클러스터링 기반 사례기반추론을 이용한 추천시스템 개발
The Development of Recommender System Using Clustering-based CBR

이 희 정, 홍 태 호
부산대학교 경영학과
부산광역시 금정구 장전동 산30

초 록

웹의 급격한 확산과 더불어 고객에게 맞춤화된 정보 제공의 필요성이 높아지고 있다. 또한 전자상거래 기업은 맞춤화와 개인화 서비스를 실현하기 위해서 웹 기반의 추천시스템에 많은 관심을 가지고 있다. 협업필터링(Collaborative filtering)은 개인화된 정보필터링 기법으로 추천시스템에서 가장 많이 사용되고 있다. 본 연구에서는 MovieLens 데이터 셋의 아이템속성을 고려하여 클러스터링 기반의 사례기반추론을 통한 협업필터링 추천시스템을 개발하고 기존의 방법과 제안된 모델의 성과를 비교 분석하였다.

1. 서 론

최근 웹의 급격한 발전으로 인해 인터넷에서의 전자상거래가 더욱 활발하게 진행되고 있다. 인터넷을 통해 쇼핑을 할 뿐만 아니라, 영화나 음악 서비스도 받을 수 있게 되었다. 그러나 인터넷의 방대한 정보로 인해 정보 홍수에 빠진 사람들은 정보에 대해 까다로운 요구를 하고 있다. 이를 테면 개인화 (personalization) 또는 맞춤화(customization)된 정보를 제공 받기를 원하고 있다. 커스터마이징과 개인화 서비스는 인터넷 쇼핑몰이나 웹 서비스 제공자에게 있어 중요한 성공 요인이기 때문에 대중 맞춤(Mass customization)을 실현할 수 있는 웹 기반의 추천시스템(Recommender system)에 대한 관심이 더욱 높아지고 있다. 추천시스템은 인터넷을 통한 일대일 고객관리와 웹 개인화 서비스를 가능하게 한다. 추천시스템의 웹 개인화 기법으로 가장 널리 이용되고 있는 기법 중의 하나가 협업필터링(Collaborative filtering)이며 협업필터링과 추천 예측알고리즘에 관한 많은 연구들이 국내외에서 진행되어 왔다. 본 연구에서는 클러스터링 기반 사례기반추론을 통해 사용자간 연관성과 아이템간 연관성을 함께 고려한 협업필터링 기법을 제안하고 이에 대한 실증분석을 수행하였다. 기존의 협업필터링 기법과 제안된 기법의 성과를 제시하고 이를 통계적으로 검증하였다.

2. 문헌연구

2.1 추천시스템

추천시스템은 고객이 관심을 가지는 상품에 관한 정보, 인구통계학적 정보나 과거 구매 행동 분석을 토대로 고객의 요구에 맞는 상품을 추천해주는 시스템이다 (Sarwar et al., 2001). 또한 고객들이 구매하고자 하는 상품을 쉽게 찾을 수 있도록 도와주는 정보필터링 기술이기도 하다(Schafer et al., 1999). 추천시스템은 Amazon.com, CDnow.com 등 해외의 유수한 전자상거래 사이트에 적용되고 있으며, Ringo 음악 추천이나 Bellcore 비디오 추천에도 이용되고 있다. 전자상거래에서는 고객 개개인에게 맞춤화된 개인화 서비스가 강조되고 있는 데(Schafer et al., 2000), 추천시스템을 위한 개인화 기법에는 다음과 같은 세 가지가 있다(Kuo & Chen, 2001).

1) 규칙기반 필터링(Rule-based filtering)
인구통계학적 정보나 개인신상 정보를 사용자에게 질문하여 그 응답에 맞는 규칙을 기반으로 하여 추천하는 개인화 기법이다.

2) 협업필터링(Collaborative filtering)
고객이 선호하는 패턴과 유사한 다른 고객들의 선호도를 이용하여 고객에게 관련된 서비스를 추천하는 개인화 기법이다.

3) 학습 에이전트(Learning agent)
웹사이트 방문기록 및 횟수, 접속장소, 시간 등 일종의 로그파일 분석을 통해 사용자의 속성, 습관, 개인의 선호를 추적하는 학습 에이전트를 이용하는 개인화 기법이다.

2.2 협업 필터링

유사 선호도를 가진 다른 고객들이 평가한 값을 기반으로 하는 협업 필터링은 가장 성공적인 정보필터링 기법으로 웹 기반의 추천시스템에서 널리 이용되고 있다. 협업 필터링은 크게

사용자기반 협업필터링(user-based collaborative filtering)과 아이템기반 협업필터링(item-based collaborative filtering)으로 구분된다(Herlocker et al., 1999). 사용자기반 협업필터링은 사용자 간의 유사성을 측정하여 선호도가 비슷한 다른 고객들이 평가한 상품을 기반으로 어떤 특정 고객이 선호할 만한 상품을 추천하는 방식이다. 아이템 기반 협업 필터링은 아이템 간의 유사성, 즉 기존의 상품들과 추천하고자 하는 상품들간의 유사성을 측정하여 어떤 특정 고객이 어떤 상품을 선호하는 지를 예측하여 추천하는 방식이다 협업필터링의 핵심은 추천하고자 하는 고객과 유사한 고객들을 찾고, 선호도가 유사한 고객군에서 높이 평가한 아이템을 추천하는 것이다. 협업필터링에서 특정 고객과 선호도가 비슷한 이웃들을 선정하는 기법에는 클러스터링, 최근접이웃, 베이지안 네트워크 등이 있다.

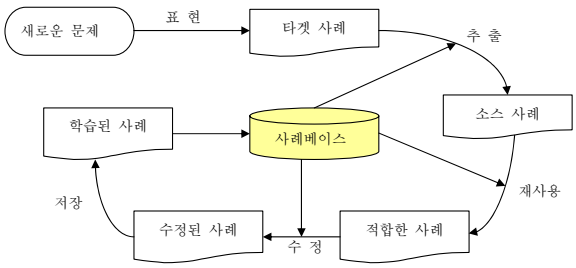
2.3 클러스터링

클러스터 기법은 고객들의 유사성을 찾아 고객세분화, 패턴인식, 추세분석 등에 활용되는 분석기법이다. 클러스터 기법은 다양한 특성을 지닌 대상들을 동질적인 집단으로 분류하는 데 이용되는 기법으로 대상들을 분류하기 위한 명확한 분류기준이 존재하지 않거나 알 수 없는 상태에서의 분류를 위하여 유용하게 이용된다.
K-means 클러스터링 알고리즘은 한 번의 군집이 묶일 때마다 각 군집별로 그 군집의 평균을 중심으로 군집 내 대상들간의 유클리디안 거리에 기반한다. 사용자의 선호도를 다차원 공간상의 점으로 표시하고, 이것의 거리를 계산함으로써 전체 사용자들의 집합을 k개의 군집으로 나누어지게 된다. 만약 사용자 a_i와 군집 k_j 사이의 거리는 아래와 같은 식으로 표현된다.

$$d_{a,k} = \sqrt{\sum_i (a_i - k_i)^2} \qquad (1)$$

2.4 사례기반추론

사례기반추론(Case-based reasoning)이란 새로운 요구에 대응하는 과거의 해답을 채택하거나, 과거의 사례를 이용하여 새로운 상황을 설명하거나, 과거의 사례로 새 해답을 평가하거나, 또는 새로운 상황을 이해하기 위해서나 새로운 문제에 대한 적당한 해답을 만들기 위해 선례로부터 추정하는 것을 의미한다(Kolondner, 1993).
사례기반추론의 추론과정은 다음과 같다. 먼저 새로운 문제가 주어졌을 때 사례집합에서 일정한 유사성의 척도에 부합하는 과거의 사례를 추출한 후, 추출된 사례를 재사용하여 해결에 이용한다. 만약 추출된 사례에 의한 해답이 새로운 문제 해결에 적합하지 않으면 이를 수정하여 새로운 해답을 제시하고 이를 다시 새로운 사례로 저장한다. 사례기반추론의 순서도는 <그림 1>과 같다(Aamodt&Plaza,1994).



<그림 1> 사례기반추론의 순서도

주어진 문제와 해결하고자 하는 문제 사이의 유사성 척도를 계산하는 가장 일반적으로 사용되는 방법 중의 하나가 최근접이웃 추출방법(Nearest-neighbor retrieval)이다. 유사도가 높은 순서대로 사례들이 추출되는 최근접이웃 추출방법은 다양한 유사도 지수를 이용하는데, 일반적으로 다음과 같은 공식에 의해 계산된다(Kolodner, 1993).

$$\frac{\sum_{i=1}^{n} W_i \times sim\,(f_i^I, f_i^R)}{\sum_{i=1}^{n} W_i} \qquad (2)$$

W_i= i번째 feature의 가중치
f_i^I= 입력 사례의 i번째 feature의 값
f_i^R= 추출 사례의 i번째 feature의 값
$Sim()$ = f_i^I 와 f_i^R의 유사도 함수

3. 추천시스템의 개발

지금까지 협업필터링을 이용한 추천시스템에 관한 많은 연구들이 진행되어 왔다. 협업필터링을 이용한 추천시스템에 관한 연구에 최근접이웃법과 클러스터링이 적용되어 왔다. Kim 등(2002)은 K-means 클러스터링 기법을 이용해 추천알고리즘에 적용하였고, Li & Kim (2000)은 클러스터링 기법을 아이템 기반 협업필터링에 응용하였다. 또한 Roh 등(2003)의 연구에서는 군집분석을 위해 SOM을 사용하고 유사도를 찾기 위해 최근접이웃법을 이용하여 기존의 협업필터링 방법과 비교 분석하여 예측 성능의 우수함을 보였다. 하지만 아이템의 속성을 활용하여 아이템속성 간에 클러스터링하고 이 클러스터안에서 사용자별 클러스터를 통한 추천방법에 관한 연구는 없었다. 기존의

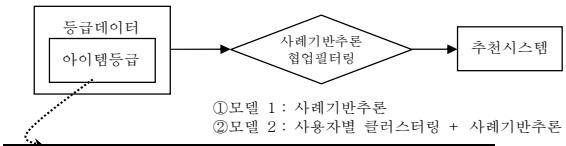
연구에서는 어떤 아이템이 여러 개의 속성을 가지고 있음에도 불구하고 여러 속성을 가진 아이템을 오직 하나의 속성만을 부여해 클러스터링 하였다. 하나의 특정 속성으로만 정의(define)하게 되면 실제의 다중속성의 성격이 희석되어 결국에는 추천 예측력이 떨어지게 된다. 본 연구에서는 협업필터링을 위해 기존의 방법인 사용자 기반 사례기반추론과 사용자 클러스터링에 기반한 사례기반추론을 적용하였다. 또한 추천시스템 개발을 위해 사용자와 아이템 클러스터링 기반의 사례기반추론을 새롭게 제안한다.

1)CBR_CF: 사용자기반 사례기반추론

각 아이템에 등급을 메긴 여러 사용자들을 기반으로 사례기반 추론을 이용하여 추천한다.

2)UC_CBR_CF: 사용자 클러스터링 기반 사례기반추론

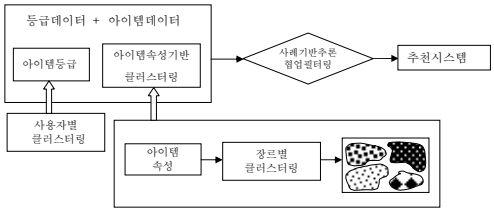
클러스터링을 통해 사용자의 소속그룹을 결정하고, 그에 따라 소속그룹의 다른 사용자들이 선호하는 아이템을 사례기반추론을 이용하여 추천한다.



사용자	아이템1	아이템2	…	아이템n
34	5	4	2	2
67	4	3	5	4
98	3	4	2	1
125	3	4	3	2

<그림2> CBR_CF, UC_CBR_CF

3)제안된 UC_IC_CBR_CF: 사용자와 아이템속성 클러스터링 기반 사례기반추론



<그림3> 제안된 UC_IC_CBR_CF모델

CBR_CF와 UC_CBR_CF가 사용자 간의 유사성을 고려한 반면에, 제안된 UC_CBR_CF 에서는 아이템 간의 유사성도 같이 고려하는 방식이다. 즉, 유사한 아이템끼리 묶기 위해 아이템의 속성 기반으로 클러스터링 한다. 클러스터링 한 다음 어떤 사용자가 어떤 종류의 아이템을 더 선호하는지 그 소속그룹을 정하고 정해진 그룹 내에서 사례기반추론을 통해 유사한 사용자들이 선호하는 아이템을 추천해주는 추천시스템 개발 방법론이다.

4. 실증분석

4.1 데이터

본 연구에 사용된 데이터는 GroupLens에서 공개되어진 MovieLens의 웹 기반 추천시스템의 영화 데이터 셋을 이용하여 실험을 수행하였다. 이 데이터 셋은 총 943명의 사용자가 1,628편의 영화에 대해 1~5점 척도 점수로 부여된 총 100,000개로 구성되어 있다. 각 사용자들이 평가한 영화는 최소 20편이고 영화의 장르는 액션, 어드벤처, 애니메이션, 어린이, 코메디, 범죄, 다큐멘터리, 드라마, 판타지, 공포, 뮤지컬, 필름누와르, 미스터리, 로맨스, 공상과학, 스릴러, 전쟁, 웨스턴으로 총 18가지로 구분되어 있다.

4.2 실험결과 및 분석

영화 데이터를 이용해 아이템의 속성 즉, 장르속성을 기반으로 K-means 클러스터링을 수행한다. 18개 장르의 총 1,628편 영화를 사전 군집 수 4개로 지정하여 클러스터링 수행결과는 <표1>과 같다.

<표1>장르 클러스터링 수행 결과

군집1	액션 성격의 스케일이 큰 영화
군집2	범죄적인 요소가 있는 오싹하고 무서운 영화
군집3	로맨틱, 서정적인 드라마 영화
군집4	신나게 웃을 수 있는 코믹한 영화

장르별 클러스터링이 필요한 이유는, 각 영화는 여러 개의 장르 속성(예를 들어,“Toy Story”의 장르는 어린이, 코메디, 에니메이션)을 지니고 있었기 때문에 단지 하나의 장르로 단정짓기에는 애매모호한 점이 많다. 사용자별 클러스터링을 위해 각 그룹마다 5편의 영화를 선택하여 총 20편의 영화를 선정하고 이 영화들에 대해 등급을 메긴 131명의 사용자들을 클러스터링 시켜 소속군집을 정한다.

사례기반추론에서 사용되는 이웃의 수를 5로 정하고 유사한 사례를 추출한다. 각 모델마다 성과를 보면 <표2>와 같다.

<표2> 각 모델별 성과비교

모 델	RMSE[1]	MAE[2]	MAPE(%)[3]
CBR_CF	0.683	0.493	12.90 (%)
UC_CBR_CF	0.561	0.429	10.78 (%)
UC_IC_CBR_CF	0.500	0.350	9.16 (%)

1)RMSE(root mean squared error) 2)MAE(mean absolute error) 3)MAPE(mean absolute percent error)

UC_IC_CBR_CF의 MAE는 0.350, CBR_CF는 0.493, UC_CBR_CF는 0.429로 제안된 모델이 우수함을 보이고 있다. 이에 대한 통계적 유의성을 검증하기 위하여 대응표본 *t*검정(Paired samples *t*-test)을 수행하였다. 그 결과 CBR_CF와 UC_IC_CBR_CF간의 *t*값을 보면 유의수준 5%수준에서 유의하게 나타났고, UC_CBR_CF와 IC_CBR_CF간의 *t*값도 유의수준 5%수준에서 유의하게 나타났다. 따라서 제안된 모델이 기존의 다른 모델보다 우수함을 보였다.

<표3> 각 모델간의 대응표본검정

	UC_CBR_CF	UC_IC_CBR_CF
CBR_CF	*t*값=0.806	*t*값=1.875*
UC_CBR_CF		*t*값=1.784*

* 유의수준 5%에서 유의함.

5. 결론 및 향후연구

추천시스템과 협업필터링에 관한 선행연구의 대부분은 높은 예측 성능을 지닌 예측 알고리즘에 초점을 두었다. 본 연구에서는 아이템의 속성을 활용하여 클러스터링 기반 사례기반추론 방법을 제시하였다.

K-means 클러스터링 기법을 통해 장르 특성별로 군집을 구분하고 사용자의 소속그룹 내에서 사례기반추론을 통해 유사 선호도를 가진 이웃들의 평가를 토대로 영화를 추천하는 방법이 기존의 방법보다 실증분석 결과 통계적으로 우수한 성과를 보였다. 영화나 상품을 추천할 때 사용자간의 유사성 뿐만 아니라 아이템간의 유사성, 즉 아이템의 속성을 함께 고려하는 것이 좀 더 정확하게 예측하여 추천할 수 있을 것이다. 본 연구에서는 아이템의 속성 중 장르속성 하나만을 가지고 클러스터링 하였으나, 향후 연구에서는 좀 더 다양한 아이템의 속성을 활용하여 클러스터링 기반의 사례기반추론을 이용한 추천시스템의 연구가 필요하다고 하겠다.

6. 참 고 문 헌

[1] Aamodt, A., Plaza, E., "Case-based reasoning:Foundational issues, methodological variations, and system approaches," *Artificial Intelligence Communication*, 7(1), 1994, 39-59.

[2] Herlocker, J., Konstan, J., Borchers, A. and Riedl, J., "An Algorithmic Framework for Performing Collaborative Filtering," *Proceedings of the 1999 Conference on Research and Development in Information Retrieval*. Aug. 1999.

[3] Kim, T.H., Ryu, Y.S., Park, S.I. and Yang, S.B.,"An Improved Recommendation Algorithm in Collaborative Filtering," *Lecture notes in computer science,* no.2455, 2002, pp.254-261.

[4] Kolodner, J. L., "Case-Based Reasoning," Los Altos, CA: Morgan Kaufmann, 1993.

[5] Kuo, Y.F., Chen, L.S., "Personalization technology application to Internet content provider," *Expert Systems with Applications*,v.21 no4, 2001, pp.203-215.

[6] Li, Q., Kim, B.M., "Clustering Approach for Hybrid Recommender System, " *Proceedings of IEEE/WIC International Conference on Web Intelligence,* 2003, pp. 33-39.

[7] Roh, T.H., Oh, K.J., Han, I.G., "The collaborative filtering recommendation based on SOM cluster-indexing CBR," *Expert Systems with Applications,* v.25 no3, 2003, pp.413-423.

[8] Sarwar, B. M., Karypis, G., Konstan, J. A. and Riedl, J.,"Item-Based Collaborative Filtering Recommendation Algorithms," *Proceedings of the 10th International World Wide Web Conference,* 2001, *pp.285-295.*

[9] Schafer, J.B., Konstan, J., Riedl, J., "Recommender Systems in E-Commerce," *Proceedings of the ACM Conference on Electronic Commerce*, November 3-5, 1999.

[10] GroupLens, http://www.grouplens.org.