협력 필터링을 이용한 P2P 파일 추천 시스템

민수홍, 조동섭
이화여자대학교 과학기술대학원 컴퓨터학과

P2P File Recommendation System using Collaborating Filtering

Su-Hong Min, Dong-Sub Cho
Dept. of Computer Science and Engineering, Ewha Womans University

Abstract - 최근 P2P 모델을 기반으로 한 애플리케이션의 등장으로 다양한 자원을 효율적으로 이용할 수 있게 되었다. P2P에서는 여러 대의 클라이언트를 상호 긴밀하게 연결함으로써 한 대의 서버에 다수의 클라이언트를 연결했을 때 보다 확실한 네트워크의 효과를 기대할 수 있다. 그러나, 기존의 P2P 모델의 경우, 다수의 피어가 네트워크에 참여하여 방대한 양의 자원을 공유할 경우, 원하는 자원을 검색하는데 많은 시간이 소요되는 문제점이 있다. 본 논문에서는 자원 검색의 비효율적인 문제를 해결하고자 협력 필터링 알고리즘을 이용해 P2P 파일 추천 시스템을 제안하고자 한다. 제안한 P2P 시스템은 피어(Peer)들을 유사한 패턴을 갖는 가상 그룹으로 형성해, 그룹 내에서 유용한 자원들을 검색 없이 공유할 수 있도록 하였으며, 자원의 선호도를 기반으로 요청한 자원 외에 추천 시스템을 통해 선호하는 자원을 예측해 제공할 수 있도록 하였다.

1. 서 론

대부분의 인터넷 서비스는 전통적으로 분산된 클라이언트/서버 모델을 기반으로 하며, 현재 우리가 사용하는 대부분의 인터넷 애플리케이션(WWW, FTP, telnet 그리고 e-mail 등)들이 이 모델을 기반으로 하고 있다. 그러나 기하급수적으로 그 규모가 커져 가는 인터넷에 비해 기존의 클라이언트/서버 모델은 모든 서비스가 서버에 집중되어 있어, 인터넷의 자원인 정보, 대역폭, 컴퓨팅 자원을 활용하는데 있어 한계점이 있다. 이러한 문제를 해결하고자 P2P 모델이 등장하였다. P2P는 네트워크에 참여하는 모든 클라이언트가 서버의 역할을 동시에 수행한다. 따라서, 중앙 집중형 서버가 없는 환경에서 하나의 클라이언트에서 다른 클라이언트로 매우 다양한 경로를 통해 상호간 통신을 할 수 있다. 각각의 클라이언트들은 자신의 자원에 대한 접근을 다른 클라이언트에 허가함으로써 P2P 네트워크에 참여할 수 있도록 한다 [1,2]. 그러나 P2P 모델의 경우, 불특정 다수에 의해 많은 양의 자원이 공유되기 때문에 원하는 자원을 찾는데, 검색 시간이 많이 소요되며, 사용자의 유용한 정보를 검색하는 데 비효율적이다. 예를 들어, 소리바다나 냅스터의 경우, 수많은 피어가 네트워크 상에 연결되어 있기 때문에 그들이 공유하고 있는 자원의 양 또한 방대하다. 따라서, 이러한 연결망으로 된 P2P 서비스를 이용해 사용자가 원격 피어에게 자원을 요청할 경우, 원하는 자원을 검색하는데 상당한 시간이 소요된다. 본 논문에서는 이와 같은 문제점을 해결하기 위해 전자 상거래 서비스에서 구매 촉진을 위해 사용하는 협력 필터링 알고리즘을 이용하였다. 디렉토리 서버는 유사한 패턴을 갖는 사용자들에 대해서 가상 그룹을 형성하여, 검색을 거치지 않고도 유용한 자원을 공유할 수 있도록 한다. 또한, 추천 시스템의 예측 값을 이용해 사용자가 검색을 통해 요청한 자원 외에 선호도가 높은 자원을 추천하여 제공하도록 한다.

본 논문의 구성은 다음과 같다.

2장에서는 관련 연구로서 제안한 P2P 모델과 사용자 기반 협력 필터링에 대해 기술하고, 3장에서는 협력 필터링 알고리즘을 사용하여 상관계수에 따른 가상 그룹을 형성하고, 예측에 의해 추천 리스트를 제공하는 방법에 대해 기술한다. 마지막 4장에서는 결론 및 향후 방향에 대해 기술한다.

2. 관련 연구

이 장에서는 본 논문에서 제안한 P2P 모델과 추천 시스템을 위해 사용된 협력 필터링에 대해 알아보고자 한다.

2.1 중앙집중형 P2P 모델

중앙집중형 P2P 모델은 클라이언트 상호간 효율적으로 통신하고 필요한 정보를 전달하기 위해 중간에 서버를 두는 방식이다. 클라이언트가 서로 통신을 하기 위해 최초로 IP 주소를 서버에 인덱스하며, 그 이후 서버의 도움 없이 클라이언트 상호간 직접 정보를 전달하게 된다. 파일을 공유하기 위해 사용자가 서버에게 질의를 전송하면, 서버는 이에 대한 응답으로 접속된 피어들의 목록과 파일의 위치 정보를 제공해 준다. 요청을 하는 피어는 요청한 파일의 목록에 근거해서 피어들에게 개별적으로 접근한다. 일단 요청을 하는 피어가 원하는 파일의 위치를 파악하면, 서버는 피어들의 연결을 개시하고, 피어는 네트워크를 통해서 원하는 파일을 다운로드 한다. 이 모델은 분산형 모델에 비해 피어들이 다른 피어들의 목록을 얻기가 쉬우며, 원하는 자원을 서버의 검색기능을 통해 손쉽게 찾을 수 있어 편리하다. 이와 같은 방식을 사용하는 애플리케이션으로는 냅스터, 소리바다, MSN 메신저 등이 있다 [1,2].

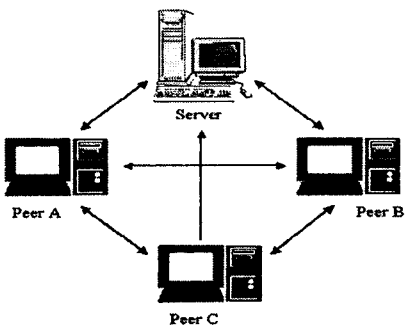

그림 1 중앙집중형 모델

이 논문은 2003년도 두뇌한국21사업에 의하여 지원되었음

2.2 사용자 기반 협력적 필터링

사용자 기반 협력 필터링은 전자 상거래에서 구매 촉진을 위해 사용된다. 이는 특정 고객의 상품에 대한 선호도를 예측하기 위하여 대부분의 경우, 식(2)에 나타나 있는 피어슨 상관 계수를 이용하여 유사한 선호도를 가지는 이웃들을 정하고, 식(1)에 의해 예측 선호도 값을 계산한다.

$$U_x = \overline{U} + \frac{\sum_{J \in Raters} (Jx - \overline{J}) r_{UJ}}{\sum_{J \in Raters} |r_{UJ}|} \quad (1)$$

여기서,

$$r_{UJ} = \frac{\sum (U - \overline{U})(J - \overline{J})}{\sqrt{\sum (U - \overline{U}) \cdot \sum (J - \overline{J})^2}}, \; -1 \le r_{UJ} \le 1 \quad (2)$$

U_x 는 상품 x에 대한 고객 u의 예측된 선호도이고, r_{UJ} 는 고객 u와 j의 상관관계를 나타내며 두 고객 모두 선호도를 표시한 상품에 대해서만 계산된다. 여기서, Jx는 상품 x에 대한 고객 j의 선호도를 나타내며, $\overline{J}$ 는 고객 j의 평균 선호도를 의미한다. r_{UJ} 가 1에 가까울수록, 두 고객의 선호도 경향이 매우 유사함을 나타내고, -1에 가까울수록 반대의 선호 경향을 나타낸다. Raters는 테스트 상품에 대해 선호도를 표시한 고객들을 의미한다.

협력 필터링 방법을 적용한 Tapestry 는 협력 필터링 방법을 가장 먼저 적용한 문서 필터링 시스템으로 워크그룹과 같은 공동체 구성원들의 의견에 기반 하여 추천을 해주므로 개인화 된 추천 서비스는 제공해 주지 못한다. 최근 몇 년 동안에는 특히 자동화된 협력 필터링 시스템이 많이 개발되었는데, 그 중 GroupLens research system은 고객과 유사 선호도를 가지는 이웃들의 의견에 기반 하여 유즈넷 뉴스와 영화에 대한 추천을 수행함으로써, Tapestry 의 문제점을 보완하면서 성능을 인정받은 시스템이다. GroupLens를 포함한 대부분의 협력 필터링 기법을 사용하는 추천 시스템들로 Ringo, Video Recommendation 등이 있으며, 이들은 모두 피어슨 상관계수를 사용하여 유사 선호도를 가지는 이웃들을 결정한다 [3,4,5].

3. 협력 필터링을 이용한 P2P 파일 추천 시스템

본 논문은 기존의 P2P 디렉토리 서버를 확장해 P2P를 이용하는 사용자를 가상 그룹을 만들어, 유사한 패턴을 갖는 사용자를 연결시켜 주며, 또한 선호도가 높은 자원을 추천 시스템을 통해 제공한다. 예를 들어, 대학과 같은 큰 도메인 안에 다수의 피어가 네트워크로 연결되어 있다면, 이는 학과, 학생, 교수 등과 같은 유사한 패턴을 가진 사용자들을 가상 그룹을 형성하여 유용한 자원들을 공유할 수 있도록 한다.

피어가 P2P 애플리케이션을 통해 로그인 할 경우, 디렉토리 서버는 새로 접근한 피어와 기존의 사용자들과의 상관 관계를 구해 유사한 패턴을 갖는 가상의 사용자 그룹에 속하도록 한다. 그룹에 속한 피어는 원격 피어들의 목록과 공유하는 자원들의 리스트를 제공받는다. 이때, 서버는 비슷한 패턴을 가진 피어들의 파일 목록을 우선순위로 제공한다. 피어가 애플리케이션을 통해 자원을 검색하면, 서버는 결과 값으로 요청한 자원 외에 추천 리스트를 함께 제공한다.

본 논문에서 제안하는 P2P 파일 추천 시스템은 특정 도메인 안에 다수의 피어가 네트워크 상에 연결되어 있으며, 유사한 패턴을 갖는 사용자 그룹을 형성할 수 있도록 사용자의 패턴을 구별할 수 있는 환경 하에 있다고 가정한다.

3.1 상관관계에 따른 그룹 형성

사용자가 P2P로 연결된 네트워크에 접근하면, 디렉토리 서버는 이전에 로그인한 피어들이 제공한 자원들에 대한 선호도를 기반으로 유사한 패턴을 갖는 피어 그룹을 찾는다. 서버는 피어슨 상관계수를 사용해 원격 피어들이 공유하고 있는 자원에 따라 다른 피어와의 유사성을 결정하여 그룹을 형성한다. 이때, 물리적인 네트워크 구조에는 변화가 없으며, 서버에 의해 가상적인 연결을 통해 그룹이 형성된다. P2P 서비스를 이용하는 사용자는 참여와 탈퇴가 빈번하게 일어날 수 있음을 고려해야 한다. 따라서, 물리적인 네트워크 구조를 변경할 경우, 시간에 따라 네트워크 구조가 자주 변경되어야 하는 문제점이 있다.

상관 분석이란 두 변수간에 얼마나 밀접한 관계를 가지고 있는가를 분석하는 방법으로 두 변인간의 상호관련성에 대해 측정하는 통계적 방법이다 [6]. 사용자 기반 협력 필터링 알고리즘에서는 문서에 대한 평가를 예측하기 위해서 상관 관계 계수를 사용한다. 이 값은 -1과 1 사이의 값을 가지게 된다. 상관관계 값이 1이면 Perfect positive relationship 이라고 하며, 값이 -1이면 Perfect negative relationship 이라고 한다. 만약 값이 0이면 relationship이 존재할 수도 존재하지 않을 수도 있다 [7].

다음은 상관계수에 대한 해석은 다음과 같다.

표 1 상관계수의 일반적인 의미

<0.2 관계가 거의 없는 경우(little if any correlation)
0.2-0.4 낮은 상관관계(low correlation)
0.4-0.7 비교적 높은 상관관계(moderate correlation)
0.7-0.9 높은 상관관계(high correlation)
>0.9 매우 높은 상관관계(very high correaltion)

협력 필터링 알고리즘에서 사용자들 사이의 상관 관계를 구하는 식은 다음과 같다.

$$W_{a,u} = \frac{\sum_{i \in I_a \cap I_u} (r_{ai} - r_a^*) \times (r_{ui} - r_u^*)}{\sigma_a \sigma_u} \quad (3)$$

위 식 (3)에서 r_{ai} 는 사용자 a가 공유하는 자원 I에 대해 평가한 선호도 값을 나타내며, r_{a^*}는 사용자 a의 선호도들의 평균을 나타낸다. 위 식의 분모는 두 사용자 a와 u가 모두 선호도를 표시한 자원에 대한 값을 구한다. 또한 σ_a는 사용자 a의 표준편차이다. 이를 쉽게 표로 나타내면 다음과 같다.

표 2 사용자-공유자원 행렬

	item 1	item 2	item 3	...	item i-1	item i	평균
user 1		2	4		5	3	r_1^*
user 2	4		5	...	2	4	r_2^*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
user u	2	1	3	...	1		r_u^*
평균	r^*_1	r^*_2	r^*_3	...	r^*_{i-1}	r^*_i	r^{**}

다음은 상관관계를 구하는 알고리즘이다.

```
search(a, table, &n_resource, 1);
for(i=0; i<count; i++) {
    search2(find2[i].code, name[i], table, &n_resource);
    if (status!=0) {
        u1 += (status-t1)*(find2[i].score-t2);
    }
}
    dev1 = sqrt(dev1);
    dev2 = sqrt(dev2);
    corr = u1/(dev1*dev2);
```

그림 2 상관 관계

함수 search2는 두 사용자가 같은 자원에 대해 선호도를 제공한 자원이 있는지 찾는 함수이다. 동시에 선호도를 제공한 자원이 있으면 그 자원에 대한 선호도를 돌려준다. 또한 dev1 과 dev2는 두 사용자의 표준편차를 나타낸다. 이 값들을 가지고 최종적으로 두 사용자의 상관관계를 구한다.

3.2 예측에 의한 추천 리스트 제공

다른 사용자들과의 상관 관계를 모두 구하고 난 후, 유사한 패턴을 지닌 사용자 그룹이 정해지면, 그룹 내에서 유용한 자원에 대한 선호도를 예측해 추천 리스트를 제공한다. 사용자는 P2P 애플리케이션의 검색 창을 이용해 원하는 자원을 검색하며, 서버는 이를 사용자 그룹 내에서 검색하며, 요청한 자원과 함께 유용한 자원을 기존의 사용자들의 자원에 대한 선호도를 바탕으로 예측해서 추천 리스트를 제공한다. 선호도의 예측에는 상관관계를 포함한 평균값을 이용한다. 추천 리스트를 제공하기 위한 선호도를 구하는 식은 다음과 같다.

$$P_{ai} = r_{a*} + \frac{\sum_{u \in U_i} W_{a,u}(r_{ui} - r_{u*})}{\sum_{u \in U_i} W_{a,u}} \quad (4)$$

위와 같은 방법으로 예측 값을 구하는 예를 들면, 사용자 행렬은 다음과 같다.

표 3 예제 행렬

	1	2	3	4	5	6
A	1	5		2	4	
B	4	2		4	1	2
C	2	4	3			5
D	2	4		5	1	

예를 들어, 사용자 A의 아이템 6에 대한 예측 값을 구하려고 한다면 우선 이미 아이템 6에 대한 편가를 제공한 사용자 B, C와의 상관 관계를 구해야한다. 먼저 A와 B 사이의 상관 관계는 아이템 1,2,4,5를 가지고 위의 식 (3)을 가지고 계산한다. 위의 식의 결과에 따라 A와 B의 상관 관계는 -0.8이며, A와 C의 상관 관계는 1이 된다. 이렇게 구한 상관 관계를 기반으로 사용자 A와 아이템 6에 대한 예측 값을 구한다. 예측 값은 식 (4)를 가지고 구하며, 따라서, 사용자 A의 아이템 6에 대한 예측 값은 4.11이 된다.

다음은 예측 값을 계산하는 알고리즘이다.

```
for(a=0; a(총 자원수; a++)
    search2(a, name[i],table, $n_resource);
    search(a, table, &n_resource,1);

    if (status == 0) {
            for(i=0; i(count; I++) {
            corr - f_corr(x,y);
            up+=corr*( r_ui − r_u* );
            dn+=corr ;
            }
           p = t1 + up/dn;
    }
}
```

그림 3 예측 값 계산

함수 search2는 현재 사용자가 이 자원에 대해 선호도를 제공했는지 찾는 함수이며, 자원에 대해 선호도를 제공하지 않았을 경우, search 함수를 사용하여 기존에 그 자원에 선호도를 제공한 사용자를 찾아 상관관계르 구하고 최종적인 예측 값을 구한다.

3. 결 론

본 논문에서는 기존의 P2P 모델의 검색의 효율성을 문제를 해결하기 위해, 협력 필터링 알고리즘을 사용하는 P2P 파일 추천 시스템을 제안하였다. 제안한 P2P 시스템은 유사한 패턴을 갖는 사용자들을 협력 필터링 알고리즘의 상관 계수를 사용해 가상으로 그룹을 형성하여, 검색 없이 유용한 자원들을 공유할 수 있도록 하였다. 또한, 기존의 사용자들의 자원에 대한 선호도를 조사하여, 사용자가 요청한 자원 외에 추천 시스템을 통해 선호하는 자원을 예측해 제공할 수 있도록 하였다.

향후 연구로는 본 논문에서 제안한 P2P 시스템에서 자원 검색의 효율성 부분에 대해 성능 평가를 할 예정이다.

[참 고 문 헌]

[1] Andy Oram, "Peer-to-Peer" O'Relly, March 2001.

[2] Dreamtech Software Team, "Peer-to-Peer Application Development", John Wiley & Sons, Nov. 2001.

[3] Badrul M. Sarwar, George Karypis, Joseph a. Konstan, John T. Riedle, "Application of Dimensionality Reduction in Recommender System-A Case Study", *ACM WebKDD 2000 Web Mining for E-Commerce Workshop*, 2000.

[4] Daniel Billנ, Michael J. Pazzani, "Learning Collaborative Information Filters", *Proceedings of ICML*, pp. 46-53, 1998

[5] 박지선, 김택헌, 류영석, 양성봉, "추천 시스템을 위한 2-way 협동적 필터링 방법을 이용한 예측 알고리즘", 한 국정보과학회 논문지 B, VOL.29, pp. 0669 ~ 0675, 2002.

[6] http://www.nararesearch.com/

[7] 양재영, 최충민, "협동 에이전트 시스템", 전자공학회지 26권 1호, pp 25-33, 1999