<PROGRESS MEETING>

장석규

[TITLE]

- 취향에 관계없이 인기있는 컨텐츠의 영향을 줄인 CF 추천 시스템

[HYPOTHESIS]

영상 추천은 다른 도메인에 비해 화제가 되는 컨텐츠를 취향에 관계 없이 보는 것 같다. 현재 영화 추천 시스템은 해당 컨텐츠들도 함께 고려하고 있지만 그런 컨텐츠는 오히려 개인화된 추천에 방해가 되는 것 같다.

[PROOF]

1. 개인화된 추천에서도 사용자와 상호작용했던 모든 컨텐츠를 고려하고 있는가?

2. 해당 컨텐츠가 추천에 방해가 되는가?(성능에 악영향을 주는가?)

[CHALLENGE]

자신의 취향에 맞으면서 인기있는 컨텐츠를 볼 수 있고, 취향에 맞지 않음에도 인기있는 컨텐츠를 볼 수 있다. 전자의 경우에는 사용자의 특성을 더 잘 표현하는 데이터로 활용될 수 있지만 후자는 오히려 사용자 특성에 맞는 콘텐츠 추천에 방해가 된다. 즉 이 두가지를 구분해 해당 데이터를 사용할 것인지 하지 않을 것인지 판단할 필요가 있다.

‘취향에 관계 없이 인기있는 컨텐츠’를 어떻게 판단할 것인가

- A 사용자(추천을 제공해야 할 대상)가 지금껏 상호작용했던 컨텐츠를 C = {a,b,c,d,...}라고 해보자. 이때 해당 컨텐츠에 대해 각각 취향에 관계없이 인기 있었는가를 판단한다.

- 내가 생각한 메소드: C가 지금껏 본 컨텐츠 C = {a,b,c,d,...}에 대해 각 컨텐츠가 취향에 관계없이 인기있었는가를 파악하기 위해 컨텐츠 각각에 대해서 다음 연산을 수행한다:

    1. a에 대해 해당 노드의 엣지를 랜덤 샘플링한다.
    2. 랜덤 샘플링한 엣지와 연결된 노드들(영상 시청자들)에 대해 A 사용자와의 similarity를 계산한다.
    3. A 사용자와 유사하지 않은 사용자가 많다면 해당 노드를 마스킹한다.

- [랜덤 샘플링하는 수와 similarity 메소드, 많다의 기준을 구체적으로 정해야할 것 같다]

[RELATED RESEARCH]

Constructing a user similarity network to remove adverse influence of popular objects for personalized recommendation (2013)

- 해당 논문에서는 인기있는 객체의 부정적인 영향을 극복하기 위한 네트워크 기반 협력 필터링 접근법을 제안한다. 각 사용자에 대해 다른 사용자 간 가장 약한 관계를 걸러내기 위해 knn 을 이용해 k-최근접 이웃 네트워크를 얻거나 임계값에 따라 사용자간 약한 관계를 걸러내어 필터링 네트워크를 생성한다.

- 인기있는 컨텐츠를 다른 사용자와의 관계 중 약한 관계로 정의 (공통 관심사 중 약한 관계는 관심이 없다고 할 수 있으니까)

- 단점: 약한 관계를 임계값으로 설정하여 사용하는 것은 매우 실험적이고 상대적으로 약한 관계라고 해서 무의미한 관계가 아닐 수 있다. 또 약한 관계가 인기있는 컨텐츠였다 라고 하는 것은 비약이라고 생각한다.

[comment]
콜드 스타트 문제에서 인기있었던 아이템을 학습과정에서 뺀다.

- 취향에 관계없이 인기있는 콘텐츠를 어떻게 정의할 것인가

- 취향에 관계없는 아이템의 특성

- 취향에 관계된 영역

- 아이템 소비 = 자기 취향 + 대중 취향 선행연구와 무엇이 다른가
- popularity bias 를 제거하면 사용자 선호만 남게될 것이다.

이번 주에는 각 데이터 도메인별로 차이와 문제가 무엇인지 생각하고 주제에 대해 생각해보기로 했다.

영상 추천에서 취향에 관계없이 인기있는 컨텐츠를 마스킹하여 CF 필터링에 적용하면 개인화된 추천에 더 좋은 성능을 제공할 것이라고 생각한다. 개인화된 추천은 개인의 특성을 더욱 많이 반영해야 한다고 생각하는데, 영상같은 경우는 취향이 아니더라도 인기가 많다면 한번 보는 경우가 많다고 생각했다. 이에 관련 논문을 찾아본 결과 현재 개인화된 추천에서 사용자와 상호작용했던 모든 컨텐츠를 고려하고 있고 해당 컨텐츠가 추천에 방해가 될 수 있음을 언급했다. 해당 연구의 경우 약한 관계를 쳐내고 knn을 이용해 유사한 이웃 네트워크를 구축하는 것이 인기있는 컨텐츠의 영향을 최소화할 수 있다고 하였다. 그러나 해당 의견이 성능 향상에는 도움이 될지는 몰라도 인기있는 컨텐츠가 아니거나 유의미한 관계가 될 수도 있고, k 를 하나의 값으로 설정하는 것도 유연하지 않을 것이라고 생각해 새로운 방법을 제안해보고 싶다. 내가 제안하고자 하는 방식은 하단에 첨부하였다.

다음주에는 이번 주제를 좀 더 구체화하거나 새로운 주제를 찾아봐야 할 것 같다.