카테고리 상관도 기반 콘텐츠 추천 기법[1)]

최상민○, 한요섭
연세대학교 컴퓨터과학과
jerassi@yonsei.ac.kr, emmous@cs.yonsei.ac.kr

Contents Recommendation based on Category Correlations

Sang-Min Choi○, Yo-Sub Han
Dept. of Computer Science, Yonsei University

1. 서 론

최근 웹2.0의 등장과 인터넷 업체들의 다양한 기능제공으로 인터넷 사용자는 정보 이용자인 동시에 정보 제공자가 되었으며, 웹상에 막대한 정보가 존재하게 되었다. 정보 이용자는 인터넷 검색을 통해 보다 많은 정보를 손쉽게 얻을 수 있게 되었다. 하지만, 인터넷 검색을 통해 얻은 정보는 문제점도 지니고 있었다. 넘쳐나는 검색 결과 때문에 정작 이용자들이 원하는 정보를 얻기 위해서는 많은 시간을 들여야 했다. 이에 따라 보다 효율적인 검색을 위하여 최근 추천 시스템에 대한 연구가 여러 대학에서 진행되고 있다.

일반적인 추천 시스템은 1990년대 이 후 미네소타 대학을 중심으로 발전된 협력적 여과 방식을 기반으로 한다[1][2][3]. 일반적으로 사용되는 방식은 사용자 선호도의 상관도를 이용하여 유사사용자 그룹인 '이웃'을 찾은 뒤, 이 '이웃'의 선호도를 기반으로 추천할 아이템을 결정하는 방식이다. 일반적인 방식은 충분한 양의 사용자 선호도가 필요하다. 왜냐하면 이 방식에서 사용자 사이의 상관도는 통계적 접근을 이용하여 얻기 때문이다. 이는 사용자의 선호도가 충분치 않은 경우, 추천된 정보에 대한 신뢰도가 떨어지게 되는 희소성의 문제와 새로운 사용자나 아이템이 추가되는 경우, 추천의 대상에서 제외되는 콜드 스타트 문제 등을 야기한다[1][2][4].

본 논문에서는 이러한 문제점을 해결하기 위하여 기존의 사용자 선호도를 기반으로 작동하는 추천시스템이 아닌 카테고리의 상관도와 최소한의 사용자 정보를 기반으로 추천이 이루어지는 방식을 제안한다.

본 논문에서는 GroupLens의 3,883편의 영화 database[5]를 이용하여 장르상관도를 구하였다.

2. 본 론

본 논문에서 제안하는 방식으로 영화를 추천하기 위해 첫 번째로 장르끼리의 상관도를 도출해야한다. 장르의 상관도는 영화정보에 포함된 장르조합을 바탕으로 도출한다. 데이터베이스에 존재하는 3,883의 영화는 각각 하나 이상의 장르를 갖는다. 이러한 장르 조합에서 차례대로 기준장르를 선택하여 나머지 장르들과의 수를 계산한다. 가령, 장르 조합이 G1|G2|G5인 경우, 우선 G1을 기준장르로 선택한다. 그리고 나머지 장르인 G2와 G5 사이의 수를 1씩 증가시킨다. 다음으로 G2를 기준장르로 선택하여 G5와의 수를 1 증가시킨다. 즉, n개의 장르조합으로 구성된 영화에서 장르 카운팅은 첫 번째 위치부터 순서대로 기준장르로 삼아 두 번째 위치한 장르부터 n번째 장르까지의 수를 1씩 증가시키는 방식이다. 기준장르는 n번째 장르까지의 카운팅이 끝나고 나면 두 번째 위치한 장르로 재선택된다. 결국, 기준장르는 첫 번째 위

1) 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2010-0009168)

치부터 n-1번째 위치한 장르까지 선택되며 n개의 장르 조합의 경우 총 nC_2번의 카운팅이 이루어진다. 이와 같은 장르 카운팅을 데이터베이스에 존재하는 모든 영화의 장르에 적용하여 장르상관도를 얻어낸다.

다음으로 장르상관도를 이용한 추천을 위하여 식(1)과 입력받은 사용자의 장르 선호도 정보를 사용한다.

$$Tr_1 = \frac{\sum_{i \in up} (\sum_{j \in mg} r_{i_n, j_m} \cdot \mu R_M)}{n(up)} \quad (1)$$

식 (1)에서 up는 사용자가 입력한 선호 장르집합이고 mg는 각 영화가 지닌 장르조합의 집합이다. r_{i_n, j_m}는 장르i_n과 장르j_m과의 상관도이다. 그리고 μR_M은 영화 M의 평균 선호도를 의미한다. 즉, Tr_1은 영화가 지닌 장르조합과 사용자가 입력한 선호 장르와의 상관도를 해당영화의 평균 선호도에 적용한 결과이다. 이 때 장르i_n과 장르j_m이 같을 경우, 상관도를 1로 적용한다.

모든 영화의 장르조합에 장르상관도를 적용하여 평균선호도 값을 변경한다. 즉, 평균선호도에 장르 상관도를 적용한 데이터를 얻고 이를 내림차순으로 정렬하여 상위에 위치한 영화들을 추천한다.

실험은 서로 다른 장르 선호도를 가진 사용자 10명의 선호 장르를 입력받아 추천받은 영화 중 만족하는 영화의 개수를 선택하는 방식으로 진행하였다. 본 논문에서 제안하는 방식과의 비교를 위해 각 영화가 지닌 장르조합에 입력받은 선호 장르가 존재하는 비율을 구하여 각 영화의 평균 선호도에 적용한 실험도 동시에 진행하였다. 실험결과 대부분의 사용자들이 비교를 위한 방식의 결과에 비해 본 논문에서 제안하는 방식의 결과에 비교적 높은 선호를 나타내었다. 그 중 몇 가지를 살펴보면, 장르 11, 13, 16, 17을 선호장르로 입력한 사용자는 본 논문의 제안방식의 결과 10개 중 9개에 만족하였고 비교를 위한 방식의 결과 10개 중 6개에 만족하였다. 그리고 장르 10, 11을 선호장르로 입력한 사용자는 본 논문의 제안방식의 결과 10개중 8개, 비교를 위한 방식의 결과 10개 중 4개에 만족하였다.

3. 결론

기존의 시스템은 희소성 문제 및 콜드 스타트 문제 등을 지니고 있다[1][2][4]. 이는 특정 수 이상의 사용자 선호도 입력을 받아야 추천을 받을 수 있다는 점과 데이터베이스 상에 충분한 양의 선호도가 존재해야 추천의 정확도가 증가한다는 문제이다. 이 문제들은 모두 사용자의 입력과 관련된 문제로서 본 논문에서는 사용자의 최소 입력을 바탕으로 추천이 이루어지는 방식을 제안하였다. 최소 입력은 사용자의 선호 장르이며 시스템은 미리 계산해 놓은 장르끼리의 상관도와 각 영화의 평균 선호도에 선호 장르 정보를 적용하여 추천하는 방식으로 동작한다. 이로써 사용자는 특정 수 이상의 선호도를 입력하는 수고를 덜게 되었다. 즉, 최소 입력만을 가지고 추천을 할 수 있다는 가능성을 제시하였다.

참고문헌

[1] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, “Analysis of Recommendation Algorithms for E-Commerce“, The ACM E-Commerce 2000 Conference, 2000

[2] Daniel Bill, Michael J. Pazzani, “Learning Collaborative Information filters“, Proceedings of ICML, 1998

[3] Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P., and Rie, J., “GroupLens: An Open Architecture for Collaborative Filtering of Netnews“, Proceedings of ACM CSCW94 Conference on Computer Supported Cooperative Work, 1994

[4] Ji-Sun Park, Taek-Hun Kim, Young-Suk Ryu, Sung-Bong Yang, “A Predictive Algorithm Using 2-wway Collaborative Filtering for Recommender Systems“, Dept. of Computer Science, Yonsei University, 2000

[5] http://www.grouplens.org/node/12