태그 확장과 시간 정보를 이용한 아이템 추천 방법
(Item Recommendation using Tag Expansion and Temporal Information)

김 현 우 † 김 형 주 ††
(Hyunwoo Kim) (Hyoung-Joo Kim)

요 약 대부분의 추천 시스템에서는 cold-start 문제가 발생한다. Cold-start 문제란, 추천 시스템에 새롭게 등장한 사용자에게 정확한 추천을 제공하지 못하는 문제를 의미한다. 이러한 문제는 새롭게 등장한 사용자의 정보가 충분하지 않아서 추천 시스템이 정확한 사용자 프로필을 작성하지 못하기 때문에 발생한다. 본 연구에서는 cold-start 문제를 완화시키고 사용자의 선호도를 예측하는 추천 방법을 제안하였다. 소셜 태깅 시스템에서는 태그 정보를 추천 과정에 활용할 수 있기 때문에, 사용자의 태그 정보를 기반으로 태그 확장을 통해서 cold-start 문제를 해결할 수 있다. 이를 위해, 사용자의 태그 셋으로 사용자 프로파일을 구성하고 자연언어처리에 사용되는 n-gram 모델을 이용하여 태그 셋을 확장하였으며 아이템의 인기도와 시간 정보를 추가적으로 활용하였다. 실험을 통해 본 연구에서 제안하는 추천 방법이 사용자의 선호도를 잘 반영하여 cold-start 사용자에게 정확한 추천이 가능함을 확인하였다.
키워드 : 추천, 소셜 태깅, cold-start 문제

Abstract Most recommender systems have cold-start problem. The system generates poor recommendations to new users, because new users do not provide enough information to make user profile. In this paper, we propose a recommendation method which alleviates cold-start problem and predicts user's interests. In social tagging system, tagging information can be used in recommendation process. We investigate tag expansion to solve cold-start problem. A user's tag set constitutes the user profile and it is expanded by n-gram model in natural language processing. We also take an item's tag popularity and temporal information into account. The experimental results show that proposed approach recommends items precisely with tag expansion to cold-start users. The system provides better recommendations reflecting the user's interests.
Key words : Recommendation, Social tagging, Cold-start problem

1. 서 론

최근, 웹과 미디어의 발달로 인해 수많은 아이템이 생성되고 있다. 이러한 상황에서 자신이 원하는 아이템을 찾기 위해 사용자가 모든 아이템을 보고 판단하는 일은 불가능하다. 이 문제를 해결하기 위해 많은 추천 방법론들이 연구되고 있고, 정보 과다(information overload) 현상으로 인해 정확한 추천 방법의 중요성이 날로 높아지고 있다.

소셜 태깅 환경에서는 아이템에 대한 키워드인 태그를 사용자가 입력한다. 이러한 소셜 태깅 환경은 사용자가 입력한 태그를 추천 방법론의 도구 중 하나로 사용할 수 있기 때문에 추천 방법론을 연구하는 연구자들에게 새로운 기회를 제공한다. 기존의 추천 시스템은 아이템의 내용 분석과 사용자의 평점을 기반으로 추천을 진행해 왔다. 하지만 소셜 태깅 환경의 태그는 평점에 비해 해당 아이템에 대한 추가적인 정보를 제공한다. 이처럼 사용

· 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 20110030812). 본 연구는 BK-21 정보기술 사업단의 연구결과로 수행되었음

† 비 회 원 : 서울대학교 컴퓨터공학부
hwkim@idb.snu.ac.kr
(Corresponding author임)
†† 종신회원 : 서울대학교 컴퓨터공학부 교수
hjk@snu.ac.kr
논문접수 : 2012년 1월 2일
심사완료 : 2012년 4월 20일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제18권 제7호(2012.7)

자의 태그를 분석하여 사용자 프로파일을 생성하고 선호도를 측정하는데 사용할 수 있기 때문에 소셜 태깅 환경에서의 장점을 활용한 많은 연구들이 진행되고 있다.

Cold-start 문제는 추천 시스템에 새롭게 등장한 사용자에게 추천을 하고자 할 때 발생한다. 이미 많은 아이템에 평점을 입력한 활동적인 사용자는 축적된 정보가 많기 때문에 추천 시스템이 사용자의 선호도를 파악하기 위한 프로파일을 생성할 수 있는 정보가 충분하다. 하지만 새로운 사용자는 추천 시스템이 사용자 프로파일을 생성하고 사용자의 선호도를 측정하는데 사용할 정보가 존재하지 않기 때문에 정확한 추천이 어렵다. 이러한 cold-start 사용자는 추천 시스템의 전체적인 성능을 저하시키는 요인 중 하나이다.

추천 시스템에서의 cold-start 문제를 해결하기 위해 본 연구에서는 태그 확장과 시간 정보를 활용한 방법을 제안하였다. 소셜 태깅 시스템의 cold-start 사용자가 입력한 태그의 수는 활동적인 사용자가 입력한 태그의 수에 비해 상대적으로 적다. 하지만 cold-start 사용자의 태그 셋을 확장하면 추천 시스템이 활용할 수 있는 정보를 풍부하게 제공할 수 있다. 본 연구에서는 사용자의 초기 태그 셋을 확장하기 위해서 자연언어처리 분야에서 사용되는 bigram 방법을 이용하였다. 이러한 방법을 통해 확장된 태그들은 기존의 사용자 태그 셋에 추가되어 사용자와 아이템 사이의 관련성을 정의하는데 도움을 준다. 사용자 태그 셋에 포함되어 있는 각 태그는 사용자가 태그를 얼마나 사용했는지 빈도수에 의해서 점수가 매겨지는데, 사용자가 태그를 더 많이 사용할수록 태그의 점수가 높아진다. 주어진 사용자에 대해서, 한 아이템의 점수는 그 아이템에 입력되어 있는 태그의 점수의 평균이 되고, 이 평균 점수를 계산하는 과정에서 아이템 자체의 인기도를 함께 고려한다. 모든 아이템의 점수가 계산되면 점수가 높은 순서대로 사용자에게 아이템을 추천해 준다. 같은 태그 셋으로 태그된 두 아이템의 점수는 같기 때문에 이러한 두 아이템을 구분하기 위한 기준이 필요하게 된다. 이를 위해, 본 연구에서는 아이템 자체의 시간 정보를 이용하였다.

본 논문은 다음과 같이 구성되어 있다. 2장에서 관련 연구를 설명하고, 본 연구에서 제안한 태그 확장 방법과 추천 알고리즘에 대해서 3장에서 설명한다. 4장에서 실험결과를 분석하고, 5장에서 결론을 맺고 논문을 마무리 한다.

2. 관련연구

최근, 소셜 태깅 환경에서의 추천 방법론에 관한 많은 연구들이 진행되고 있다. Guan[1]은 optimal semantic space를 학습하여 문서를 추천하는 방법을 제안하였다. Optimal semantic space는 그래프 구조의 연결성을 보존하는 공간이다. 사용자, 태그, 문서를 하나의 공간에 표현하여 서로 연관성이 높은 것들을 그래프 공간 상에서 서로 가까운 곳에 위치시키는 방법이다. 주어진 사용자에 대해서 학습된 공간상에서 사용자와 가장 가까운 문서를 추천해주는 방법을 사용하였다. Guy[2]는 소셜 네트워크와 태그 정보를 기반으로 한 추천 방법을 제안하였다. 이 연구에서는 사용자의 태그 정보뿐만 아니라 사용자의 소셜 네트워크를 분석하여 추천에 활용하였다. 사용자와 태그 사이의 연관성을 측정하기 위해서 간접적 태그와 다른 사용자가 해당 사용자에게 입력한 태그 정보를 활용하였다. 또한, Konstas[3]도 소셜 네트워크의 요소들을 추천 시스템에 활용하였다. 소셜 네트워크를 효율적으로 표현하기 위해서 RWR(Random Walk with Restarts) 모델을 사용하였다. RWR 모델은 데이터베이스로부터 직접 사용자의 선호도를 예측할 수 있다는 장점이 있다. Kim[4]은 소셜 태깅 시스템의 정보만 이용하는 것이 아니라, 사용자의 태그 정보와 평점 정보를 함께 이용하는 방법을 제안하였다. 이 연구에서는 태그와 평점을 함께 연결하여 사용함으로써 추천 결과의 다양성을 보존하고 cold-start 문제를 해결하였다. 이를 위해, 사용자와 선호하는 아이템이 비슷한 사용자와 선호하지 않는 아이템이 비슷한 사용자 모두를 고려한 협업 필터링(collaborative filtering) 방법을 사용하였다.

Kim[5]은 cold-start 문제를 해결하기 위해서 태그 확장을 사용하는 방법을 제안하였다. 이 연구에서는 협업 필터링 방법을 사용하여 사용자와 가장 가까운 이웃을 선정한 뒤, 그 이웃들의 태그로 확장하는 방법을 제안하였다. 사용자-아이템, 사용자-태그, 아이템-태그 사이의 관계를 기반으로 입력한 태그가 비슷하거나, 입력한 아이템이 비슷한 사용자를 선정하여 태그를 확장하고, 이를 아이템 추천에 활용하였다. Liang[6]은 태그의 품질에 관한 연구를 진행하였다. 이 연구에서는 태그를 적절한 아이템을 찾는 매개수단으로 보고, 아이템의 내용을 기반으로 태그를 확장하였다. 사용자-태그, 아이템-태그 사이의 관계뿐만 아니라 각 사용자-태그와 아이템의 관계, 즉(사용자-태그)-아이템의 관계를 활용하였다. Heymann[7]은 아이템의 직접적인 추천이 아닌, 태그를 예측하는 과정에서 태그 확장을 이용한 방법을 제안하였다. 아이템의 문자 정보를 분석하여 출현 가능성이 높은 태그를 예측하였다.

3. 태그 확장을 이용한 아이템 추천

본 연구에서는 cold-start 문제를 해결하기 위해서 사용자의 초기 태그 셋을 확장하는 방법을 제안하였다. 사용자의 초기 태그 셋은 사용자가 입력한 태그 집합으로

부터 도출된다. 사용자 u가 태그 t_1, …, t_n을 아이템에 입력했다면, 사용자 u의 초기 태그 셋 $TS_i(u)$는 다음과 같이 정의된다.

$$TS_i(u) = \{t_1, ..., t_n\}$$

3.1 Bigram을 이용한 태그 확장

사람의 사고방식은 연상작용을 이용한다. 하나의 생각에서 전혀 다른 생각으로 뛰어넘는 것이 아니라, 이전 생각과 연관된 다른 생각으로 사고를 이어나간다. 사용자가 태그를 입력하는 과정은 아이템에 대한 사용자의 사고방식의 표현으로 볼 수 있다. 그렇기 때문에 본 연구에서는 태그를 입력하는 과정에서도 이와 같은 연상작용이 일어날 것이라고 가정하였다. 바로 옆에 입력된 태그가 그렇지 않은 태그보다 관련성이 높을 것이다. 예를 들어 새로운 아이폰의 출시에 관한 뉴스 기사에 스티브잡스, 애플, 아이폰, 출시와 같은 순서로 태그가 입력 되었다면, 스티브잡스라는 태그는 출시보다는 애플과 관련성이 더 높을 것이다.

본 연구에서는 이러한 점에 착안하여 자연언어처리 연구분야에서 사용되는 bigram 모델을 사용하여 초기 태그 셋을 확장하였다. 소셜 태깅 환경에서 두 태그의 관련성을 측정하기 위해서 많이 사용되는 방법은 두 태그의 동시출현 빈도를 측정하는 것이다. 두 태그가 하나의 아이템에 함께 입력된 경우가 많다는 것은 그만큼 두 태그가 관련이 높다는 것이다. 동시출현 빈도를 측정하는 경우, 사용자가 한 아이템에 태그 t_1, t_2, t_3, t_4를 순서대로 입력했다면 (t_1, t_2), (t_1, t_3), (t_1, t_4), (t_2, t_3), (t_2, t_4), (t_3, t_4)가 함께 등장한 것으로 본다. 같은 경우에 대해서 bigram 모델은 (t_1, t_2), (t_2, t_3), (t_3, t_4)를 함께 등장하는 것으로 보고 두 태그 사이의 관련성을 측정하는데 사용한다. 위의 예제에서 bigram 모델에서는 태그 t_1과 t_2의 관련성이 t_1과 t_4의 관련성보다 더 높다고 보는 것이다. 사용자가 한 아이템에 많은 수의 태그를 입력하는 경우, 가장 처음으로 입력한 태그와 가장 마지막으로 입력한 태그의 관련성은 더욱 감소할 것이다. 동시출현 빈도를 이용하면 관련성의 감소를 반영하지 못하지만, bigram 모델은 이러한 관련성의 감소를 반영하여 측정할 수 있다.

Bigram 모델에서 두 태그 사이의 관련성은 전체 데이터에서 bigram 쌍의 빈도수로 정의된다. (t_1, t_2)의 빈도수가 10,000번이고 (t_1, t_3)의 빈도수가 10번이라면, 태그 t_1은 t_2와는 관련성이 높지만 t_3와는 관련성이 낮다고 볼 수 있다. 이러한 경우, 본 연구에서 제안하는 태그 확장 방법에서는 t_1을 사용한 사용자의 태그 셋에 t_2를 추가하게 된다. 초기 태그 셋에 어떤 태그를 추가하고 어떤 태그를 추가하지 않을지 정하기 위해서 조건부 확률을 도입하였다.

$$P(t_2|t_1) = \frac{P(t_1|t_2)}{P(t_1)}$$

$P(t_1)$은 t_1이 등장할 확률이고, 이 값은 전체 데이터에서 t_1이 등장한 빈도수와 같다. $P(t_1, t_2)$는 태그 쌍 (t_1, t_2)의 확률이고, 이 값은 전체 데이터에서 t_1과 t_2가 함께 등장한 bigram 쌍의 빈도수와 같다. 주어진 태그 t_1에 대한 확장 태그 셋 $ET(t_1)$은 다음과 같이 정의된다.

$$ET(t_1) = \{t_k \mid P(t_k|t_1) > threshold\}$$

태그 t_1과 함께 쓰인 태그 중 bigram 쌍의 조건부 확률이 임계값(threshold)을 넘는 t_k를 사용자의 태그 셋에 추가한다. 임계값을 얼마로 설정하느냐에 따라 확장되는 태그의 수가 정해지기 때문에 임계값을 너무 높게 설정하면 추가되는 태그가 거의 없을 것이고, 임계값을 너무 낮게 설정하면 거의 모든 태그가 추가될 것이다. 확장 태그가 추천 성능에 영향을 끼치므로 실험을 통해 적절한 임계값을 설정하는 것이 중요하다.

사용자 u의 초기 태그 셋에 포함되어 있는 태그를 기반으로 확장된 태그 셋 $TS_e(u)$는 다음과 같이 정의된다.

$$TS_e(u) = \{t_k \mid P(t_k|t_1) > threshold, t_1 \in TS_i(u)\}$$

$TS_i(u)$는 사용자의 초기 태그 셋이고 $TS_e(u)$는 $TS_i(u)$에 포함된 태그 중에서 조건부 확률이 임계값을 넘는 태그 t_l들의 집합이다. 이렇게 생성된 확장 태그 셋 $TS_e(u)$는 초기 태그 셋 $TS_i(u)$와 함께 사용자 프로필을 생성하여 아이템을 추천하는데 이용된다.

3.2 태그 스코어

사용자의 태그 셋에 포함되어 있는 각 태그의 점수는 사용자가 그 태그를 얼마나 많이 사용하였는지를 기반으로 한 확률 값으로 정해진다. 사용자가 태그 t_1을 한 번밖에 사용하지 않았다면 t_1의 점수는 낮을 것이고 사용자가 t_2를 많이 사용했다면 t_2의 점수는 높아질 것이다. 본 연구에서는 [8]에서와 같은 방식으로 태그에 점수를 부여하였다. 주어진 사용자 u에 대한 태그 t_1의 점수 w_{u,t_1}은 다음과 같이 정의된다.

$$w_{u,t_1} = \frac{freq(u, t_1)}{\sum_{t=1}^{k} freq(u, t_1)}$$

k는 사용자 u가 사용한 서로 다른 태그의 수를 의미하고, $freq(u, t_1)$은 사용자 u가 태그 t_1을 사용한 횟수를 의미한다. 주어진 사용자 u에 대한 사용자의 태그 셋과 점수의 집합인 $STS(u)$는 다음과 같이 정의된다.

$$STS(u) = \{<t_m, w_{u,t_m}> \mid t_m \in (TS_i(u) \cup TS_e(u))\}$$

초기 태그 셋 혹은 확장 태그 셋에 포함된 모든 태그 tm에 대해서 점수를 부여하여 태그와 점수를 함께 포함한 집합이다. 초기 태그 셋에 포함된 태그는 사용 빈도를 기준으로 점수를 부여할 수 있지만 확장된 태그의 경우, 사용자가 사용하지 않은 태그이기 때문에 사용 빈

도를 기준으로 점수를 부여할 수 없다. 그래서 다음과 같이 기존의 태그가 가진 점수의 일정 비율(ratio)을 확장된 태그에 부여한다.

$$w_{u,t_m} = w_{u,t_l} \times ratio,$$

where $t_m \in TS_e(u)$ and $P(t_m \mid t_l) > threshold$

태그 t_m이 태그 t_l에 의해서 확장되었다면 t_m의 점수는 t_l의 점수에 일정 비율(ratio)을 곱한 값이 된다. 비율이 1이면 초기 태그와 같은 점수를 부여받게 된다.

3.3 랭킹 알고리즘

본 연구에서 제안하는 추천 알고리즘은 사용자 u와 아이템 p의 관련성을 계산하여 관련성이 높은 아이템을 사용자에게 추천한다. u와 p의 관련성은 u의 STS(u)와 p에 입력된 태그 사이의 관련성으로 정해진다. 사용자 u와 아이템 p 사이의 관련성 R(u, p)는 다음과 같이 정의된다.

$$R(u,p) = \sum_{t_p \in MS(p)} \frac{w_{u,t_p}}{N}$$

MS(p)는 아이템 p에 입력된 태그의 멀티 셋이다. MS(p)에는 같은 태그가 여러 번 등장할 수 있다. N은 p에 입력된 태그 t_p 중에서 태그 점수 값 w_{u,t_p}가 0이 아닌 태그의 수이다.

소셜 태깅 모델 중, 협력 태깅 모델은 하나의 아이템에 서로 다른 사용자가 태그를 입력할 수 있는 모델을 의미한다. 이러한 모델은 사용자들의 태그 정보가 그 아이템에 대한 폭소노미(folksonomy)를 생성할 수 있게 한다. 다른 태그에 비해서 많이 사용된 태그는 해당 아이템을 대표하는 키워드라고 할 수 있다. 본 연구에서 제안한 관련성 척도의 경우, 이러한 협력 태깅 모델의 특성을 잘 반영할 수 있다.

사용자 u의 태그를 바탕으로 두 아이템에 같은 태그가 입력되어 있다면, u에 대한 두 아이템의 점수는 같을 것이다. 혹은, 두 아이템에 서로 다른 태그가 입력되어 있지만 점수가 같은 경우도 발생할 수 있다. 본 연구에서는 이처럼 같은 점수를 가지는 아이템을 구분하기 위해서 시간 정보를 이용하였다. Dia[9]가 아이템의 최신성의 중요함에 대한 언급한 것과 같이, 본 연구에서는 이러한 점에 착안하여 같은 점수를 가지는 아이템에 대해서 최근에 생성된 아이템에 대해서 더 높은 점수를 부여하는 방식을 적용하였다. 이러한 방법은 cold-start 사용자나 아이템에 특히 유용할 수 있다. 왜냐하면, cold-start 사용자의 경우, 사용자가 입력한 태그의 수가 적기 때문에 추천된 아이템이 같은 점수를 갖게 될 가능성이 높다. 그리고 cold-start 아이템의 경우, 아이템에 입력된 태그의 수가 적기 때문에 이러한 아이템이 사용자에게 추천되면 같은 점수를 갖는 아이템이 많아질 가능성이 높다. 이처럼 같은 점수를 갖는 아이템을 구분하는 기준이 필요하게 되는데, 시간 정보가 그 역할을 하게 된다.

알고리즘 1에서 본 연구에서 제안하는 전체 알고리즘에 대해서 간략히 서술하였다.

알고리즘 1. 태그 확장과 시간 정보를 이용한 아이템 추천

Input: user u
Output: L: a list of ranked items
IT ← $TS_i(u)$
foreach tag $t_l \in$ IT **do**
  Find bigram pair of (t_k, t_l)
  if $P(t_k \mid t_l) > threshold$ **then**
    $TS_e(u) \leftarrow t_k$
  end
end
ETS ← $TS_i(u) \cup TS_e(u)$
foreach tag $t_m \in$ ETS **do**
  Calculate the score of tag t_m
  $STS(u) \leftarrow < t_m, w_{u,t_m} >$
end
P ← items
foreach $p_n \in$ P **do**
  Calculate the score of p_n with $STS(u)$
end
L ← Sorted list of $p_n \in$ P with the score and temporal information

사용자가 입력으로 들어오면, 그 사용자가 사용한 태그를 기준으로 임계값이 넘는 태그들을 태그 셋에 추가한다. 하지만 추가된 태그에 대해서 다시 태그 확장을 수행하지는 않는다. 그 다음으로, 확장된 태그 셋에 있는 태그에 대해서 점수를 계산하고, 그 점수를 기반으로 각 아이템에 점수를 부여한다. 마지막으로, 부여된 점수와 시간정보를 기반으로 높은 순서대로 사용자에게 아이템을 추천한다.

4. 성능평가

본 연구에서는 CiteULike[10]에서 제공하는 데이터 셋을 사용하였다. CiteULike는 논문과 같은 학술문서를 저장하고 태그를 입력할 수 있는 소셜 태깅 서비스를 제공한다. CiteULike 데이터셋은 사용자, 아이템, 태그, 시간으로 구성되어 있다. 총 3,051,409개의 아이템, 633,483개의 서로 다른 태그, 89,461명의 사용자, 14,028,761개의 태깅 정보로 이루어져 있다.

4.1 데이터 분포

우선, CiteULike 데이터 셋의 사용자가 사용한 서로 다른 태그의 수에 따른 사용자의 분포와 입력한 아이템의 수에 따른 사용자의 분포를 분석하였다. 그림 1, 2를 통해 두 분포가 멱법칙(power law)을 따른다는 것을 알 수 있다. 그림 1에서 34,111명의 사용자가 하나의 태그만을 사용하였고, 8,987명의 사용자가 두 종류의 태그

표 1 임계값에 따른 그룹별 평균 확장 태그의 수

그룹 1	임계값	0.01	0.02	0.03	0.04	0.05	0.07	0.1	0.2	0.3
	평균확장태그수	29.7	13.1	8.1	5.4	3.7	2.5	1.7	0.8	0.2
그룹 2	임계값	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	평균확장태그수	15.7	4.2	1.9	0.7	0.2	0.1	0.07	0.04	0.03
그룹 3	임계값	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	평균확장태그수	135	36.6	15.4	5.0	1.1	0.76	0.43	0.29	0.24

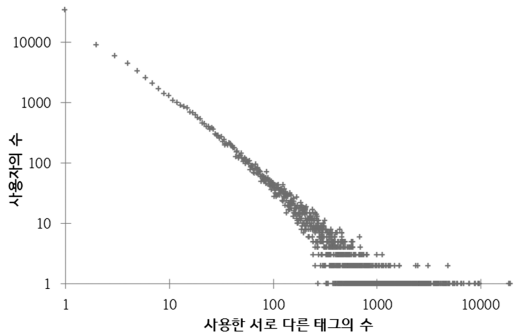


그림 1 사용자가 사용한 서로 다른 태그의 수에 대한 사용자 분포

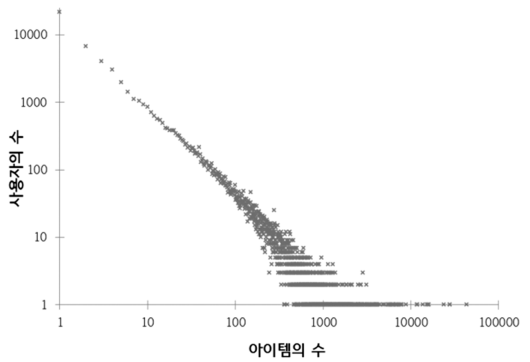


그림 2 입력한 아이템 수에 대한 사용자 분포

를 이용해서 아이템을 입력하였다. 전체 사용자의 73%가 10개 이하의 서로 다른 태그를 사용하였고, 100개 이상의 태그를 사용한 사용자는 5%에 불과하였다. 이와 마찬가지로, 그림 2에서 21,688명의 사용자가 하나의 아이템을 입력하였고, 전체 사용자의 67%가 10개 이하의 아이템을 입력하였다. 100개 이상의 아이템을 입력한 사용자는 전체 사용자의 9%에 그쳤다. 이를 통해, 대부분의 사용자들이 cold-start 사용자이고 아이템 역시 cold-start 아이템이라는 것을 알 수 있다.

4.2 사용자 분류

사용자의 태깅 패턴에 따른 분석을 위해, 사용한 태그의 수에 따라 사용자를 분류하였다. 사용자를 세 그룹으로 나누어서 사용한 태그의 수가 10개 이하인 사용자는 그룹 1, 11개 이상 100개 이하인 사용자는 그룹 2, 101개 이상의 태그를 사용한 사용자는 그룹 3으로 분류하였다. 각 그룹의 사용자에 대해서 임계값에 따라서 확장되는 태그의 수가 어떻게 변화하는지 측정하고, 그 결과를 표 1에 정리하였다. 각 그룹의 결과는 서로 다른 확장 태그의 수를 나타냈다. 임계값이 0.1인 경우 그룹 1은 평균 1.7개의 태그만 확장되었다. 그룹 2와 그룹 3은 각각 15.7개와 135개의 평균 태그가 확장되었다. 그룹 3의 사용자가 그룹 1의 사용자보다 훨씬 더 많은 태그를 사용했기 때문에 이와 같은 결과가 나타났다. 본 연구에서는 cold-start 사용자인 그룹 1의 사용자에 집중하여 실험을 진행하였다.

4.3 실험결과

우선, 사용자가 관심 있어 할만한 아이템을 추천하기 위해 어떤 임계값과 비율을 적용해야 가장 좋은 결과를 얻어낼 수 있는지를 측정하는 실험을 진행하였다. 이를 위해 사용자가 입력한 아이템 중 80%를 이용해서 사용자 프로파일을 작성하고, 나머지 20%의 아이템을 정답으로 간주하였다. 80%의 학습 데이터를 기반으로 $TS_i(u)$를 생성하고, 추천 시스템으로부터 추천된 결과는 20%의 정답 데이터와 비교하여 정확도(precision)와 재현율(recall)을 측정하였다. 서로 다른 임계값에 따른 정확도와 재현율 측정 결과를 그림 3, 4에 그래프로 나타내었다.

이 실험에서는 20개의 아이템을 사용자에게 추천하여

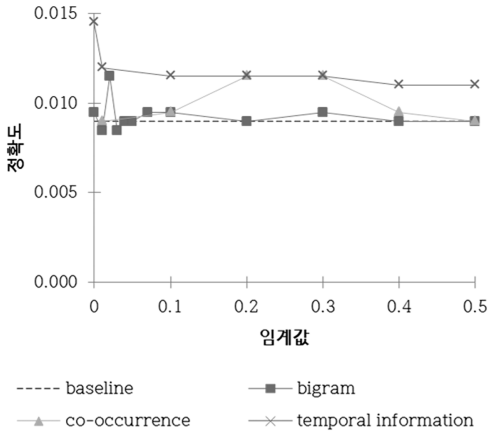


그림 3 임계값에 따른 정확도 변화

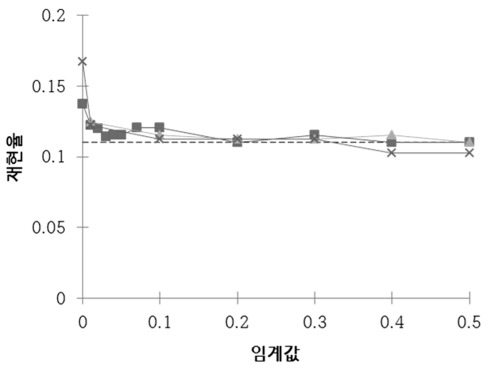


그림 4 임계값에 따른 재현율 변화

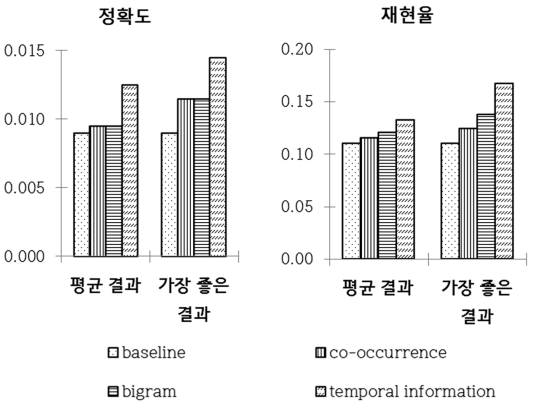


그림 5 태그 확장 방법론에 따른 정확도와 재현율의 평균 결과와 각 방법론에 따른 가장 좋은 결과

정확도와 재현율을 측정하였다. 위 그래프의 정확도와 재현율 값은 무작위로 선택된 100명의 사용자의 결과를 평균 낸 값이다. 그래프에서 점선으로 나타난 baseline은 태그 확장 방법을 사용하지 않은 결과이다. 이 baseline을 기준으로 방법론을 평가하게 된다. Co-occurrence는 bigram이 아닌 동시출현 빈도를 이용한 태그 확장 방법의 결과이다. Bigram은 태그 확장을 하는데 bigram 모델을 사용한 결과이고, temporal information은 bigram 모델과 함께 시간 정보를 사용하여 점수가 같은 아이템을 시간순으로 순위를 결정해서 추천한 결과이다. 임계값이 0.4보다 큰 경우, bigram 모델의 정확도와 재현율이 baseline과 같은 것을 볼 수 있다. 이와 같은 결과가 나타나는 이유는, 임계값이 너무 높아서 태그가 전혀 확장되지 않기 때문에 태그 확장을 사용하지 않는 경우와 같은 결과를 나타내었다. 그림 3에서 temporal information이 다른 방법론에 비해서 높은 정확도를 보임을 알 수 있다. 하지만 재현율의 경우 임계값이 0.4보다 큰 경우 baseline 방법보다 낮은 값을 보이고 있다. 이는 임계값이 0.4보다 큰 경우 태그 확장이 전혀 일어나지 않기 때문에 시간 정보만을 이용해서 순위를 새롭게 정하는 것은 추천 결과의 재현율을 떨어뜨린다는 것을 의미한다. 이와 같은 실험 결과로 미루어 볼 때, 적절한 임계값의 설정이 중요하다는 것을 알 수 있다.

다음으로 확장된 태그에 점수를 얼마나 반영할 것인지 결정하는 수치인 비율(ratio)에 대한 실험을 진행하였다. 비율이 0.6보다 낮은 경우, 각 비율에 따른 전체 시스템의 추천 성능은 차이를 보이지 않았다. 비율이 0.7보다 큰 경우, 추천 결과의 정확도와 재현율은 적절하지 않은 비율 값으로 인해 약간 감소하는 경향을 보였다. 실험 결과에서 비율 값이 1인 경우 가장 낮은 정확도와 재현율을 나타내었다. 이러한 결과가 나타나는 이유는 비율 값이 1인 경우, 확장된 태그의 점수는 그 태그가 확장될 수 있게 된 초기 태그의 점수와 같은 점수를 부여받게 되는데, 확장된 태그가 사용자 프로파일을 작성하는데 너무 많은 기여를 하기 때문으로 분석된다.

그림 5는 네 가지 태그 확장 방법론에 따른 정확도와 재현율의 결과를 나타낸다. 그림 5의 왼쪽 그래프는 서로 다른 임계값과 비율 값에 따른 정확도를 평균 낸 결과와 가장 좋은 결과를 나타낸 값이고, 오른쪽 그래프는 서로 다른 임계값과 비율 값에 따른 재현율을 평균 낸 결과와 가장 좋은 결과를 나타낸 값이다. Baseline 방법의 결과는 태그 확장을 사용하지 않아서 임계값과 비율에 영향을 받지 않기 때문에 같은 결과를 나타낸다. Co-occurrence와 bigram의 경우, 정확도에서는 큰 차이를 보이지 않지만 재현율에서는 bigram 모델이 더 나은 결과를 나타내고 있다.

그림 5에서 볼 수 있듯이 bigram 모델과 시간 정보를 함께 사용한 방법이 가장 좋은 추천 성능을 나타냈다. 이를 통해 시간 정보의 사용이 추천 결과의 전체적인 성능 향상을 가지고 온다는 것을 알 수 있다. 이러한 성능 향상이 일어나는 이유는 3장에서 언급한 바와 같이 cold-start 사용자의 태깅 정보가 많지 않기 때문에 같은 점수를 가지는 아이템이 많이 등장하게 되는데, 이러한 아이템들을 잘 구분하였기 때문으로 분석된다.

지금까지의 실험 결과를 바탕으로 본 연구에서 제안한 태그 확장과 시간정보를 이용한 아이템 추천 방법이 실제 데이터 상에서 좋은 성능을 보임을 알 수 있다.

5. 결론 및 향후연구

본 연구에서는 소셜 태깅 시스템에서의 추천 방법론을 제안하였다. 본 연구에서 제안한 추천 방법론은 사용자의 태그 셋을 확장하여 cold-start 문제를 해결하고자 하였으며, 제안된 방법론은 bigram 기반의 태그 관련도 측정 기법으로 효과적이고 정확하게 태그 셋을 확장하

여 좋은 결과를 나타내었다. 아이템의 순위를 정하기 위해서 사용자의 태그 선호도와 각 아이템에서의 태그 인기도를 반영하고, 같은 태그 셋을 가지고 있는 아이템을 구분하기 위해 시간 정보를 활용하였다. 성능 평가를 통해 제안된 방법론이 실제 데이터 상에서 잘 동작함을 알 수 있었다.

본 연구에서는 사용자가 사용한 태그를 하나씩 살펴보고 그 태그와 관련된 태그들을 추가하여 태그 셋을 확장하였다. 추가적으로, 사용자가 사용한 모든 태그를 한번에 고려해서 태그를 확장하는 방법을 사용한다면 사용자의 선호도를 더욱 정확하게 판단할 수 있을 것이다. 또한 시간 정보를 이용하여 태그의 점수를 측정하는데 사용할 수 있을 것이다. 사용자가 최근에 사용한 태그는 더 높은 점수를 부여하는 방식을 사용하면, 사용자의 최근 관심사가 반영된 사용자 프로파일 작성이 가능할 것이다. 또한, 확장된 태그의 점수를 부여하는 방식을 다르게 적용할 수도 있을 것이다. 기존의 태그에 일정 비율을 곱하는 것이 아닌, 그 태그가 확장된 bigram 모델의 확률값을 곱하는 방식을 취하게 되면 적절한 임계값이나 비율 값을 찾지 않고 태그의 점수를 정하는 방식으로 본 연구를 확장해 나갈 수 있다.

참 고 문 헌

[1] Z. Guan, C. Wang, J. Bu, C. Chen, K. Yang, D. Cai, X. He, "Document recommendation in social tagging services," *Proc. of the 19th ACM International Conference on World Wide Web*, pp.391-400, 2010.

[2] I. Guy, N. Zwerdling, I. Ronen, D. Carmel, E. Uziel, "Social media recommendation based on people and tags," *Proc. of the 33rd International ACM SIGIR Conference*, pp.194-201, 2010.

[3] I. Konstas, V. Stathopoulos, J.M. Jose, "On social network and collaborative recommendation," *Proc. of the 32nd International ACM SIGIR Conference*, pp.195-202, 2009.

[4] H.N. Kim, A. Alkhaldi, A.E. Saddik, G.S. Jo, "Collaborative user modeling with user-generated tags for social recommender system," *Expert Systems with Application 38*, pp.8488-8496, 2011.

[5] H.N. Kim, A.T. Ji, I. Ha, G.S. Jo, "Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation," *Electronic Commerce Research and Applications 9*, pp.73-83, 2010.

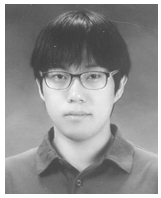
[6] H. Liang, Y. Xu, Y. Li, R. Nayak, X. Tao, "Connecting users and items with weighted tags for personalized item recommendations," *Proc. of the 21st ACM Conference on Hypertext and Hypermedia*, pp.51-60, 2010.

[7] P. Heymann, D. Ramage, H. Garcia-Molina, "Social tag prediction," *Proc. of the 31rd International ACM SIGIR Conference*, pp.531-538, 2008.

[8] N. Zheng, Q. Li, "A recommender system based on tag and time information for social tagging systems," *Expert Systems with Applications 38*, pp.4575-4587, 2011.

[9] N. Dia, D.B. Davison, "Freshness matters: in flowers, food, and web authority," *Proc. of the 33rd International ACM SIGIR Conference*, pp.114-121, 2010.

[10] CiteULike, http://www.citeulike.org



김 현 우

2007년 KAIST 전산학과(학사). 2007년~현재 서울대학교 컴퓨터공학부 석박통합과정 재학중. 관심분야는 추천, 태깅, 데이터베이스

김 형 주

정보과학회논문지 : 컴퓨팅의 실제 및 레터 제 18 권 제 3 호 참조