

콘텐츠 정보 지식구조를 이용한 협업 추천 시스템

김준우, 박준영, 이문용

한국과학기술원

e-mail : junu@kaist.ac.kr, j.park89@kaist.ac.kr, munyi@kaist.ac.kr

Content Knowledge Structure based Collaborative Filtering Recommender Systems

Junu Kim, Juneyoung Park, Mun Y. Yi
Korea Advanced Institute of Science and Technology

요약

애플리케이션에서 고객들에 의해 생성된 평가정보는 해당 콘텐츠에 대한 고객별 선호도 정보로 볼 수 있기 때문에, 개인에게 맞춤형 추천 시스템을 설계하기 위해서 매우 중요하다. 현재 추천 시스템 분야에서 가장 많이 사용되고 있는 사용자 기반 추천 시스템은 사용자의 평점 정보만을 가지고 유사도를 측정하여 추천에 사용하고 있다. 그러나 이러한 평점 정보만을 가지고 사용자 유사도를 도출하는 것은 정밀하지 못할 수 있다.

따라서 본 연구에서는 사용자의 평점 정보 뿐만 아니라 콘텐츠의 내용을 활용하여 사용자의 선호 콘텐츠를 지식구조의 형태로 나타냄으로써 콘텐츠와 사용자의 관계를 유기적으로 표현하였다. 이와 같은 사용자의 지식구조를 바탕으로 사용자간의 유사도를 평가하고 추천에 활용하였고, 실험 결과 제시된 방법으로 더 우수한 성능을 얻을 수 있는 것으로 나타났다.

1. 서론

최근 콘텐츠 산업 전반에 걸쳐 사용자의 니즈를 자동으로 발견하고 해당 사용자에게 적합한 콘텐츠를 추천해주는 추천시스템이 활발하게 사용되고 있다. 특히, 사용자들이 남긴 평점 정보 기반의 추천 기법들은 아마존, 넷플릭스 등의 콘텐츠 제공 서비스들에서도 널리 사용되고 있는 성공적인 추천 전략이다. 사용자가 남긴 정보를 추천에 활용하는 방법에는 크게 사용자 기반 협업 필터링(User-based Collaborative Filtering) 기법과 아이템 기반 협업 필터링(Item-based Collaborative filtering) 기법이 있다.[6]

그 중 본 연구에서는 사용자 기반 협업 필터링 기법에서 유저간의 유사도를 계산하는 pearson 상관계수를 대체하는 유사도를 계산하여 모델을 정교화 하고자 하였다. 이와 관련한 추천 시스템 분야에서는 사용자가 콘텐츠를 사용한 시간, 대중성, 신뢰도 등을 활용하는 기법도 소개되었다.[4][7]

그러나 위와 같은 것들은 실제 다른 여러 콘텐츠들에서는 추천이나 공감 등 대중성이나 신뢰도를 측정할 수 있는 요소가 없을 수 있기 때문에 공통적으로 적용할 수 없는 한계가 있다. 이에 본 연구에서는 콘텐츠의 내용을 유의미한 형태의 지식 구조로 재구성하여 세부 내용 간의 상관관계를 적극 활용하여 사용자간의 유사도를 구한다. 또한 사용자들간의 관심 콘텐츠의 내용이나 선호도의 유사도를 더욱 정밀하게 추출함으로써 콘텐츠들의 추천 정확도를 향상하는 방법을 제안한다.

본 논문에서 제안하는 방법을 통해 기여하는 바는 텍스트를 통해 사용자간의 유사도를 사용하고자 하는 분야 어디든 본 논문의 기법을 사용할 수 있다는 것과 텍스트 가중치에 사용자의 선호도를 반영함으로써 보다 개인의 성향에 밀접한 콘텐츠의 추천이 가능하다는 것이다.

2. 관련연구

2.1 협업추천 시스템

추천 시스템은 고객별로 상이한 선호도를 감안하여 해당 고객이 만족할 수 있도록 맞춤화된 상품 목록을 생성, 이를 고객에게 제안하는 기법으로서 많은 선행 연구를 통해 다양한 추천 방법론 혹은 시스템들이 제안되어왔다. 그 중에서도 협업 필터링을 근간으로 하는 협업 추천 시스템이 가장 성공적인 기법으로 평가되고 있다.[6] 협업 추천 시스템은 고객 프로필에 따라 비슷한 이웃들을 찾고 이웃들이 구매하지 않은 상품 중 구매 가능성이 높은 상품을 찾아 추천하는 기법이다. 비슷한 이웃을 찾기 위해서는 고객 프로필을 필요로 한다. 고객 프로필은 온라인 상점을 방문하는 m 명의 고객들이 상점에서 취급되는 n 개의 상품에 대해 가지는 선호도 정보를 표현하기 위해 논리적으로 $m \times n$ 의 행렬 형태를 취한다.[6]

*본 연구는 중소기업청의 창업성장기술개발사업의 일환으로 수행하였음. [S2295562, 한류콘텐츠 사용자를 위한 소셜테이스트 분석 기반 다중 콘텐츠 검색/추천 플랫폼 개발]

2.2 지식구조

지식구조란 일반적으로 특정 분야에 대한 인간의 지식을 의미 있고 유기적인 형태로 나타내는 것을 의미 하며 유사한 의미로 멘탈 모델(Mental Model), 인지 스키마(Cognitive Schema) 등으로도 불린다. 단순한 개념, 단어 들의 집합이 아니라 이러한 객체들의 의미있는 관계를 정립함으로써 성립된다. 다수의 선행 연구에서 인간의 지식을 자동으로 지식구조로 변환시키고자 하였고 자연어 처리 기법들을 활용하여 가능케 하였다.[2] 지식구조가 인간의 지식의 형상을 표현한다는 점은 추천 및 검색 분야에서 다양한 가능성을 제시해 왔고, 영화, 애플리케이션 등 다양한 도메인에서 그 가능성을 증명하였다.[1][2][3]

3. 콘텐츠 정보 기반 사용자 지식구조

콘텐츠 정보에 기반하여 사용자간의 유사도를 도출하기 위해 기존 연구에서 제시된 방법[2]을 이용하여 콘텐츠를 지식구조화 하였다. 본 연구에서는 오픈 소스 한글 형태소 분석기 KMORAN2.4 을 사용하여 콘텐츠의 텍스트 중 명사를 지식구조의 개념으로 활용하였고 개념들 간의 연관관계를 도출하여 지식구조를 형성하고 지식구조간의 비교를 통해 사용자간의 유사도를 도출하였다.

3.1 사용자가 이용한 콘텐츠 키워드 추출

전통적인 키워드 추출 기법은 문서 콘텐츠의 내용을 이용하는 통계적 기법이 있다. 통계적 기법의 보편적인 기법으로 정보검색 분야에서 사용되는 TF(Term Frequency), TF*IDF(Inverse Document Frequency) 가 주로 사용된다. [2]

본 연구에서는 콘텐츠를 나타내는 텍스트 정보가 충분치 않아 TF 를 기준으로 키워드를 추출하였다.

3.2 키워드 간 관계 추출

키워드로 추출된 단어 간 관계를 통해 사용자 간의 지식구조를 구축하고 개념간의 공기 정보를 통해 연관 관계를 계산하였다. 연관 관계란 개념적으로 밀접한 관련이 있으나 동의어나 유사 동의어인 등과 관계에 포함되지 않는 두 단어 간 의미적, 심리적 연관 정도를 나타낸다. 연관 관계 추출 기술로서 전통적으로 단어의 공기 정보(Co-occurrence)를 이용하는 방법이 있다. 공기 정보란 두 단어가 동일한 문서, 문장, 구 등에 같이 발생하는 현상을 말하며, 더 자주 발생 할수록 두 단어가 밀접한 관계를 가지고 있다는 전제에 기반하며, 키워드 추출의 시초가 된 Salton 의 1989년 연구에서 측정 방법이 제시 되었다.[5]

본 논문에서는 콘텐츠 간의 핵심 개념 간 연관관계를 추출하기 위해서, 단어 쌍의 공기 정보(Co-occurrence)를 이용한다.

3.3 사용자 유사도 분석

사용자 기반 협업 추천시스템은 가장 기본적으로 Pearson 상관계수로써 사용자 유사도를 사용한다. 본 연구에서는 유사도를 측정하는 세가지 방법을 제안한다. 첫번째는 키워드 간의 공기 정보를 이용하는 방법이고 두번째는 공기 정보(Co-occurrence)에 사용자의 평점 정보를 매핑한 방법이다. 마지막으로 세번째는 카테고리별 유사도를 이용한 방법이다. 공기 정보를 이용하는 것은 위 두가지 방법과 동일하지만 여기에서 공기정보의 키워드로써 사용되는 것이 사용자간의 공통된 단어가 아니라, 미리 뽑아놓은 카테고리별 단어들과의 공통된 단어이다.

3.3.1 사용자간의 유사도 행렬 유사도 추출

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

<그림 1 코사인 유사도>

사용자간의 유사도를 계산하기 위하여 본 논문에서는 코사인 Cosine similarity(그림 1)를 사용하였다.

예를 들어 사용자(1)이 본 콘텐츠에서 뽑아낸 키워드가 W1, W2, W3, W4, W5 라고 하고, 사용자(2)로부터 추출한 키워드가 W2, W3, W4, W7, W8 이라고 하면 유사도를 구하기 위하여 공통 키워드인 W2, W3, W4 를 최종적으로 행렬의 구성 요소로 사용한다. 이를 통해서 W2, W3, W4 로 이루어진 각 사용자의 3x3 행렬을 통해서 코사인 유사도를 구하고 이를 사용자간의 유사도 지표로써 사용한다.

3.3.2 공기 정보 행렬 유사도 추출

지식구조를 형성하는 가장 기본적인 형태에서 사용자들간의 공통 된 개념의 공기 정보를 도출하여 행렬을 나타내면 표 1,2 와 같다.

사용자(1)	W1	W2	W3
W1	-	3	2
W2	-		1
W3	-	-	-

<표 1 사용자(1)의 공기 정보 행렬>

사용자(2)	W1	W2	W3
W1	-	3	2
W2	-		1
W3	-	-	-

<표 2 사용자(2)의 공기 정보 행렬>

이를 통해서 두 사용자의 코사인 유사도를 구할 수 있다. 사용자(1)과 사용자(2)의 코사인 유사도는 1.0 이다.

3.3.3 사용자 선호도 반영 유사도 추출

3.3.2 의 공기 정보의 연관 관계는 단순히 동시 발생 횟수를 카운트 한다. 그렇게 되면 평점을 1 점을 주고 본 콘텐츠와 5 점을 주고 본 동일한 콘텐츠 간의 차이를 반영할 수 없다. 콘텐츠가 구성하고 있는 키워드가 동일하기 때문에 공기 정보로는 사용자의 선호도의 차이를 알 수 없기 때문이다. 이러한 차이를 반영하기 위하여 2 번째 방법으로 사용자가 평가한 선호도를 공기 정보에 포함한다.

3.3.2 와 같은 형태에서 연관 관계 점수에 가중치로써 사용자의 선호도를 부여한 표의 예는 아래와 같다.

사용자(1)	W1	W2	W3
W1	-	65	40
W2	-		3
W3	-	-	-
사용자(2)	W1	W2	W3
W1	-	30	4.5
W2	-		60
W3	-	-	-

<표 3 연관 관계에 선호도를 반영한 행렬>

표 3에서 나타내는 것은 동일한 콘텐츠를 본 사용자들 간의 유사도가 선호도에 따라 달라질 수 있음을 보여준다.

위의 행렬에서 코사인 유사도를 도출하면 0.45 가 계산된다. 이는 동일한 콘텐츠를 보더라도 선호도의 차이를 반영할 수 있기 때문에 더욱 정밀하게 사용자 간의 유사도를 계산할 수 있다.

3.3.4 카테고리 정보 행렬 유사도 추출

학습을 위한 데이터 셋에서 카테고리 별로 텍스트를 추출하였다. 그런 후, 총 12 개의 카테고리 (동물, 드라마, 패션, 음식, 게임, 유머, 인간관계, 이슈, 음악, 스포츠, 연예, 여행) 별로 키워드를 추출하였다. 이를 활용하기 위해 사용자의 텍스트들에서 각 카테고리 별로 추출된 키워드들에 대한 공통된 단어를 추출하고 카테고리 별 유사도를 도출하게 된다.

먼저 카테고리 별 키워드로써 사용될 단어들을 추출한다. 그런 다음, 사용자에게서 뽑아낸 키워드들과 각 카테고리에서 추출한 키워드들 간의 공통된 단어를 선정한다. 이렇게 추출된 공통된 단어들을 기준으로 카테고리의 텍스트 정보에서 공기 정보 행렬을 도출하고, 사용자 텍스트 정보에서 또한 공기 정보 행렬을 계산한다. 이를 통해서 카테고리별 관심도(유사도)가 나오게 되고 총 12 개 카테고리에 대한 관심도(유사도)가 계산된다. 모든 카테고리에 대한 정보가 나오게 되면 이를 통해서 사용자간의 코사인 유사도를 계산함으로써 사용자간의 카테고리 유사도를 도출할 수 있다. 이를 통해서 사용자간의 유사한 카테고리 취향을 도출할 수 있다.

예를 들어 음악이라는 카테고리에서 30 개의 키워드가 선정되었다고 하면, 사용자에게서 뽑아낸 키워드들 중에서 이 30 개의 키워드와 공통된 단어를 뽑아 이를 통해 음악 카테고리 공기 정보 행렬, 사용

자가 갖고 있는 음악 카테고리에 대한 공기 정보 행렬을 각각 계산한다. 그런 후, 코사인 유사도를 계산하여 카테고리 1 개에 대한 유사도를 도출한다. 이런 과정을 모두 거치면 사용자마다 12 개의 카테고리에 대한 유사도가 나오게 되는데 이렇게 도출된 관심도를 행렬을 통해서 사용자간의 유사도를 추출한다.

앞서 설명된 것처럼 카테고리에 따라서 사용자간의 유사도를 도출하는 방식은 3.3.2, 3.3.3 의 방법과 약간의 차이가 있다.

음악	W1	W2	W3
	W4	W5	W6
	W7	W8	W9
사용자(1)	W1	W2	W3
	W10	W11	W12
	W13	W14	W15

<표 4 카테고리 “음악”에 대한 공통된 키워드 >

표 4를 참조하면, 음악 카테고리안에 있는 키워드와 사용자에게서 뽑아낸 키워드 중 공통된 것은 W1, W2, W3 다. 이를 통해서 음악의 모든 텍스트 정보에 있는 W1, W2, W3 의 공기 정보 행렬을 도출한다. 마찬가지로 사용자(1)의 텍스트 정보를 통해 W1, W2, W3 의 공기 정보 행렬을 추출한다. 이렇게 사용자(1)의 “음악” 카테고리에 대한 관심도를 구할 수 있고 총 12 개의 카테고리에 대한 관심도를 계산할 수 있다. 이렇게 사용자마다 12 개의 카테고리가 계산이 되면 이를 이용하여 사용자간의 코사인 유사도를 계산함으로써 새로운 사용자 유사도 지표로써 사용하게 된다.

이 방법을 통해서 사용자가 어느 카테고리에 관심이 있는지를 도출해, 사용자간의 유사도를 계산할 수 있다.

4. 실험

4.1 실험 설계

제시된 모델들의 성능을 검증하고 기존의 모델과의 비교를 통해 성능 향상을 실험하기 위하여 미디어 아티클 45865 건에 대하여 986 명의 사용자가 2016년 1월 26 일부터 2016년 2월 16 일까지 약 3 주간 사용하며 남긴 평가 정보 46457 건을 이용하여 실험을 진행하였다. 46457 건의 데이터 중 42707 개를 모델들의 트레이닝 셋으로 이용하였고, 3750 개를 테스트 셋으로 활용하여 모델들의 성능을 평가하였다.

실험은 먼저 사용자 기반 협업 필터링에서 가장 기본적으로 사용되고 있는 pearson 상관계수를 기준으로 하였다. 이렇게 해서 먼저 제안한 모델(1)과 선호도를 반영한 모델(2), 카테고리 별 유사도를 반영한 모델(3)을 비교하였다. 모델간의 성능차이를 분석하기 위하여 본 연구에서는 RMSE(root mean square error) 와 MAE(Mean Absolute Error) 측정 지표를 이용하였다.

실험 평가에 활용된 RMSE 와 MAE 는 추천 방법론 (혹은 구현시스템)에 의해 예측된 선호도와 실제 선호도 간의 오차를 계산해주는 것으로서 추천시스템의 예측 정확성 평가에 널리 활용되는 지표이며 0 에 가까울 수록 좋은 성능을 나타내는 것이다.

4.2 실험 결과

평가 표	RMSE	MAE
기본 모델	1.281	1.034
제안 모델(1)	1.082	0.777
제안 모델(2)	1.070	0.779
제안 모델(3)	1.060	0.804

< 표 4 모델간의 평가 표>

표 4에서 보는 것과 같이 기본 모델보다 본 연구에서 제안한 모델들이 좋은 성능을 보였다. 이는 제안된 방법으로 측정된 유사도를 통해 사용자 선호도의 예측 정확도를 향상시킨다는 것을 의미한다.

제안 모델(1)을 통해 콘텐츠 기반의 사용자 유사도가 사용자 선호도 예측 정확도를 향상 시킬 수 있다는 점과 사용자 선호도의 융합으로 더욱 정확한 예측이 가능하다는 점은 정확한 추천에 고무적인 결과이다. 또한 카테고리 정보와 같이 사용자의 취향을 구분 짓는 정보를 통해서 한층 더 정확한 예측이 가능했다.

위의 실험 결과에서 보이듯이 제안모델(2) 와 제안모델(3)의 경우, MAE 평가지표에서 제안모델(1)보다 성능이 낮게 평가되었으나 RMSE 평가지표에서는 더 높은 성능을 보였다. 그러나 사용자의 평가가 급격하게 변화하는 미디어 데이터의 특성상 예측 점수의 오류에 더 민감한 지표인 RMSE 가 본 실험에 더 적합하다고 고려할 수 있다.

5. 결 론

본 논문에서는 콘텐츠의 핵심 개념과 그 개념들 간의 관계도로 구축되는 지식구조를 활용한 사용자 기반 협업 추천 시스템을 제안하였고 실험 결과, 기존의 기법과 비교하여 더 우수한 성능을 보였다. 평점 정보뿐만 아니라 사용자가 이용한 콘텐츠의 내용을 반영하여 사용자의 선호도를 분석하고 선호도가 유사한 사용자들을 통해 우수한 추천이 가능했을 뿐만 아니라, 카테고리 정보와 같은 사전에 분석이 가능한 콘텐츠의 특성을 적용할 수 있는 가능성을 실험을 통해 증명할 수 있었다.

추후 연구에서는 콘텐츠를 통한 사용자간의 유사도 뿐만 아니라 추천할 콘텐츠 자체에 대한 사용자의 평가를 추측하여 추천에 융합하여 활용할 수 있는 다방면적인 후속 연구가 진행되어야 한다.

참고문헌

- [1] Bae, Donghwan, et al. "AppTrends: A graph-based mobile app recommendation system using usage history." Big Data and Smart Computing (BigComp), 2015 International Conference on. IEEE, 2015.
- [2] Kim, Hyung W., and Mun Y. Yi. "Empirical validation of an automated method of knowledge structure creation from single documents." ICT and Knowledge Engineering (ICT & Knowledge Engineering), 2011 9th International Conference on. IEEE, 2012.
- [3] Kim, Sansung, et al. "Exploiting Knowledge Structure for Proximity-aware Movie Retrieval Model." Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014.
- [4] Lee, Grace E., Keejun Han, and Y. Yi Mun. "Incorporating Distinct Opinions in Content Recommender System." Information Retrieval Technology. Springer International Publishing, 2015. 109-120.
- [5] Salton, Gerard. "Automatic text processing: The transformation, analysis, and retrieval of." Reading: Addison-Wesley (1989).
- [6] Sarwar, Badrul, et al. "Item-based collaborative filtering recommendation algorithms." Proceedings of the 10th international conference on World Wide Web. ACM, 2001.
- [7] Yang, Xiwang, et al. "A survey of collaborative filtering based social recommender systems." Computer Communications 41 (2014): 1-10