사용자 평판도와 태그 네트워크 확산을 활용한 개인화 추천 기법[1)]

윤상민, 홍현기, 김건우, 이동호[2)]
한양대학교 컴퓨터공학과
e-mail: {peerage, route, kgwhsy, dhlee72}@hanyang.ac.kr

A Personalized Recommendation Method exploiting User Reputation and Tag Network Diffusion

Sang-Min Yun, Hyun-Ki Hong, Gun-Woo Kim, Dong-Ho Lee
Dept of Computer Science & Engineering, Hanyang University

요 약

최근 인터넷의 급격한 발전으로 사용자의 관심사에 적합한 정보를 제공하는 추천 시스템에 대한 필요성이 증가하고 있다. 이에 따라 태그를 활용하여 추천 시스템의 성능을 향상시키려는 연구가 최근 활발하게 진행되고 있다. 하지만 태그를 활용하는 추천 시스템은 악의적인 사용자에 의해 달린 스팸 태그로 인해 부적합한 아이템을 제공한다는 문제점을 가지고 있다. 본 논문에서는 이러한 문제를 해결하기 위해 사용자 평판을 활용한 아이템 추천 기법을 제안한다. 이 기법은 먼저 사용자의 태깅 활동을 분석하여 사용자 평판을 측정한다. 다음으로 태그 네트워크를 구축한 후 사용자 평판을 고려하여 태그의 영향력을 계산하고 이를 기반으로 아이템을 추천한다.

1. 서론

인터넷이 발전함에 따라 사용자가 이용할 수 있는 정보의 양이 급격하게 증가하고 있다. 방대한 양의 정보 중 사용자가 원하는 정보를 찾는 것은 매우 어렵기 때문에 사용자의 관심사에 부합되는 정보나 아이템을 제공하는 추천 시스템의 중요성이 점점 더 높아지고 있다.

현재 추천 시스템의 성능 향상을 위해 태그를 활용하는 연구가 활발히 진행되고 있다. 태그는 태깅된 아이템에 대해 사용자들이 생각하는 일반적인 정보들을 나타내기 때문이다. 하지만 최근 많은 연구들을 통해 태그를 활용하는 추천 시스템이 스팸 태그와 악의적인 사용자(스패머)에 취약함이 밝혀졌다 [1, 6]. 예를 들어, [그림 1]은 스패머의 스팸 태그 생성과 이에 따른 추천 리스트의 왜곡을 나타낸다. 스패머가 다양한 고양이 아이템에 고양이와는 관련 없는 “신발”이라는 태그를 태깅했을 때 관심사가 신발인 사용자의 추천 리스트에 고양이 아이템이 포함되는 왜곡이 발생한다. 이처럼 스패머는 아이템과 관련 없는 태그를 아이템에 태깅함으로써 정상적인 사용자에게 부적합한 아이템을 제공하여 태그 기반 추천 시스템의 성능을 떨어뜨린다.

본 논문에서는 태그 기반 아이템 추천에서 나타나는 문제점을 해결하기 위해 사용자 평판을 활용하여 아이템을 추천하는 기법을 제안한다. 이 기법은 사용자가 태깅한 아이템의 태그 간 유사도를 기반으로 사용자 평판을 측정한다. 이후 태그 간 관계를 기반으로 태그 네트워크를 구성한 후 사용자 평판이 적용된 페이지랭크 알고리즘을 통해 측정된 태그의 영향력을 해당 태그가 태깅된 아이템에 확산하여 아이템을 추천한다.

[그림 1] 스패머의 스팸 태그 생성과 왜곡된 추천 리스트

2. 관련 연구

2.1. 아이템 추천 기법

콘텐츠 기반 필터링. 주로 사용자가 과거에 평점을 부여한 아이템과 유사한 아이템을 사용자에게 추천한다. [2]에서는 사용자에게 입력받은 논문을 분석하여 해당 논문에 포함된 용어를 사용해 잠재적인 질의를 생성한다. 다음으로 생성된 잠재 질의를 검색 엔진에 검색한 후 반환되는

1) 이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2016R1D1A1A09918271, 다중 저장소 지속성 환경에서 빅데이터 기술을 활용한 개인 맞춤형 소셜 멀티미디어 태깅 및 태그 관리 시스템)
2) 교신저자

결과를 후보 논문으로 가정한다. 이후 입력 받은 논문과 후보 논문들의 내용을 비교해 순위를 측정하여 추천 리스트를 구성하고 사용자에게 결과 값으로 반환한다.

협업 필터링. 협업 필터링은 사용자의 행동을 모으고 분석하여 해당 사용자와 비슷한 사용자를 기반으로 사용자의 관심사를 예측한다. [3]에서는 사용자가 아이템에 부여한 평점을 기반으로 사용자-아이템 평점 매트릭스를 구성하고 이를 클러스터링하여 같은 클러스터에 속한 사용자들은 비슷한 관심사를 가진다는 가정을 통해 사용자가 평가하지 않은 아이템에 대한 선호도를 예측해서 아이템을 추천한다.

네트워크 기반 아이템. [4]에서는 사용자의 아이템에 대한 선호도를 나타내는 과거 기록을 통해 사용자 유사도 네트워크를 구축한다. 이후 네트워크를 기반으로 후보 아이템에 대한 점수를 계산하고 내림차순으로 정렬하여 추천 리스트를 구성한다. [5]에서는 사용자가 서로 다른 태그로 한 아이템에 태깅했을 때 두 태그가 관련 있다고 보고 이를 기준으로 태그 네트워크를 구성한다. 이 네트워크에서 페이지랭크 또는 HITS 알고리즘을 수행하여 태그의 영향력을 측정하고 태그-아이템의 이분 그래프로 영향력을 확산하여 아이템의 점수를 매긴 후 사용자에게 아이템을 추천한다. 이 방법은 태그 간 관계를 통해 태그의 중요도를 얻어 낸다는 장점이 있지만 활동량이 많은 사용자나 스팸 태그를 많이 사용하는 사용자에 의해 결과 값에 스팸 태그가 태깅된 아이템이 주를 이루게 되는 단점이 있다.

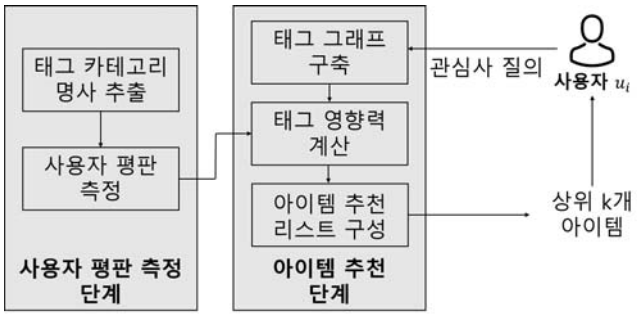
2.2. 사용자 평판 측정 기법

사용자 평판은 스팸 태그나 스패머의 검출을 위해 사용자 평판을 기반으로 스팸으로 의심되는 아이템, 태그 혹은 사용자의 영향력을 낮추는 방식으로 사용된다. [6]에서는 사용자들이 제공한 태그에 대한 피드백을 통해 태그의 스팸 여부를 판별한다. 이 후 스팸 태그를 태깅한 사용자에게 패널티 점수를, 정상적인 태그를 태깅한 사용자에게 보상 점수를 부여한다. 이 때 연속적으로 스팸 태그를 태깅한 사용자는 부가적인 패널티 점수를 받게 된다. 이를 통해 사용자 평판을 측정하고 평판에 따라 태그의 중요도를 산정한다.

3. 제안하는 시스템

[그림 2]는 사용자 평판을 활용한 아이템 추천 기법의 전체 과정을 보여준다. 기법은 크게 사용자 평판 측정 단계, 아이템 추천 단계로 구성된다.

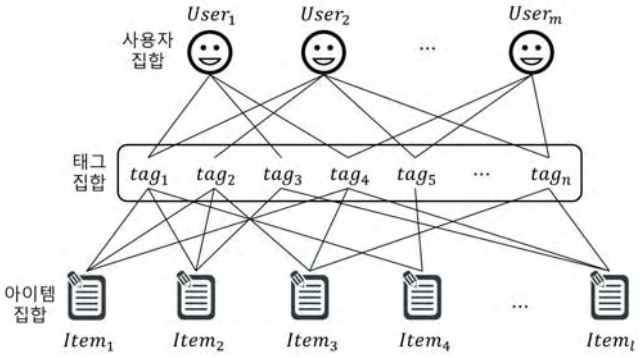
사용자 평판 측정 단계에서는 위키피디아에서 사용자가 태깅한 아이템의 각 태그로 검색된 페이지가 속한 카테고리의 명사를 추출한다. 그 후 사용자가 한 아이템 I에 태깅한 태그와 I에 태깅된 다른 태그 간 카테고리 명사의 비교를 통해 사용자가 태깅한 태그와 다른 태그들과의 평균 유사도를 계산한다. 이러한 방식으로 사용자가 태깅한 태그들의 평균 유사도를 각각 계산하고 이에 대한 평균값을 사용자의 평판으로 정의한다.



[그림 2] 제안하는 기법의 전체 과정

아이템 추천 단계에서는 방향 태그 네트워크를 구축하여 사용자 평판이 고려된 페이지랭크를 통해 네트워크의 각 태그의 영향력을 계산한다. 이후 해당 태그가 태깅된 아이템에 확산하여 아이템의 점수를 계산하고 이를 내림차순으로 정렬하여 상위 k개의 아이템을 사용자에게 제공한다.

3.1. 사용자 평판 측정 단계



[그림2] 소셜 태깅 시스템에서 사용자의 태깅 활동

[그림 2]에서 볼 수 있듯이 소셜 태깅 시스템에서는 일반적으로 사용자가 간단한 단어 형태인 태그를 아이템에 할당할 수 있다. 이러한 태그는 아이템에 대한 이해를 높일 수 있으며 아이템 공유 및 검색을 가능케 한다. 또한, 태그는 사용자의 흥미나 선호도를 파악하는 데 사용될 수 있다.

본 논문에서는 신뢰할 수 있는 사용자인지를 판별하기 위해 태그를 활용하여 사용자 평판을 측정한다. 먼저 해당 사용자가 사용한 태그들을 의미적으로 분석하기 위해 위키피디아에서 사용자가 태깅한 아이템의 태그들을 검색하여 나온 페이지의 카테고리를 추출한다. 위키피디아 카테고리는 단어뿐만 아니라 구(phrase)의 형태로도 구성될 수 있기 때문에 카테고리를 통해 단순히 비교하는 것은 두 태그의 연관성을 분석하는 데에 적합하지 않다. 따라서 문

장 속 단어들의 품사를 분석해주는 품사 태거[7]를 사용해 카테고리에 존재하는 명사만을 추출한다. 이후 각 태그의 위키피디아 카테고리에서 추출된 명사 집합에 자카드 유사도를 적용하여 사용자가 태깅한 태그와 아이템에 태깅된 다른 태그 간 유사도를 계산한다. 다음으로 이전에 계산한 유사도들의 합을 통해 사용자가 사용한 태그의 아이템에 대한 평균 유사도를 계산한다. 마지막으로 사용자가 사용한 태그들의 아이템에 대한 평균 유사도 값을 모두 더하여 사용자의 평판을 측정한다.

두 태그 간 유사도 $Sim(t_i, t_j)$는 자카드 유사도를 이용해 계산된다.

$$Sim(t_i, t_j) = \frac{|CL_{t_i} \cap CL_{t_j}|}{|CL_{t_i} \cup CL_{t_j}|} \quad (1)$$

여기서 CL_t는 태그 t로 검색된 위키피디아 페이지의 카테고리에서 추출된 명사 리스트이다.

식(1)의 유사도를 기반으로 어떤 사용자 u가 아이템 I에 태그 t로 태깅했다고 가정할 때 태그 t의 아이템 I에 대한 평균 유사도 $ASim_I(t)$는 아이템 I에서 태그 t와 t가 아닌 태그 t_j 간 유사도의 평균값으로 계산되며 식은 다음과 같다.

$$ASim_I(t) = \frac{\sum_{t_j \in Tag_I, t_j \neq t} Sim(t, t_j)}{|Tag_I|} \quad (2)$$

여기서 Tag_I는 아이템 I에 태깅된 태그 집합을 나타낸다.

사용자가 사용한 태그의 평균 유사도 값을 더해 사용자 u_i의 평판 $Rep(u_i)$를 측정할 수 있다. 식은 다음과 같다.

$$Rep(u_i) = \frac{\sum_{t_j \in Tag_{u_i}} ASim_{I_{t_j}}(t_j)}{|Tag_{u_i}|} \quad (3)$$

여기서 Tag_{u_i}는 사용자 u_i가 사용한 태그들의 집합을 나타내고 아이템 I_{t_j}는 태그 t_j가 태깅된 아이템을 의미한다.

3.2. 아이템 추천 단계

아이템 추천 단계에서는 태그 네트워크를 구축하고 태그 영향력을 계산한다. 이후 아이템 추천 리스트를 구성하여 사용자에게 아이템을 추천한다.

먼저 아이템 추천을 위해 태그를 방향을 가진 네트워크 $G=(V,E)$ 로 나타낸다. 서로 다른 두 태그가 한 아이템에 태깅되었을 때 두 태그는 관련 되어 있다고 가정한다. 여기서 $V=\{t_1, t_2, \cdots, t_n\}$ 는 태그를 나타내고, $E=\{e_{11}, e_{13}, e_{14}, e_{24}, \cdots\}$ 이다. e_{ij}는 t_i에서 t_j로의 간선을 의미하며 식은 다음과 같다.

$$e_{ij} = \frac{co-tagging(t_i, t_j)}{co-tagging_{All}(t_i)} \quad (4)$$

여기서 $co-tagging_{All}(t_i)$은 태그 t_i와 t_i를 제외한 다른 모든 태그와 동시에 태깅된 아이템의 수를 나타내고, $co-tagging(t_i, t_j)$는 태그 t_i와 t_j가 동시에 태깅된 아이템의 수이다.

이러한 방법으로 태그 네트워크를 구축한 후 네트워크에서 사용자의 평판을 고려한 변형된 페이지랭크 알고리즘을 적용하여 각 태그의 영향력을 계산한다.

태그 t_i의 k번째 반복에서의 $PR_i(k)$는 다음과 같다.

$$PR_i(k) = \left\{ d * \sum_{j \in E_{t_j}} e_{ji} PR_j(k-1) + (1-d) * PR_i(0) \right\} * \frac{\sum_{u_k \in User_{t_i}} Rep(u_k)}{\sum_{u_j \in User_{All}} Rep(u_j)} \quad (5)$$

여기서 $User_{All}$은 전체 사용자 집합을 뜻하고 $User_{t_i}$는 태그 t_i를 사용한 사용자의 집합을 의미하며 d는 감쇠비(damping factor)로 연결된 노드 간 영향력을 조절하는 변수이다. 또한 $PR_i(0)$는 태그 t_i의 최초 영향력 값을 의미하며 기존의 페이지랭크 알고리즘의 초기 값 설정과 달리 태그의 사용빈도를 고려하기 위해 [5]에서 소개된 다음의 TF-IUF 식을 이용하여 계산된다.

$$PR_i(0) = \frac{tf_{t_i}}{\max(tf_{t_j})} * \log\left(\frac{|User_{All}|}{|User_{t_i}+1|}\right) \quad (6)$$

위 식에서 tf_{t_i}는 태그 t_i가 사용된 횟수를 뜻하고 $\max(tf_{t_i})$는 제일 많이 사용된 태그의 사용횟수를 뜻한다.

위의 과정을 통해 계산된 태그의 영향력을 해당 태그가 태깅된 아이템에 확산하여 아이템의 점수를 계산한다.

이를 통해 계산되는 아이템 I_k의 점수 $Score(I_k)$는 다음과 같다.

$$Score(I_k) = \sum_{t_i \in Tag_{I_k}} \left(\frac{freq_{I_k}(t_i)}{freq_{All}(t_i)} * PR_i \right) \quad (7)$$

식(7)에서 $freq_{All}(t_i)$는 태그 t_i가 태깅된 모든 횟수를 뜻하며 $freq_{I_k}(t_i)$는 태그 t_i가 아이템 I_k에 태깅된 횟수를 의미한다.

이렇게 계산된 아이템들을 점수에 따라 내림차순으로 정렬하여 사용자에 대한 아이템 추천 리스트를 구성하고 상위 k개의 아이템을 사용자에게 추천 아이템으로 제공한다.

4. 결론

본 논문에서는 사용자에게 신뢰도 있는 아이템 추천을 위해 사용자의 평판을 측정한 후 태그 영향력 계산 시 이를 적용하고 아이템에 태그의 영향력을 아이템에 반영하여 순위를 측정함으로써 악의적인 사용자의 영향력을 줄

여 정확성을 높일 수 있는 아이템 추천 기법을 제시하였다. 향후 연구로는 기존 연구들과의 성능 비교를 통해 제안한 기법의 성능을 평가하는 실험을 진행할 예정이다.

참고문헌

[1] Wang, Yonggang, et al. "ReSpam: A novel reputation based mechanism of defending against tag spam in social computing." Service Oriented System Engineering (SOSE), 2014 IEEE 8th International Symposium on. IEEE, 2014.

[2] Nascimento, Cristiano, et al. "A source independent framework for research paper recommendation." Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries. ACM, 2011.

[3] Huang, Chuangguang, and Jian Yin. "Effective association clusters filtering to cold-start recommendations." Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on. Vol. 5. IEEE, 2010.

[4] Gan, Mingxin, and Rui Jiang. "Constructing a user similarity network to remove adverse influence of popular objects for personalized recommendation." Expert Systems with Applications 40.10 (2013): 4044-4053.

[5] Mao, Jin, et al. "Profiling users with tag networks in diffusion-based personalized recommendation." Journal of Information Science 42.5 (2016): 711-722.

[6] Wang, Yongang, et al. "Dspam: Defending against spam in tagging systems via users' reliability." Parallel and Distributed Systems (ICPADS), 2010 IEEE 16th International Conference on. IEEE, 2010.

[7] https://nlp.stanford.edu/software/tagger.html