Pure P2P 환경에서 컨텍스트 정보에 기반을 둔 협력적 필터링

A Collaborative Filtering based Context Information in Pure P2P Environments

이세일, 이상용*
공주대학교 컴퓨터공학과, 공주대학교 컴퓨터공학부*
Se-Il Lee, Sang-Yong Lee
Dept. of Computer Engineering, Division of Computer Science & Engineering,
Kongju National University
E-mail : lsilhr@kongju.ac.kr

요 약

Pure P2P 환경에서는 축적된 자료를 사용하지 않고 실시간 정보를 사용하여 소수의 서비스 항목만으로도 협력적 필터링을 제공할 수 있어야 한다. 그러나 지역에서 수집된 소수의 서비스 항목만으로 협력적 필터링을 할 경우 추천 서비스의 질이 떨어지게 되므로, 사용자의 컨텍스트 정보를 이용하여 추천 서비스의 질을 높일 수 있는 방법이 연구되어야 한다. 하지만 사용자 컨텍스트 정보는 다량의 정보가 순간에 인식될 수 있기 때문에 확장성 문제(Scalability Problem)가 발생하고, 영역과 아이템에 따라 차별화된 서비스를 지원하기에는 한계성을 가지고 있다.

이러한 문제점을 해결하기 위하여 본 연구에서는 SOM을 이용하여 컨텍스트 정보를 서비스 영역별로 클러스터링(Clustering)하여, 사용자별로 분류함으로 확장성 문제를 해결하였다. 또한, 분류된 자료들 중 서비스 요구자와 비슷한 분류에 있는 사용자들의 컨텍스트 정보들을 정량화하여 협력적 필터링함으로 사용자에게 적합한 서비스를 지원할 수 있다.

1. 서론

유비쿼터스 환경에서 모바일이나 PDA를 이용하여 필요한 정보를 추천 받기 위해서는 서버에 저장되어 있는 서비스 이력 정보와 추천 시스템을 조합하여 원하는 정보를 얻게 된다[1]. 그러나 서버는 여러 사람들의 추천한 내용들로 인하여 막대한 메모리 공간이 필요하다. 이를 해결하는 방법 중에 하나는 pure P2P를 이용하여 실시간으로 정보를 얻어 원하는 정보를 추천하는 것이다. 그러나 이것 또한 다음과 같은 문제점이 발생한다. 서비스 요구자 자신의 서비스 항목들과 유사 사용자들의 서비스 항목들을 비교하여 추천을 해야 되지만, pure P2P에서는 축적된 정보가 사용되지 않으므로 단일의 서비스 정보만을 이용하여 유사 사용자와 비교하게 되면 정확하지 않은 결과를 초래할 수 있다. 이러한 문제점을 해결하는 방법은 서비스 정보 및 다른 컨텍스트 정보를 이용하여, 추천 시스템에서 가장 많이 사용하고 있는 협력적 필터링을 이용하면 추천의 질을 높일 수 있다[2]. 하지만 사용자들의 컨텍스트 정보는 영역과 아이템에 따라 차별화된 서비스를 지원하기에는 한계성이 있으며, 또한 일정한 지역에서 많은 정보 수집으로 인한 확장성 문제(scalability problem)도 발생하게 된다.

본 논문에서는 이러한 문제점을 해결하기 위하여 SOM(Self-Organizing Maps)을 이용하여 컨텍스트 정보를 서비스 영역별로 클러스터링(Clustering)하여 사용자별로 분류한다. 그리고 분류된 자료들 중 사용자와 비슷한 분류에 있는 컨텍스트 정보들을 정량화하여 협력적 필터링하게 된다. 그러므로 서비스 요구자와 비슷한 컨텍스트 정보만 분류되어 확장성 문제를 해결하였다. 또한, 단일의 서비스 항목만을 사용하는 것이 아니라 다른 컨텍스트 정보들을 정량화하여 협력적 필터링하게 되므로 서비스 요구자에게 적합한 서비스를 지원하게 된다.

본 논문의 구성은 2장에서 관련연구, 3장에서는 SOM을 이용한 컨텍스트 정보 분석, 4장에서는 실험 및 평가한 내용을 기술하였다. 그리고 마지막으로 5장에서는 결론에 대하여 언급한다.

2. 관련 연구

2.1 협력적 필터링

협력적 필터링(Collaborative Filtering)은 방대한 양의 정보 중 필요로 하는 정보만을 추천하기 위한 방법으로 가장 많이 사용되고 있다. 협력적 필터링에서는 사용자들이 좋아할 만한 항목을 예측하기 위하여 비슷한 선호도를 가지는 다른 고객들의 항목에 대한 평가에 근거하여 f=항목을 추천한다[3].

예를 들어 5명의 사용자와 4개의 항목에 대하여 각각 1~5점까지 점수를 부여한다고 하면, 표 1과 같은 테이블이 된다.

	항목1	항목2	항목3	항목4
사용자1	3	5	2	1
사용자2		2	4	5
사용자3	2	3	5	4
사용자4	5	5	3	
사용자5	?	5	1	2

표 1. 평가 항목

사용자5의 항목1을 추천 받기 위해서 관련성을 따져보면, 사용자5와 가장 관련이 있는 사용자1을 참조하는 것이 좋다는 것을 알 수 있다. 이러한 과정을 일반화한 식이 식(1)과 식(2)이며, 각각 선호도와 유사도를 나타낸다.

$$P_{x,b} = \bar{r}_x + \frac{\sum_{y=1}^{n} w(x,y)(r_{y,b} - \bar{r}_y)}{\sum_{y=1}^{n} w(x,y)} \quad (1)$$

$$W_{x,y} = \frac{\sum_{a=1}^{n}(r_{x,a} - \bar{r}_x)(r_{y,a} - \bar{r}_y)}{\sqrt{\sum_{a=1}^{n}(r_{x,a} - \bar{r}_x)^2}\sqrt{\sum_{a=1}^{n}(r_{y,a} - \bar{r}_y)^2}} \quad (2)$$

$P_{x,b}$는 사용자 x와 항목 b에 대한 선호도를 예측한 값이고, $\bar{r}_x$는 사용자 x의 선호도 평균값이다. $r_{y,b}$는 사용자 y가 항목 b에 대하여 평가한 값이며, n은 결정된 이웃의 수이다. w(x, y)는 사용자 x와 사용자 y의 유사도 가중치이다.

2.2 컨텍스트

유비쿼터스 컴퓨팅에서 언급되는 컨텍스트(Context) 정의는 사용자가 처한 환경에서 사용자의 현재 위치, 행동 및 작업 등의 사용자에 대한 정보와 그 정보들의 지속적인 변화를 말한다. 또한 Xerox PARC의 Schilt는 사용자와 오브젝트에 관련된 신원 및 오브젝트 정보로 정의하였고, GATECH의 Abowd는 사용자, 공간, 오브젝트 등의 개체와 관련된 모든 정보로 정의하였다[4].

2.3 Pure P2P

P2P란 Peer to Peer의 준말이며, 클라이언트 간 정송과 관리를 서버 없이, 또는 단일한 관리 서버를 통해서 클라이언트와 클라이언트가 직접 연결하는 분산 컴퓨팅 구조를 말한다. P2P 기술은 중앙에 관리 서버가 존재하는 hybrid P2P와 중앙 서버가 존재하지 않는 pure P2P로 나눌 수 있다. 이 pure P2P 방식의 최초의 시스템은 Gnutella이다. Gnutella는 회선중심의 물리적인 연결성을 바탕으로 하는 라우팅이 아니라, 라우터가 하는 행동을 노드 자신이 가지고 있고 경유하는 데이터를 노드 자신과 인접한 노드들에게 정보를 전달하는 논리적인 라우팅을 하는 것이다. pure P2P는 연결성과 확장 가능성이 높지만 관리적인 측면과 보안성이 단점으로 지적되고 있다[5].

3. SOM을 이용한 컨텍스트 정보 분석

Pure P2P 환경에서 컨텍스트 정보를 분석하고, 그 정보를 협력적 필터링에 적용하기 위한 순서를 보면 다음과 같다.

① 센서들로부터 컨텍스트 정보를 받아들여 통합시킨다.
② 필요한 컨텍스트 정보만을 분석하여 분류한다.
③ 분류된 컨텍스트 정보는 SOM을 이용하여 서비스 영역별로 분류된다.
④ 일정한 값을 만들기 위해 정량화한 후 협력적 필터링하여 사용자에게 서비스를 추천한다.

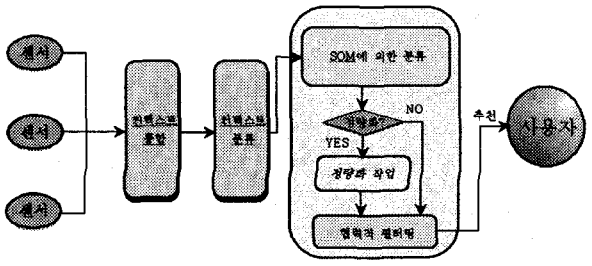

그림 1. 흐름도

3.1 컨텍스트 인식

서비스 요구자가 모바일이나 PDA를 이용하여 원하는 서비스를 받으려면 pure P2P 환경에서는 사용자의 실시간 컨텍스트 정보를 필요로 한다. 서비스 추천을 위해 먼저 컨텍스트 통합 에이전트(Integration Agent)는 자신의 프로파일 정보와 센서로부터 읽어들인 사용자 컨텍스트 정보 그리고 로컬 P2P에서는 기본적인 정보만을 통합한다. 통합된 컨텍스트 정보는 서비스에 필요한 컨텍스트 정보만을 분류한다. 필요한 컨텍스트 정보는 표 2와 같다.

분류	내용
사용자 ID	사용자 구분
성별	남/여
나이	나이(10단위로 분류)
장르	영화장르(클러스터링)
제목	영화제목(추천)
장소	서비스 장소
시간	서비스 시간
동행인	의도(친구, 애인,...)

표 2. 서비스에 필요한 컨텍스트 정보

이 분류된 컨텍스트 정보는 장르별로 클러스터링하

기 위해 SOM에 의한 분류 단계로 이동한다.

3.2 SOM에 의한 클러스터링 및 컨텍스트 정량화 단계

사용자들로부터 수집된 컨텍스트 정보가 많아지면 확장성의 문제가 발생하게 된다. 이 문제의 해결을 위해, 첫 번째 단계로 사용자에 직접 필요한 서비스 정보만을 위해 SOM을 이용하여 장르별로 클러스터링한다.

SOM은 Kohonen이 제안한 신경망 기반의 알고리즘이며, 입력층과 출력층 사이의 연결선은 연결강도를 나타내며 양층 사이의 완전 연결되어 있는 구조이다. 또한, 연결강도 벡터와 입력벡터가 통상 0에서 1사이의 정규화된 값을 사용하며, 연결강도 벡터와 입력 벡터의 거리가 가장 가까운 뉴런만이 출력을 낼 수 있다[6]. 장르별로 분류하기 위한 수행 알고리즘은 다음과 같다.

① 가중치를 초기화한다.(N : 입력, M : 출력)
② 새로운 입력을 제시한다.(장르별)
③ 입력과 출력간의 유사성을 계산한다.

$$s_b = \sum_{a=0}^{n-1}(R_a(t) - V_{(ab}(t))^2 \qquad (3)$$

$R_a(t)$는 시간 t에서 a번째 노드의 입력이고, $V_{ab}(t)$는 시간 t일 때, a부터 b까지 가중치이다.

④ 가장 유사성이 높은 노드 b를 선택한다.
⑤ 유사성이 높은 노드와 이웃 간의 가중치를 갱신한다.

$$V_{ab}(t+1) = V_{ab}(t) + \eta(t)(R_a(t) - V_{ab}(t)) \qquad (4)$$

$\eta(t)$는 학습률이다.

⑥ 2단계부터 다시 반복한다.

SOM을 이용한 분류는 그림 2와 같다.

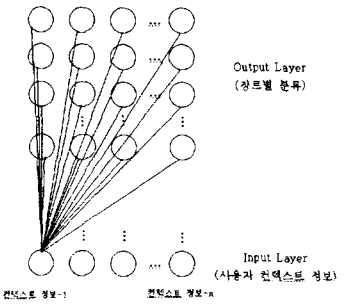

그림 2. SOM을 이용한 장르별 분류

모든 컨텍스트 정보는 바로 이용되지 않고, 분류된 후 사용하게 되므로 확장성 문제를 개선할 수 있다.

이 분류된 컨텍스트 정보를 직접 협력적 필터링에 사용하기 위해서는 값들의 정량화가 필요하다. 정량화에 사용되는 컨텍스트 정보는 나이, 성별, 동행인, 서비스 시간 등이고, 정량화된 값은 0.1과 1.0 사이의 값을 가진다. 컨텍스트 정보 중 나이를 정량화하기 위해서는 서비스 요구자와 비슷한 나이의 사용자에게는 높은 점수를 부여하며, 나이차가 많이 날수록 낮은 점수를 부여한다. 아래 식 (5)은 나이별로 정량화하는 식이며, 최저 0.2부터 최고 1.0까지 5단계로 값이 주어진다.

$$J = 1 - |S_{Age} - U_{Age}| * 0.2 \qquad (5)$$

S_{Age}는 서비스 요구자의 나이를 10으로 나눈 몫이고, U_{Age}는 비교되는 서비스 요구자의 나이를 10으로 나눈 몫이다. 또한, 성별의 구분은 자료 수집 결과 비슷한 취향을 보여 각각 0.8과 1.0의 값을 준다. 모바일의 그룹 정보를 이용하여 동행인과 함께 영화관에 왔는가에 따라 0.2부터 1.0으로 평가한다. 시간에 따라 영화 보는 취향이 조금씩 다르기 때문에 0.1부터 0.4까지 값을 준다. 0.1부터 0.4까지 주는 이유는 나이 차이나 동행인에 비하여 관련성 면에서 적은 영향을 받기 때문에 값의 차이를 적게 주었다.

3.3 컨텍스트 정보를 이용한 협력적 필터링

테이블의 구성이 표 1과 같으면 협력적 필터링을 이용하여 추천을 하는데 문제가 발생하지 않는다. 그러나 pure P2P에서는 축적된 자료가 없기 때문에 표 3과 같이 서로 관련성이 없는 문제가 발생하며, 또한 사용자들의 서비스에 대한 만족도를 숫자로 표현하기 어렵다.

항목 / 사용자	서비스1	서비스2	서비스3	서비스4
사용자1	형사	동막골	떠나라	사랑
사용자2	편지	별		
사용자3	쉬리			
서비스 요구자				

표 3. Pure P2P의 비교 테이블

그러나 컨텍스트를 이용한 협력적 필터링 방법에서는 서비스된 동류의 항목이 하나만 있어도 다른 컨텍스트 정보들과 함께 이용하여 서비스 요구자에게 추천 서비스를 할 수 있다. 표 4는 협력적 필터링을 하기 위한 컨텍스트 정보의 예이다.

항목 / 사용자	나이	성별	의도	Time	제목 (서비스)
사용자1	26	남	혼자	오후	형사
사용자2	59	남	가족	오후	금자씨
사용자3	26	여	혼자	저녁	외출
사용자4	31	남	애인	저녁	동막골
서비스 요구자	24	남	친구	저녁	?

표 4. 컨텍스트 정보

위와 같은 컨텍스트 정보를 가지고, 표 5와 같이 정량화한다.

항목 / 사용자	나이	성별	의도	Time	제목
사용자1	1.0	1.0	0.8	0.3	형사
사용자2	0.4	1.0	0.4	0.3	금자씨
사용자3	1.0	0.8	0.8	0.4	외출
사용자4	0.8	1.0	0.6	0.4	동막골
서비스 요구자	1.0	1.0	1.0	0.4	?

표 5. 정량화한 값

정량화된 값에 협력적 필터링을 적용하면 서비스 요구자와 가장 유사성이 높은 사용자1을 참조하여, 서비스 정보를 추천하게 된다. 따라서 서비스 요구자는 질적으로 향상된 서비스를 제공 받을 수 있다.

4. 평가

본 논문은 펜티엄 IV, 2.8Ghz, 512MB의 환경에서 C#과 J2ME, WIPI를 이용하여 설계하고 실험하였다.

컨텍스트 정보들 중 본 논문에 사용된 정보는 나이, 동행인, 시간, 성별이다(표 6).

나이	19이하	20~29	30~39	40~49	50이상
동행인	동료	친구	싱글	애인	가족
시간	오전	오후	저녁	심야	
성별	남	여			

표 6. 분류된 컨텍스트 정보

나이는 관련성이 많아서 정량화하기가 쉽지만 동행인, 시간 그리고 성별에 대한 관련성을 정량화하는 것은 쉽지 않기 때문에 427명에게 설문조사하여 관련성을 조사하였다. 설문조사한 인원은 남자 241명과 여자 186명 이었다. 동행인의 경우에는 가장 보고 싶은 장르를 선택하게 하여 상위 항목의 관련성을 비교하여 나온 결과이다. '싱글'을 기준으로 '동료'와 '친구' 쪽으로 이동할수록 액션과 코미디 부분에 더 가까웠으며, '애인'은 로맨스와 드라마 부분에 더 친밀한 것을 알 수 있었고, 가족은 가족드라마나 드라마 쪽에 가까움을 알 수 있었다.

동행인과의 경우보다 관계가 약하지만 시간도 관련성이 있다는 것을 알 수 있었다. 오전 쪽으로 가까울수록 액션과 코미디를 선호하였고, 심야 쪽으로 이동할수록 로맨스나 공포 쪽을 선호하는 것을 알 수 있었다.

평가 식으로는 예측의 정확성을 평가하기 위하여 MAE(Mean Absolute Error)를 사용하였다. v_i는 예측 선호도이며, r_i는 실제 선호도이고, N은 총 예측 회수이다.

$$|E| = \frac{\sum |v_i - r_i|}{N} \qquad (6)$$

실험 결과 컨텍스트 정보에 기반을 둔 SOM을 이용한 협력적 필터링 방법(CFS)은 기존의 GroupLen방식(CF)보다 월등하게 오차가 적음을 알 수 있었고, Naive Bayesian 알고리즘을 적용한 협력적 필터링 방식(CFN)[2]보다는 평균 0.0044정도 오차가 적음을 알 수 있었다.

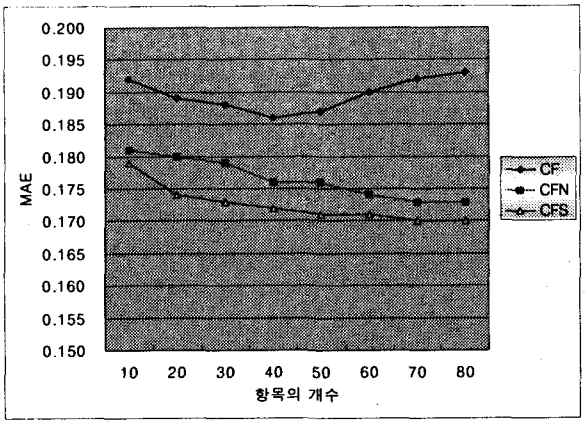

그림 3. 항목의 개수에 따른 MAE

5. 결론

Pure P2P 환경에서 소수의 항목만으로 협력적 필터링을 할 경우 서비스의 질이 떨어지는 문제점이 발생하게 되므로, 컨텍스트 정보를 이용하였다. 그러나 많은 정보 수집으로 인한 확장성 문제가 발생하므로, SOM을 사용하여 서비스 영역별로 클러스터링하여 확장성 문제를 어느 정도 개선하였다. 또한 사용자의 컨텍스트 정보는 영역과 아이템에 따라 차별화된 서비스를 지원하기에 힘든 한계성을 가지고 있으므로, 나이, 동행인, 시간, 성별등과 같은 컨텍스트 정보를 정량화하여 협력적 필터링함으로 서비스의 질적 향상을 가져올 수 있었다.

제안된 방법은 WIPI 에뮬레이터를 사용하여 실험한 결과 서비스 지원 성능 평가 면에서 CF보다는 9.9%, CFN보다는 2.3% 정도 우수함을 보였다.

참고자료

[1] 구미숙, 황정희, 최남규, 정두영, 류근호, "유비쿼터스 상거래 환경의 컨텍스트 기반 점진적 선호 분석 기법", 정보처리학회논문지, D 제11-D권 제7호, 2004.12

[2] 이세일, 이상용, "P2P 모바일 에이전트의 컨텍스트 정보를 이용한 협력적 필터링", 퍼지 및 지능 시스템 학회 논문지, vol.15, No.5, pp.643-648, 2005

[3] Sarwar, B. et al., "Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System", Proc. ACM CSCW 98, pp.345-345, 1998

[4] S.Jang, W.Woo, "ubi-UCAM: A Unified Context-Aware Application Model.", LNAI(Context03), pp.178-189, 2003

[5] 김용곤, 김동현, 이성주, "P2P 비즈니스 모델 분석에 관한 연구", 한국퍼지및지능시스템학회, 춘계학술대회 학술발표 논문집, pp.203-206, 2001

[6] T. Kohonen, Self-Organizing Maps, Springer-Verlag, Berlin, 1995