 지표를 사용하여 평가할 수 있다.  본 작업에서는 상위 N점수를 사용자에게 추천하는 상위 N개 추천에 초점을 맞춘다.  우리는 또한 우리의 방법을 다른 상위 N 권장 알고리즘과 비교한다.  따라서 상위 N 권장 설정에 공통적인 정밀도 @ N(또는 P@N )과 평균 평균 정밀도(MAP@N )를 사용하였다.  순위에 따른 항목 리스트가 주어지면, 정밀도 @N은 상위 N위 사이에서 관련되는 항목의 비율이다. 우리는 사용자들의 10배 교차 검증을 이용했다. 구체적으로는 사용자 집합을 유사한 크기의 10개의 하위 집합으로 나눈다. 테스트 세트로 하나의 서브셋을 보유했고, 리메인딩이 트레이닝 세트로 하위셋을 보유했다. 여러 테스트 세트에 대해 이 절차를 10회 반복했으며 평균 측정값을 10배 이상 보고하였다. 테스트 세트의 각 사용자에 대해 정격 항목(타임스탬프를 기반으로 한 마지막 정격 항목)의 20%를 보류했다. 그런 다음 시스템에 의해 만들어진 사전 받아쓰기를 보류된 데이터와 비교하여 평가 지표를 계산한다. |
| 논문 의의 | 데이터 희박성으로부터 영향받을 위험이 낮은 상황별 세그먼트를 고려할 수 있다. 이 방법은 또한 사용자 수 + 항목과 등급에 선형적인 계산 복잡성을 가지고 있는데, 이것은 많은 상황적 요소에서 분할된 결과 항목이 많은 도메인에서 적용가능성에 중요하다. |
| 논문의 한계점 및 기타 | 다른 domain에 적용할 필요성이 있다. |