온라인 커뮤니티에서 전문성과 대중성에 기반한 사용자 생성 컨텐츠의 랭킹

이지훈, 신효섭
건국대학교 신기술융합학과
e-mail : jeehoon.lee01@gmail.com, hsshin@konkuk.ac.kr

Ranking User-Generated Contents Based on Expertise and Popularity in Online Communities

Jeehoon Lee, Hyoseop Shin
Dept. of Advanced Technology Fusion, Konkuk University

요 약

오늘날 웹 상에는 수 많은 온라인 커뮤니티들이 존재하고, 그 안에서 유저들이 올린 게시글(이하 포스트)를 효과적으로 검색하는 것은 중요한 이슈가 되고 있다. 만약 검색하는 유저들이 각기 다른 성향을 갖고 있다면 그에 맞는 검색 결과를 제공하는 것이 효과적인 검색의 한 예라 할 수 있겠다. 이 논문에서는 이러한 유저 성향 기반의 효과적인 검색을 위하여 유저의 "전문성"과 "대중성"을 정의하고 그에 기반한 포스트 랭킹을 한다. 또한 서로 다른 유저의 성향은 매우 다른 검색 결과를 나타낸다는 우리의 주장을 실험결과로 뒷받침 한다.

1. 서론

Flikcr 와 YouTube 같은 대규모 온라인 커뮤니티에는 유저가 작성한 수백만 개의 포스트(post) 들이 존재한다. 그 안에서 효과적인 포스트 검색을 하는 것은 점차 중요한 이슈가 되고 있다. 특히 검색 유저가 각기 다른 성향을 갖고 있다면, 그들은 서로 다른 검색 결과를 기대 할 것이다. 예를 들어 어떤 유저가 인기 있고 대중적인 포스트를 좋아한다면, 검색 시 대중적인 포스트가 상위 랭크 되고, 반대로 유저가 전문성이 요구되는 포스트를 좋아한다면, 검색 시 전문성이 높은 포스트가 상위 랭크 되는 것이 바람직할 것이다. 이 논문에서는 이러한 유저 성향에 기반한 포스트 랭킹의 개념을 정의하고, 실제로 이런 접근이 얼마나 타당한지 실험 결과로 증명한다.

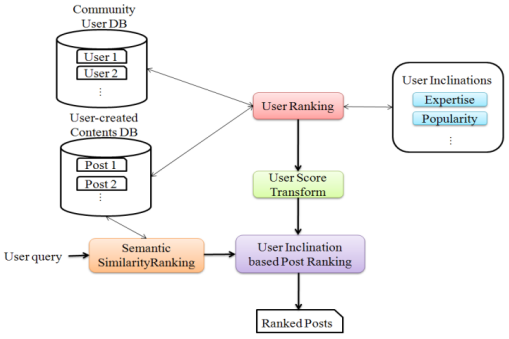
2. 유저 성향 기반의 포스트 랭킹

한 포스트의 점수는 의미적 유사성과 유저 성향에 대한 점수의 합으로 이루어 진다. 이에 대한 공식은 다음과 같다.

$$\text{PostScore}(query, post) = \alpha * \text{SemanticSimilarity}(query, post) + (1-\alpha) * \text{UserInclinationPostScore}(post), \quad (0 \le \alpha \le 1)$$

의미적 유사성의 점수는 TF*IDF 기반의 문서 가중치로 이루어 진다. 그리고 유저 성향에 대한 포스트 점수는 해당 포스트에 기여한 유저의 점수를 합함으로써 얻게 된다. 여기서의 기여란 포스트에 대하여 코멘트를 하거나 좋아하는 사진에 추가, 조회, 추천 등 해당 포스트에 관한 일련의 활동을 의미한다. 여기서 유저의 점수는 유저의 성향에 기반하는데, 이 논문에서는 전문성(expertise)과 대중성(popularity)이라는 두 가지 측면에서 접근하게 된다.

그림 1 은 위와 같은 과정에 대한 흐름도이다. 먼저 오프라인 과정에서 유저의 특정한 성향에 기반한 유저 랭킹 점수[1]가 계산되고, UserScore Transform 과정을 거치게 된다. 이러한 전처리 과정 후에 유저의 쿼리를 입력으로 받아 의미적 유사성에 대한 점수와 유저 성향에 대한 점수를 합하여 포스트의 점수를 매기게 되고, 이 점수를 바탕으로 랭킹을 하게 된다.



(그림 1) 유저 성향에 기반한 포스트 랭킹 흐름도

3. 유저의 전문성과 대중성

커뮤니티에서 사진이나 동영상은 작성한 유저의 성향을 반영한다. 이러한 유저의 성향은 커뮤니티에서 카테고리를 형성하고 이에 기반한 랭킹 방법이 여러 커뮤니티에서 제공 되고 있다. 하지만 카테고리는 단순히 주제에 대한 분류만을 제공 할 뿐이어서 사진이나 동영상에 대한 전문성이나 대중성을 판단 할 수는 없다. 이에 논문에서는 카테고리 분류가 아닌 전문성과 대중성을 정의하고, 이에 기반한 포스트 랭킹을 하려고 한다.

만약 포스트가 전문적인 지식이나 스킬(skill)을 요구하는 것이라고 할 때, 피드백을 주는 유저들은 해당 포스트에 대해 비전문적인 유저 보다는 전문성을 갖고 있는 유저들이 더 많을 것이다. 이와 반대로 포스트가 전문성이 없는 것이라면, 피드백을 주는 유저들은 전문성이 없는 유저들이 더 많을 것이다.

이에 유저의 전문성에 대하여 Expertise Rank(ER)을 정의한다. 어떠한 유저 u 에 대한 ER 값은 유저 u 가 작성한 포스트에 코멘트나 좋아하는 사진에 추가 등으로 피드백을 준 모든 유저의 ER 값에 영향을 받는다.

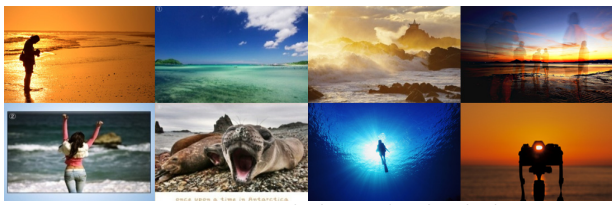
$$ER(u) = d * \sum_{v} \frac{|C_{A_{u,v}}|}{|C_v|} ER(v) + (1 - d)$$

여기서 $|A_u|$는 유저 u 가 작성한 모든 포스트, $|C_v|$는 유저 v 가 작성한 모든 피드백 수를 의미한다. 그리고 $|C_{A_{u,v}}|$는 유저 u 가 작성한 모든 포스트에 대해 유저 v 가 피드백을 준 숫자이다. 따라서 $\sum_u |C_{A_{u,v}}| = |C_v|$의 식이 성립한다. 여기서 d 는 damping factor 이다.

반면 유저의 대중성은 전문성의 반대 개념으로 생각 될 수도 있지만 몇몇 포스트는 높은 전문성을 지닌과 동시에 대중성도 지닐 수 있다. 그에 따라 ER 의 중간 유저를 기점으로 하여 Gaussian Kernel Transform 을 이용하여 유저의 대중성 점수를 구하였다. 유저의 대중성은 Popularity Rank(PR)로 정의한다.

4. 실험 결과

대규모 웹 커뮤니티에서의 실험을 위해 SLR Club (http://www.slrclub.com)과 Flickr(http://www.flickr.com)의 데이터를 사용 하였다. SLR Club 의 데이터는 1 년 6 개월 동안 모은 데이터로서 129,941 개의 포스트와 1,191,716 개의 코멘트, 114,308 명의 유저로 구성 되어 있다. 한편 Flickr 의 데이터는 쿼리 "animal"에 대한 것으로서 4,057,176 개의 포스트와 3,836,607 개의 좋아하는 사진에 추가 한 수, 664,039 명의 유저수로 구성 되어 있다.



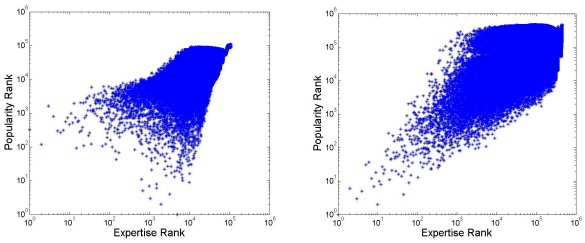
(a) SLR Club 에서 검색 된 사진들

(b) Flickr 에서 검색 된 사진들

(그림 2) 검색 결과 중 높은 전문성과 높은 대중성을 보인 사진들

그림 2-(a)는 SLR Club 에서 쿼리 "바다"에 대한 결과로서 위쪽의 4 개 그림은 전문성이 높은 사진, 그리고 아래 4 개의 사진은 대중성이 높은 사진이다. 그림에서 알 수 있듯이 전문성이 높은 사진들은 주로 분위기나 자연 경관, 사진의 구도, 후처리 기술 등이 많이 고려 된 사진 이었다. 반면에 대중성이 높은 사진들은 여자, 희귀 동물, 신비한 경험 같은 것들이 상위 랭크가 되었다.

그림 2-(b)는 Flickr 에서 쿼리 "animal"에 대한 검색 결과로서 역시 위의 네 장의 사진은 전문성이 높은 사진, 아래 네 장의 사진은 대중성이 높은 사진이다. Flickr 에 대한 결과 에서도 전문성이 높은 사진들은 역시 구도, 자연과의 배치 등이 많이 고려된 것을 알 수 있었다. 반면에 대중성이 높은 사진들은 주로 일상 생활 안의 동물의 모습이나 귀여운 동물 사진들이 주를 이루었다.



(a)SLR Club (b)Flickr

(그림 3) 전문성과 대중성에 기반한 포스트 랭킹의 분포

그림 3 은 유저의 전문성과 대중성에 기반한 포스트 랭킹의 분포를 나타낸 그래프이다. x 축은 전문성 기반의 포스트 랭크이고, y 축은 대중성 기반의 포스트 랭크이다. 먼저 그림 3-(a) SLR Club 의 포스트 랭킹 분포를 보면, 상위 랭크 된 포스트들은 넓은 폭의 분포를 보이면서 전문성과 대중성의 상관관계가 적음을 보이는 반면, 전문성과 대중성 모두를 갖지 못하는 하위 랭크 된 포스트들 일수록 밀집도가 커지는 것을 볼 수 있었다. 그러나 그림 3-(b) Flickr 의 포스트 랭킹에서는 조금 다른 양상을 보였다. 상위 랭크에서 밀집도가 높아지고 하위 랭크에서 밀집도가 낮아지는 현상이 나타났기 때문이다. 이 현상은 소재의 차이 때문이라 할 수 있겠다. "바다"라는 소재는 그 보는 바에 따라 전문성과 대중성이 극명히 갈릴 수 있는 반면, "animal"이라는 소재는 사진에 전문성이 있든 없든 그 선호하는 바의 차이가 적을 것이기 때문이다. 실제로 검색 결과에서도 중첩되는 포스트들이 많이 발견 되었다. 그래도 3-(b) 그래프에서 상위 랭크 된 포스트들의 밀집도가 SLR Club 과 비교하여 높기는 하나 어느 정도 분포를 이루는 것을 볼 수 있는데, 이 분포가 그림 2-(b)와 같은 전문성 있는 포스트와 대중성 있는 포스트를 구별해 내는 것이라 할 수 있겠다.

끝으로 네 명의 사용자에 대하여 252 장의 사진으로 전문성과 대중성에 대한 유저 스터디를 수행 하였다. 전문성 측면에서는 precision 과 recall 이 각각 63.3%, 69.2%를 보였고, 대중성 측면에서는 71.2%와 65.4%를 보였다. 이 두 가지 측면에서의 F1 score 는 66~68%의 분포를 보였다.

5. 결론 및 향후 계획

웹 커뮤니티에서 유저의 성향을 반영하는 효과적인 검색을 위해 유저의 전문성과 대중성을 정의 하고 이를 기반으로 포스트를 랭킹 하였다. 우리의 이러한 접근 방법은 웹 검색에서 유저의 프로파일(profile)을 이용하는 방법과는 다른 것이다.[2] 이러한 유저의 성향이 반영된 포스트 랭킹은 다양한 결과를 나타낸다는 것을 SLR Club 과 Flickr 데이터를 이용한 실험을 통하여 증명하였다. 앞으로도 더 다양한 유저의 성향을 포스트 랭킹에 반영하는 연구를 진행 할 것이다.

참고문헌

[1] Hyoseop Shin, Zhiwei Xu, and Eun Yi Kim, "Discovering and Browsing of Power Users by Social Relationship Analysis in Large-scale Online Communities," Proc. of 2008 IEEE/WIC/ACM conf. on Web Intelligence, Sydney, Australia, Dec. 2008.

[2] J. Luxenburger, S. Elbassuoni, and G. Weikum, "Matching task profiles and user needs in personalized web search," Proc. of 2008 ACM CIKM conference, pages 689-698, Napa Valley, USA, Sep. 2008.