사용자 프로파일 정보를 고려한 협력 필터링

*김병만 *이 경[0] *박창석 *김시관 **김주연
*금오공과대학교 컴퓨터공학부
**부천대학 전산정보처리학과
*{bmkim, liqing[0], icisl, sgkim}@se.kumoh.ac.kr, **jykim@mail.bc.ac.kr

A Collaborative Filtering Approach using User Profile

*Byeong Man Kim *Qing Li[0] *Chang-Seok Park *Si-Gwan Kim **Ju-Yeon Kim
*Dept. of Software Engineering, Kumoh National Institute of Technology
**Dept. of Computer Science, Bucheon College

요 약

엄청난 속도로 증가하고 있는 정보의 홍수 시대에서는 정보들을 선별하기 위하여 정보 필터링 기법이 필요하다. 정보 필터링은 내용 기반 방법과 협력에 의한 방법으로 분류할 수 있다. 내용 기반 기법에서는 내용에 기반을 두어 정보를 추출하는 반면 협력 기법은 대상이 되는 사용자에 대한 예측을 하기 위하여 다른 사람들의 의견들을 이용하게 된다. 본 논문에서는 기존 협력 필터링 방법의 문제점을 해결하기 위한 방법의 일환으로 내용 기반 기법과 협력 기법을 보다 유기적으로 결합시키는 연구를 수행하였다. 이를 위해 협력 필터링 틀을 그대로 유지하면서 사용자 프로파일을 효과적으로 이용하는 방법을 제안하였다. 또한, 본 논문에서 제시한 기법을 실험적으로 분석하고 기존의 필터링 기법과 비교함으로써 제시된 기법의 우수성을 보였다.

1. 서 론

최근 들어 엄청난 속도로 정보의 양이 증가하고 있다. 영화, 책, 음악, 뉴스와 특히 온라인 정보의 양은 엄청나게 증가하고 있다. 이러한 정보의 홍수 속에서 검색 시간을 줄이고 우리가 관심있는 정보에 접할 수 있도록 하는 정보의 우선 순위화에 도움이 되는 정보 필터링 기법이 필요하게 된다.

정보를 필터링하기 위하여 기존의 연구 방향은 내용 기반 기법과 협력 필터링 기법에 중점을 두었다. 내용 기반 필터링은 정보 검색의 표현과 사용자 프로파일의 내용 표현을 비교하여 사용자에게 적절한 정보를 선택을 해 준다. 내용 기반 정보 필터링은 불리안 모델, 벡터공간 모델, 확률모델, 인공신경망 모델, 퍼지집합 모델과 같은 기법을 사용하여 주제에 적절한 텍스트 정보를 찾아 주는데 아주 효과적인 것이 증명되었다. 그러나, 내용 기반 기법은 아이템(찾고자하는 정보)은 반드시 기계가 분류할 수 있는 형태(예: 텍스트)로 되어 있어야 하고 비슷한 성향을 갖는 다른 사람등의 정보를 이용하지 못한다.

협력 필터링은 다른 사람들의 관심 사항을 예측하기 위해 동료의 의견을 사용하는 기법으로 제록스 알토 연구소(PARC)의 Nichols등에 의해서 개발된 Tapestry 문서 필터링은 협력 필터링을 적용한 최초의 시스템이다 [1,2]. 미네소타 대학의 GroupLens 프로젝트는 현재로서 가장 유명한 협력 필터링 시스템 형태이다 [3]. 협력 필터링 시스템은 같은 취향이나 취미를 가진 사람들의 정보를 이용해 추천을 할 때 도움을 줄 수 있는 가장 널리 사용되는 시스템이다. 링고 시스템[4]은 음악 앨범 추천용으로, 무비렌즈 시스템[5]은 영화 추천용으로, 제터 시스템[5]은 조크 추천용으로, 그리고 플라이캐스팅은 온라인 라디오 추천용으로 개발된 협력 필터링 시스템들이다.

협력 필터링은 내용 기반 필터링의 단점을 어느 정도 해결한다. 필터링 되는 항목은 반드시 테스트 형태일 필요는 없다. 또한, 항목의 내용이 아닌 타 사용자의 항목에 대한 평가 정보에 근거를 두고 추천이 이루어지기 때문에 추천의 질을 향상시킬 수 있다. 그러나, 협력 필터링 기법은 연구적인 측면이나 실용적인 측면에서 아주 성공적이지만 효율적인 정보 필터링을 고려한다면 다음과 같은 문제점을 안고 있는 실정이다.

● 희소성 문제 (Sparsity problem) : 많은 정보 도메인에서는, 항목들의 개수는 개별 사용자들이 소화할 수 있는 개수를 훨씬 초과한다. 따라서, 모든 사용자들에 대한 모든 항목들의 평가들을 포함하고 있는 행렬들은 매우 드문드문한 분포성을 띤다. 비교적 조밀한 정보 여과 도메인들은 가끔씩 여전히 98%-99% 의 희소성을 띠는데 이는 협동 여과 예측들에 기반이 되는, 충분한 사람들에 의해서 평가되어진 문서들을 찾기 어렵게 만든다.

● 회색 양 (Gray sheep) : 소규모 혹은 심지어 중규모 사용자 그룹에서는, 순수한 협동 여과 시스템들로부터 이득을 얻지 못할 수 있는 사용자들이 존재한다. 왜냐하면, 그들의 의견들은 어떤 사용자 그룹과도 일관성 있게 일치되거나 혹은 불일치 되지 않기 때문이다. 이들 사용자들은 사용자와 시스템 초기 구동 단계 후에도 좀처럼 정확한 협동 여과 예측들을 제공받지 못하게 된다.

● 확장성 (Scalability) : 협력 필터링 분야에서 주로 사용하는 최근접 이웃 알고리즘 (Nearest Neighbor Algorithm)은 사용자와 항목 수에 비례해서 계산 시간이 비례한다. 따라서, 사용자 수와 항목 수가 수백만 되는 환경하에서는 이러한 계산 시간이 치명적일 수 있다.

내용 기반의 필터링은 사용자들의 추천의 질 문제를 제외하고는 위에 나열한 문제점을 가지고 있지는 않는다. 따라서, 보다 우수한 정보 필터링을 위하여 2가지 필터링의 기법을 결합하여 각 기법의 장점을 이용하는 것이 당연하다. 따라서, 본 논문에서는 이러한 측면에서 내용 기반 필터링과 협력 필터링 기법을 보다 유연하게 결합할 수 있는 방법을 제시하고 계산 모델에 대해 여러 가지 요소들을 변화시키면서 분석을 하였다. 또한, 기존의 협력 필터링 방법과 실험을 통해 비교, 분석하였다.

2. 제안 방법

본 논문에서는 사용자-사용자의 유사도를 계산하기 위해 사용자 프로파일 정보와 항목 평가 정보를 통합한다 (그림 1 참조). 제안 방법은 다음과 같다.

① 사용자 프로파일의 클러스터링하고 이를 이용하여 그룹 평가 행렬을 생성한다. (자세한 내용은 2.1 절 참조)

② 사용자간 유사도를 계산한다. 기존 협력 필터링에서 사용하던 사용자-항목 평가 행렬과 1단계에서 생성된 사용자-그룹 평가 행렬을 이용하여 사용자간 유사도를 구한다. (자세한 내용은 2.2 절 참조)

③ 이웃 사용자의 평균으로부터 가중치 편차 평균을 계산하여 항목에 대한 예측을 한다.

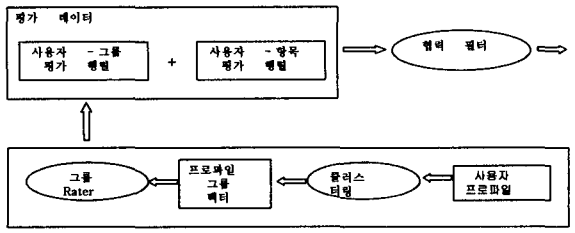

그림 1. 제안 방법 개략도

2.1 그룹 평가(Group Rating)

본 논문에서는 내용기반 필터링 방법을 협력필터링 프레임 워크 안에 수용하기 위해 그룹 평가 정보를 사용한다. 그룹 평가는 특정 사용자가 정해진 사용자 그룹들에 속할 정도를 나타내는 것이다. 사용자 그룹은 결정하는 방법은 다양하지만 본 논문에서는 사용자 프로파일의 내용을 기반으로 K-means 알고리즘 [6]을 적용하여 자동으로 형성하는 방법을 선택하였다. 그림 2는 본 논문에서 사용되는 그룹평가 알고리즘을 보여 주고 있다.

입력 : 클러스터 수 k와 사용자 프로파일 집합

출력 : k개의 클러스터와 각 사용자가 해당 클러스터에 속할 정도

(1) 각 사용자가 속할 클러스터를 무작위로 할당

(2) 클러스터에 더 이상 변화가 없을 때까지 다음의 과정을 반복

(a) 각 클러스터의 중심을 구함.

(b) 각 클러스터의 중심과 사용자 프로파일 간의 유클리디리디안 거리를 계산해서 가장 가까운 클러스터에 해당 사용자재 할당

(3) 아래의 식을 이용하여 각 사용자가 각 클러스터에 속할 정도 계산.

$$P = 1 - \frac{E(j,m)}{Max_i\ E(i,m)}$$

여기서, j는 j번째 사용자 프로파일 벡터를, m은 (2) 단계에서 구해진 m 번째 클러스터의 중심벡터를, E(j,m)은 j와 m 간의 유클리디안 거리를, $Max_i\ E(i,m)$은 모든 사용자의 프로파일 벡터와 m 번째 클러스터와의 거리 중 최대값을 의미한다

그림 2. 그룹 평가 알고리즘

단계 1과 2는 K-means 알고리즘을 적용하여 클러스터를 생성하는 단계이며 단계 3은 구해진 클러스터의 중심들과 각 사용자의 프로파일 벡터 간의 유사도를 구하는 과정이다. 보통 클러스터링의 결과는 해당 객체가 속할 그룹만 결정되지만, 단계 3에서는 해당 객체가 각 그룹에 속할 가능성이 계산되어 지고 이를 행렬 형태로 표현할 수 있다. 이 행렬을 사용자-그룹 평가 행렬이라 한다.

2.2 유사도 계산

협력 필터링 방법을 사용하여 특정 항목에 대한 특정 사용자의 평가 정도를 예측하기 위해서는 위에서 얻어진 행렬과 원래 주어진 사용자-항목 평가 행렬을 사용하여 사용자간의 유사도를 계산할 필요가 있다. 여기서는, 사용자-항목 평가 행렬을 이용하여 사용자간 유사도를 구하는 기존의 대표적인 방법 2가지를 간단히 소개하고, 본 논문에서 사용자-그룹 평가 행렬을 추가로 고려한 사용자간 평가방법을 소개한다.

2.2.1 피어슨 상관 공식을 이용한 유사도 [3]

가장 많이 사용되는 가중치 측정은 Pearson correlation coefficient 방법이다. 피어선 상호관계는 두 변수간에 존재하는 선형 관계의 정도를 측정한다. 피어선 상호관계는 선형 리그레션 모델에서 유도되며 상호관계는 선형적이며 에러는 서로 독립적이고 평균 0인 확률 분포와 독립 변수의 모든 설정에 대한 상수 편차를 가진다는 가정에 기반을 둔다. [7]

$$\mathrm{sim}(k,l) = \frac{cov(k,l)}{\sigma_k \cdot \sigma_l} = \frac{\sum_{i=1}^{m}(R_{k,i} - \overline{R_k})(R_{l,i} - \overline{R_l})}{\sqrt{\sum_{i=1}^{m}(R_{k,i} - \overline{R_k})^2}\sqrt{\sum_{i=1}^{m}(R_{l,i} - \overline{R_l})^2}}$$

여기서, m 은 항목 수를, $R_{k,i}$ 는 사용자 k 가 i 번째 항목에 대한 평가치를, $R_{l,i}$ 는 사용자 l 이 i 번째 항목에 대한 평가치를, $\overline{R_k}$ 와 $\overline{R_l}$은 사용자 k 와 l 의 m 개의 항목에 대한 평가치의 평균을 나타낸다.

2.2.2 보완 코사인 유사도(Adjust Cosine Similarity) [8]

코사인 유사도는 유사도를 계산하기 위해 한 때 가장 많이 사용된 기법이지만 단점을 가지고 있다. 서로 다른 사용자들 사이에서 평가 스케일(척도)의 차이는 아주 많이 다른 유사도를 초래한다는 것이다. 예를 들어, Bob이 가장 선호하는 영화를 4라고 평가를 생각했다면 5라는 평가는 하지 않을 것이다. 나쁜 영화에 대한 평균 평가가 2인데도 1이라고 평가할 수 있다. 그러나, Oliver는 가장 좋은 영화는 5, 나쁜 영화는 2라고 평가한다. 기존의 코사인 유사도를 사용한다면 2명에 대한 유사도는 아주 다를 것이다. 보완된 코사인 유사도는 이러한 단점을 보완할 수 있다.

$$\mathrm{sim}(K,L) = \frac{\sum_{i=1}^{m}(R_{k,i} - \overline{R_i})(R_{l,i} - \overline{R_i})}{\sqrt{\sum_{i=1}^{m}(R_{k,i} - \overline{R_i})^2}\sqrt{\sum_{i=1}^{m}(R_{l,i} - \overline{R_i})^2}}$$

여기서, m은 항목 수를, $R_{k,i}$ 는 사용자 k 가 i 번째 항목에 대한 평가치를, $R_{l,i}$ 는 사용자 l 이 i 번째 항목에 대한 평가치를, $\overline{R_i}$ 는 항목 i 에 대한 모든 사용자의 평가치 평균을 나타낸다.

2.2.3 사용자-그룹 평가 행렬을 고려한 사용자간 유사도 계산 방법

보통, 사용자-항목 평가 행렬 정보는 이산적인 값 (예, MovieLens 데이터인 경우 평가치는 1과 5상이의 정수)을 갖으며 사용자-그룹 평가 행렬은 0과 1 사이의 연속적인 값을 갖는다. 따라서, 한쪽의 값을 다른쪽의 값으로 확대 또는 축소하여 동일한 유사도 척도를 적용하는 방법도 고려할 수 있으며 각각의 행렬에 대해 다른 유사도 척도를 적용시키고 이들의 결과를 조합하여 사용하는 방법도 고려해 볼 수 있다. 본 논문에서는 아래의 세가지 방법에 대해서 그 성능 평가를 수행하였다.

- Non-enlarged Pearson : 사용자-항목 평가 행렬과 사용자-그룹 평가
- Enlarged Pearson : 사용자-그룹 평가 행렬을 사용자-항목 평가 값의 범위로 확대하고 두 개의 행렬을 하나의 행렬로 취급한 후 기존의 Peason 상관 공식을 이용
- Combination Approach : 사용자-항목 평가 행렬에는 피어슨 상관 공식을, 사용자-그룹 평가 행렬에는 보완 코사

인 유사도를 적용시킨 후 결과의 평균을 사용

2.3 협력 예측

사용자 k의 항목 i에 대한 예측을 구하기 위해 GroupLens에서 제안한 식 [3]을 사용하였다. 여기서는, 항목에 대한 예측은 이웃의 평균값으로부터 편차의 가중치 평균을 수행함으로써 계산된다. 그리고 사용자의 유사성에 기반한 가장 인접한 N 개의 이웃을 선택하기 위하여 top N 규칙을 사용한다.

$$P_{k,i} = \overline{R_k} + \frac{\sum_{u=1}^{n}(R_{u,i} - \overline{R_u}) \times \mathrm{sim}(k,u)}{\sum_{u=1}^{n}|\mathrm{sim}(k,u)|}$$

$P_{k,i}$ 는 항목 i 에 대한 사용자 k 에 대한 예측을 표시한다. n 은 사용자 k 의 최인접 이웃의 수, $R_{u,i}$ 는 항목 i 에 대한 u 평가, $\overline{R_k}$ 는 항목에 대한 사용자 k 의 평균 평가, sim(k, u)은 사용자 k 와 이웃 u 사이의 유사도, $\overline{R_u}$ 는 항목에 대한 사용자 u 의 평균 평가를 의미한다.

3. 평가

3.1 평가 환경

현재 웹에 기반을 둔 추천 시스템인 MovieLens에서 수집된 영화 평가에 대한 데이터를 이용하였다. 데이터 집합은 943명의 사용자, 1682개의 영화를 각 사용자가 적어도 20개의 항목에 대해 평가를 한 100,000개의 평가를 포함하고 있다. MovieLens 의 평가는 사용자들이 1에서 5사이의 정수값으로 직접 평가 자료를 입력하였다. 이 데이터 집합은 훈련 집합과 테스트 데이터 집합으로 구성되어 있다.

MovieLens 데이터에는 사용자 프로파일 정보가 명시적으로 주어지지 않는다. 본 실험에서는 사용자가 평가한 영화의 정보를 갖고 사용자 프로파일을 구성하였다. 즉, 사용자가 평가한 영화의 장르와 그의 가중치로 사용자의 프로파일을 구성하였다. 특정 사용자가 해당 장르를 좋아할 가중치를 계산하기 위해 더 높은 점수로 등급이 매겨지면 질 수록 사용자는 그 장르를 더 선호한다고 가정하였다. 예를 들어, 테스트 데이터 집합에서 사용자 a가 영화 1, 2, 3 과 4에 대한 등급을 매겼다고 하고 영화 1, 2 와 3은 영화 장르 "love"에 속하고 영화 4는 ""action"에 속한다고 하자. 그러면 사용자 a의 프로파일은 {love (75%), action (25%)}가 된다.

MAE (Mean Absolute Error)는 테스트 자료에서 실제 사용자 평가에 대하여 예측치를 비교함으로써 추천 시스템의 정확도를 평가하는데 가장 많이 사용되고 있다. 본 논문에서도 이 척도를 사용하였다. MAE는 모든 테스트 대상에 대해서 평가치와 예측치간의 오류를 구하고 이 오류의 절대값을 합한 후 테스트 대상의 수로 나누어 줌으로써 얻을 수 있다. MAE가 낮을수록 예측의 정확도는 좋아지게 된다.

$$MAE = \frac{\sum_{i=1}^{n}|p_i - q_i|}{n}$$

여기서, n은 평가 대상의 수를, p_i는 대상 i에 대한 예측치를, q_i는 대상 i에 대한 실 평가치를 나타낸다.

3.2 성능 비교

본 논문에서는 2.2.3 절에서 제안된 3가지 경우를 구현하여 기존의 순수 협력 필터링 방법, 즉, 사용자-항목 평가 행렬만을 이용하고 사용자간 유사도는 피어슨 상관 공식을 이용하는 방법과 비교하였다. 그림 3에서 "conventional Perason"이 기존 순수 협력 필터링 방법을 의미한다.

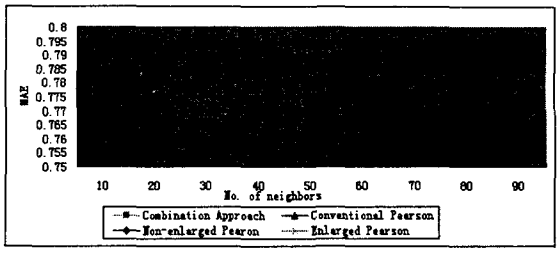

그림 3. 고전적 협력 필터링과의 성능 비교

그림 3에서 보는 바와 같이, 사용자-그룹 평가의 값을 확대시킬 경우가 그렇지 않은 경우보다 성능이 월등함을 알 수 있다. 더욱이, 사용자-그룹 평가의 값을 확대시키지 않을 경우는 사용자-그룹 평가 정보를 사용하지 않는 경우, 즉 순수 협력 필터링 보다 성능이 나쁨을 알 수 있다. 이는 그룹 평가정보가 제대로 반영이 되지 않고 오히려 부정적으로 작용하는 것으로 보인다. 또한, 두 개의 행렬에 다른 유사도 척도를 적용한 경우가 가장 성능이 우수함을 알 수 있다. 이는 두 개의 행렬이 이질적인 데 기인한 것으로 보인다.

5. 결론 및 향후 연구

본 논문에서는 기존 협력 필터링의 문제점을 해결하기 위해 사용자 프로파일을 기반으로 한 사용자 그룹 평가 정보를 추가로 이용하는 방법을 제시하였다. 실험을 통해 사용자간 유사성 계산 시 추가의 평가 정보를 어떻게 이용하는 것이 좋은지에 대해서도 살펴보았다. 결론적으로 사용자 그룹 평가정보가 예측 성능에 긍정적인 효과를 보임을 알 수 있었고 기존의 항목 평가 정보와 사용자 그룹 평가 정보에 다른 유사도 척도를 사용하는 것이 더 좋음을 알 수 있었다. 앞으로, 본 방법의 성능에 영향을 미치는 요소들에 대해 추가의 연구가 필요하며, 특히, 기존 다른 연구와의 성능 분석이 필요하다.

참고문헌

[1] Donna Harman. Overview of the third Text Retrieval Conference (TREC-3). In D. K. Harman, editor, *Overview of the Third Text Retrieval Conference (TREC-3)*, pages 1--19. *NIST, U.S. Department of Commerce*, 1994.

[2] Douglas B. Terry. A tour through tapestry. In *Proceedings of the ACM Conference on Organizational Computing Systems (COOCS)*, pages 21--30, November 1993.

[3] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. GroupLens: An open architecture for collaborative filtering of Netnews. In Richard K. Faruta and Christine M. Neuwirth, editors, In *Proceedings of the Conference on Computer Supported Cooperative Work*, pages 175--186. ACM, October 1994.

[4] Upendra S. and Patti M. Social Information Filtering: Algorithms for Automating "Word of Mouth", In *proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, pages. 210---217, 1995

[5] D. Gupta, M. Digiovanni, H. Narita, and K. Goldberg. Jester 2.0: A New Linear-Time Collaborative Filtering Algorithm Applied to Jokes. In *Proceedings of Workshop on Recommender Systems: Algorithms and Evaluation*, Aug. 1999.

[6] Han, J., and Kamber, M. Data mining: Concepts and Techniques. *New York: Morgan-Kaufman*, 2000.

[7] McClave and Frank H. Dietrich II statistics. *Dellen Publishing Company*,1988

[8] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl Item-based Collaborative Filtering Recommendation Algorithms <http://www10.org/cdrom/papers/519/index.html>. *WWW10*, May 1-5, 2001, Hong Kong.