Ⅰ 논문정보&요약

제목	신뢰관계 네트워크 분석을 활용한 추천 알고리즘 개선
저자	최 슬 비
출판년도	2017

기존 colaborative filter 알고리즘에 추가해서 소셜 네트워크 분석(이하 SNA)수치를 더하여 소비자 최적의 cf기반 하이브리드 알고리즘을 설계하였다.

Ⅱ 연구배경

협업 필터링(Collaborative Filtering)은 추천시스템을 구현하는 대표적인 추천 알고리즘으로 알려져 있으며, 유용성과 정교성 면에서 가장 우수한 성능을 보이는 알고리즘으로 평가받고 있다. 협업 필터링은 내용기반 필터링이나 지식기반 필터링과 같은 여러 가지 추천 알고리즘 기술들과 비교했을 때, 상대적으로 높은 정확도를 보인다는 장점으로 산업계나 학계에서 많은 관심과 함께 연구 및 활용되고 있지만, 사용자 평가 점수에만 기반하여 상품 및 서비스를 추천한다는 한계점을 갖는다

이러한 한계점을 극복하기 위해, 본 연구에서는 사용자의 평점뿐만 아니라 사용자 간 신뢰관계를 추가적으로 추천 알고리즘에 반영하는 새로운 접근법을 제안하고자 한다.

기존 협업 필터링은 4가지 단계로 구성되어 있다.

1. 행렬구성

사용자 평점 데이터를 기반으로 행렬을 구성

	상품 1	상품 2	상품 3	…	상품 N-1	상품 N
사용자 1	5	1	3		4	2
사용자 2	4		2		4	3
사용자 3	2	3	2		4	5
…						
사용자 M-1	3	4	4		4	1
사용자 M	3	5	3		3	?

< 그림 1 > 사용자-상품 평가점수 행렬 구성의 예

2. 유사도 산출

사용자 간 유사도를 산출하고 이를 바탕으로 선호도가 유사한 사용자를 탐색한다. 이 때, 유사도를 측정하는 방법으로 코사인 유사도(Cosine similarity)와 피어슨 상관계수(Pearson correlation coefficient, PCC)가 일반적으로 사용되는 방법인데 이 중, PCC가 가장 대표적으로 사용되며 계산식은 다음과 같다.

$$S_{x,y} = \frac{\sum_i (R_{x,i} - \overline{R_x}) \cdot (R_{y,i} - \overline{R_y})}{\sqrt{\sum_i (R_{x,i} - \overline{R_x})^2} \cdot \sqrt{\sum_i (R_{y,i} - \overline{R_y})^2}} \quad (1)$$

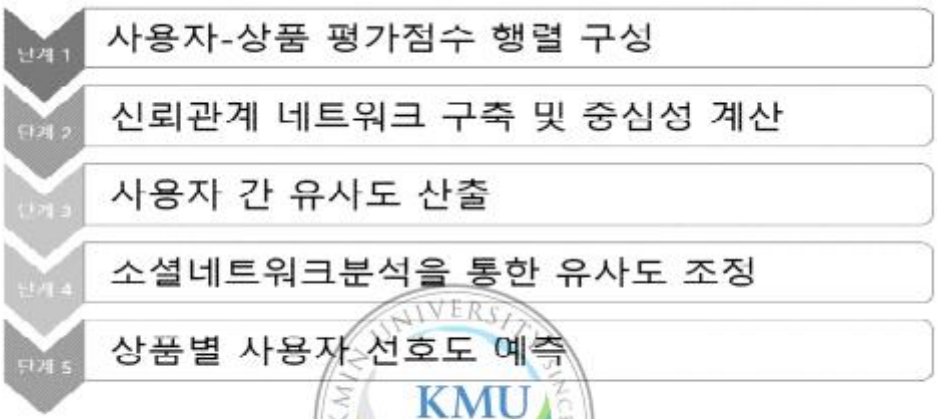
3. 추천 대상자와 선호도가 유사한 n명의 사용자 선택

4. 상품별 선호도 예측 및 이를 바탕으로 적합한 추천 상품을 최종적으로 결정

$$P_{x,i} = \overline{R_x} + \sum_{z \in NN} (R_{z,i} - \overline{R_z}) \times \frac{S_{x,z}}{\sum_{z \in NN} |S_{x,z}|} \quad (2)$$

Ⅲ 연구내용

소셜네트워크분석(Social Network Analysis, SNA)은 연결된 네트워크의 구조를 이해하기 위해서, 각 노드(node)들의 개별적인 특성에 초점을 맞추기 보다는 네트워크의 전체적인 연결구조 또는 연결 상태에 초점을 맞추고 그 패턴을 분석하는 기법.



< 그림 2 > 제안 알고리즘 단계별 절차

기존 4단계에서 1단계 추가(2단계) 여기서는 외향 중심성과 내향 중심성을 계산 (추천 신뢰가 높은사람, 개방적 성향) 이를 간단하게 측정하기 위한 방법으로는, 한 액터 와 직접적인 연결관계를 갖는 다른 액터들 개수의 합으로 구할 수 있다

(여기서는 UCINET6이라는 소프트웨어 분석 패키지 사용하여 계산)

4단계에서는 앞서 2단계에서 도출된 신뢰관계 네트워크 데이터 분석 결과와 3단계에서 계산된 사용자 간 유사도를 통합하여 전체적으로 사용자간 유사도를 재조정한다. 이를 반영한, 새로운 사용자 간 유사도를 의미하는 다음 식 (3)과 같이 산출된다.

$$S^*_{x,y} = amp_{x,y} \times S_{x,y} \quad (3)$$

사용자 간 신뢰관계 행렬(User to User Trust Matrix) 산출이나 신뢰관계를 반영하여 사용자 간 유사도를 산출하는 제안 알고리즘은 Microsoft Excel VBA 실험용 소프트웨어로 구현하였다.

위 식에서 $S^*_{x,y}$는 신뢰관계 네트워크 분석 결과를 추가로 반영하여 다시 산출된 사용자 x와 사용자 y의 유사도를 뜻한다. $amp_{x,y}$는 사용자 x와 사용자 y 사이의 유사도를 확대시키기 위한 조정계수이다.

Trust CF-All : 첫 번째 접근법은 신뢰관계 네트워크에서 사용자 y의 내향 연결정도 중심성을 고려하여 유사도를 확대하는 것이다. 예를 들어, 내향 연결정도 중심성의 값이 높은 사용자 A가 있을 때, A는 다른 사용자들에게 높은 신뢰를 받는다고 볼 수 있기 때문에 사용자 A의 추천을 더 적극적으로 수용할 가능성이 높다. 이러한 특징을 반영하여, $amp_{x,y}$를 다음의 식 (4)와 같은 방식으로 계산한다.

$$amp_{x,y} = 1 + (IC_y)^{\mu} \quad (4)$$

위의 식에서 IC_y는 사용자 y의 내향 연결정도 중심성이고 μ는 승수이다. 여기서 승수 μ는 내향 연결정도 중심성의 값을 보다 비중 있게 반영하기 위한 조정계수로써, 탐색의 시행착오(trial and error)를 거듭하면서 최적의 값을 찾아야 한다.

Trust CF-Conditional : 다음의 접근법은 첫 번째 접근법의 확장된 방법으로, 사용자 y의 내향 연결정도 중심성을 고려할 때, 외향 연결정도 중심성이 특정 임계치(threshold) p이상인 사용자 x에 대해서만 고려하는 것이다. 예를 들어, 사용자 B의 외향 연결정도 중심성 값이 높다면, 다른 사용자를 신뢰하는 성향(개방적 성향)이 강하다고 볼 수 있다. 때문에 다른 사용자들로부터 높은 신뢰를 받고 있는 사용자 A의 추천을 긍정적으로 받아들일 수 있다. 반대로 사용자 B의 외향 중심성이 특정 임계치 이하 수준

Trust CF-Search : 세 번째 접근법은 2단계에서 구축된 신뢰관계 네트워크를 직접 탐색하고, 탐색 정보를 유사도에 반영하는 접근법이다. 이 방법에서는 사용자 간 신뢰관계를 크게 2가지('직접적인 신뢰관계', '간접적인 신뢰관계')로 정의한다. 예를 들어, '사용자 A가 사용자 B를 신뢰'하고, '사용자 B가 사용자 C를 신뢰'할 때 이 경우, A와 B 그리고 B와 C는 직접적인 신뢰관계를 갖는다고 정의한다.

실험결과 첫 번째 접근법은 전통적인 협업 필터링과 비교해 성능의 개선을 가져오지 못했다. 유감스럽게도 승수의 값이 커지면 성능이 더 낮아지는 것으로 나타났다.

Trust CF-Conditional(두 번째 접근법)의 결과를 정리한 것이다. 두 번째 접근법은 이 <표 2>에서 확인할 수 있듯이, 특정 임계치를 적절하게 설정하고, 해당 임계치 이상의 외향 연결정도 중심성을 보였던 사용자에 대하여 내향 연결정도 중심성을 가중 반영하면, 예측 정확도가 어느 정도 상승한다.

Trust CF-Search(세 번째 접근법)의 실험결과이다. 결과를 살펴보면, 앞서 제시된 다른 접근법들과 비교해 월등히 우수한 예측 정확도가 산출되었음을 알 수 있다. 특히, 승수값이 커지면 커질수록 예측 정확도가 상승하는 패턴을 보인다. 이는 추천과정에서 연결정도 중심성을 고려하는 다른 접근법들에 비해 직접·간접적 신뢰 관계 탐색 정보를 고려하는 것이 훨씬 효과적이라는 사실을 시사하고 있다

Ⅳ 한계점 및 활용방안

본 연구에서 통계적으로 유의한 성과 차이를 보였던 Trust CF-Search는 사용자 간 유사도를 산출할 때 매번 네트워크를 탐색하기 때문에 많은 컴퓨팅 자원이 요구된다. 또한 본 연구가 제안하는 연구 모형은 임계치, 승수 등과 같이 모형 설계자가 임의로 설정해야 할 변수들이 포함되어 있다.

이를 우리 플랫폼에 활용하기 위해서 고려해봐야 할 것들은 다음과 같다. 첫 번째로 논문의 알고리즘을 활용하려면 네트워크의 효율성 개선이나 변수들의 최적화 방법, 카테고리 물품의 aging등을 통해 알고리즘을 강화시키는 방법이 있을 수 있고,

두 번째로는 우리가 다양한 카테고리를 이용하고 있는데 이 카테고리들이 전부 소셜 네트워크 분석에서 따르는 다른 사용자의 추천에 민감한 카테고리들이냐는 의문이 들었다. 다른 논문을 살펴본 결과(반려견 식품 구매시 인터넷을 이용하는 소비자 인식도 조사 연구 38p) 식품이나 목욕등의 생필품은 다른 사용자의 추천보다 건강을 위한 영양 밸런스나 영양 성분, 색과 냄새 등에 더 높은 구매력이 보여지는 것으로 확인이 된다. 미용물품등에는 SNA가 민감할 수 있으나 우리 플랫폼에 카테고리는 4가지이고, 각자 다른 물품들이므로 공통적으로 SNA를 적용하는 것이 맞냐는 것에 의문이 들었다.

그래서 카테고리별로 다른 알고리즘을 제안하거나, (예를 들어 사료면 영양 밸런스등을 점수화하여 해당 식에 넣는 방법을 고안) 고객의 니즈에 좀 더 맞게 새로운 알고리즘을 생각해야할 것 같다.