한국지능정보시스템학회논문지 제13권 제4호 2007년 12월(pp.65~78)

장르별 협업필터링을 이용한 영화 추천 시스템의 성능 향상*

이재식
아주대학교 e-비즈니스학부
(leejsk@ajou.ac.kr)

박석두
유비쿼터스 컨버전스 연구소
(epitus@hanmail.net)

추천시스템은 개인화 서비스를 구현하는 방법 중의 하나이다. 추천시스템은 다양한 기법을 통해 구축될 수 있는데, 최근 전자상거래 분야에서 사용되는 기법들 중에서 대표적인 것이 협업필터링이다. 협업필터링은 영화나 음악 같이 명시적인 속성만으로 그 특성을 기술하는데 한계가 있는 아이템의 추천문제에 효과적으로 적용되어 왔다. 하지만, 이 기법은 희박성, 확장성 및 투명성 등의 문제점을 가지고 있는데, 본 연구에서는 희박성과 확장성 문제를 극복하는 방안으로 장르별 협업필터링 방법을 제안한다. 장르별 협업필터링 방법은 아이템을 최종적으로 추천하기 전에 아이템의 상위 카테고리, 즉 장르에 대한 정보를 활용하는 방법이다. 본 연구에서 제안하는 방법의 실용성을 보이기 위하여, 영화 추천시스템인 GenreWise_CF를 개발하여, 공개 데이터인 MovieLens Data에 적용하여 평가하였다. 실험 결과, 본 연구에서 제안한 GenreWise_CF가 전통적인 협업 필터링을 적용하여 개발한 추천시스템인 Basic_CF보다 향상된 성능을 보였다.

논문접수일 : 2007년 07월 게재확정일 : 2007년 11월 교신저자 : 이재식

1. 서론

최근 인터넷의 사용이 보편화되면서 서비스 제공자들이 고객 정보를 실시간으로 모니터링하는 것이 가능해지면서 다양한 개인화 서비스가 제공되고 있다. 개인화 서비스란 고객들이 필요로 하는 제품이나 서비스를 명시적으로 묻지 않고 제공하는 것을 뜻한다(Mulvenna et al. 2000). 고객맞춤화나 개인화 서비스는 인터넷 서비스 제공자들의 중요한 성공요인으로 인식되고 있다(Ansari et al., 2000). 개인화 서비스는 개인에 대한 정보를 기반으로 서비스를 제공하기 때문에 서비스 제공자와 개인간의 정보교류가 원활할 때 효과적으로 이루어진다. 최근 인터넷의 사용이 보편화되면서 서비스 제공자들이 고객 정보를 실시간으로 모니터링하는 것이 가능해지면서 다양한 개인화 서비스가 제공되고 있다.

* 본 연구는 21세기 프론티어 연구개발 사업의 일환으로 추진되고 있는 정보통신부의 유비쿼터스컴퓨팅 및 네트워크 원천기반기술 개발사업의 지원에 의한 것임.

개인화 서비스 중에서 추천시스템(Recommendation System)은 목표고객에게 그가 좋아할 만한 서비스나 아이템을 추천해주는 서비스로서 Amazon이나 CD Now 등 인터넷 쇼핑몰에서 많이 사용되고 있다. 추천시스템은 다양한 기법을 통해 구현될 수 있는데 최근 전자상거래 분야에서 쓰이는 기법 중에서 대표적인 것이 협업 필터링(Collaborative Filtering)이다.

협업 필터링은 고객들의 프로파일정보를 활용하여 목표고객이 높게 평가할 것으로 예상되는 서비스나 아이템을 추천하는 기법으로 다음과 같은 과정으로 아이템을 추천한다. 먼저, 아이템들에 대한 고객의 평가치를 직간접적으로 수집하여 고객별 프로파일을 생성한다. 생성된 프로파일을 기반으로 목표고객과 유사한 성향을 보이는 고객들로 최근접 이웃을 구성한 후 최근접이웃의 평가치를 이용하여 목표고객이 평가하지 않은 아이템의 평가치를 예측한다. 이렇게 예측된 평가치를 기반으로 목표고객이 높게 평가할 것이라고 예상되는 서비스나 아이템을 추천한다. 기존의 추천시스템의 기법들은 아이템간의 연관성을 파악할 때, 아이템의 속성을 사용하였다. 하지만, 음악이나 동영상 같은 무형의 아이템들은 제목, 제작자 등의 명시적인 속성만으로 그 특성을 기술하는데 한계가 있다. 협업 필터링은 이러한 아이템들을 대상으로도 적용할 수 있다는 장점이 있다.

하지만 고객의 프로파일 정보를 기반으로 추천을 하기 때문에 유사한 성향을 보이는 최근접이웃을 찾고, 아이템에 대한 평가치를 예측하기 위해서는 고객 및 구매 데이터가 충분히 축적되어야 한다. 또한 희박성(Sparsity), 확장성(Scalability) 그리고 투명성(Transparency) 등의 한계점이 지적되었다(Sarwar, 2001 ; Li and Yamada, 2004 ; Yang et al., 2004]. 이러한 한계점들에 대해서는 제 2.3절에서 자세하게 기술한다. 그리고 협업 필터링은 고객들이 경험하고 평가한 아이템을 대상으로 추천을 하기 때문에 신제품이 추천되기가 어렵다는 지적이 제시되었다(Canny, 2002).

본 연구에서는 영화를 대상으로 추천시스템을 개발하였는데, 최종적으로 개별 영화를 추천하기 전에 영화 장르의 점수를 계산하여 이것을 활용하여 장르별로 추천을 수행함으로써 협업 필터링의 한계점에 대처하고, 또한 추천의 성능을 높였다. 본 연구는 다음과 같이 구성되었다. 제 2장에서는 협업 필터링에 대해서 기술한다. 제 3장에서는 본 연구에서 사용한 데이터인 영화 데이터의 설명 및 준비 과정을 설명하고, 전통적인 협업 필터링을 사용한 추천시스템에 대해서 기술한다. 제 4장에서는 본 연구에서 제안하는 장르별 협업 필터링을 사용한 추천시스템의 구축 및 성능에 대해서 기술하고, 제 5장에서 결론을 맺는다.

2. 협업 필터링

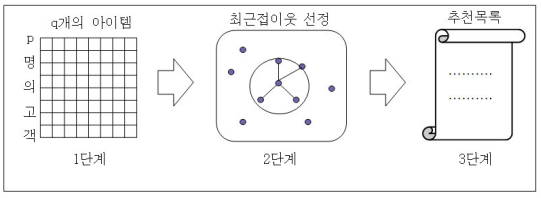
추천시스템 연구의 주된 관심사는 고객과 아이템에 대한 이용 가능한 정보를 분석하여 고객들이 관심을 가질 아이템이 무엇인지를 파악하는 것이다. 추천시스템은 보통 아이템의 속성들을 기반으로 목표고객이 관심을 가진 아이템과 비슷한 속성의 아이템을 추천하거나, 고객들이 아이템들을 경험하고 부여한 평가치들을 기반으로 목표고객이 평가하지 않은 아이템들 중에서 높게 평가할 것이라고 예상되는 아이템을 추천한다. 평가치는 고객들의 행동들을 주시하거나(Breese et al., 1998), 고객들에게 경험한 아이템에 대한 평가치를 묻는 방법을 통해서 얻어진다(Lekakos and Giaglis, 2006).

협업 필터링은 아이템에 대한 목표고객의 평가치와 다른 고객의 평가치를 이용하여 목표고객이 좋

아할 만한 아이템을 추천하는 기법이다(Resnick et al., 1994 ; Lekakos and Giaglis, 2006). 협업 필터링은 학계 및 산업계에서 널리 연구 및 적용되고 있다. 기사 추천시스템인 GroupLens(Resnick et al., 1994), 비디오 추천시스템인 Video Recommender (Hill et al., 1995), 음악 추천시스템인 Ringo(Shardanand and Maes, 1995) 및 World Wide Web에서 사용자와 관련된 정보를 찾아주는 PHOAKS(Terveen et al., 1997) 등이 개발되었으며, Amazon, CDNow, Drugstore, MovieFinder 등에서 협업 필터링 기법을 사용하여 상품을 추천하고 있다.

2.1 협업 필터링의 과정

협업 필터링기법으로 아이템을 추천하는 과정은 [그림 1]과 같이 크게 세 단계로 나뉘어진다.



[그림 1] 협업 필터링의 3단계

제 1단계 : 평가치 매트릭스의 준비

p명의 고객들이 q개의 아이템을 경험하고 부여한 평가치들을 정리하여 <표 1>과 같이 p × q의 고객 × 아이템 매트릭스를 만든다.

<표 1> 협업 필터링에서 고객 × 아이템 매트릭스

아이템 / 고객	아이템 1	아이템 2	아이템 3	아이템 4
고객 A	$R_{A,1}$	ϕ	$R_{A,3}$	$R_{A,4}$
고객 B	$R_{B,1}$	$R_{B,2}$	ϕ	$R_{B,4}$
고객 C	$R_{C,1}$	ϕ	$R_{C,3}$	ϕ

여기서 $R_{A,1}$는 고객 A가 아이템 1에 부여한 평가치를 뜻한다. Φ는 고객이 아이템을 평가하지 않았음을 뜻한다.

제 2단계 : 최근접이웃의 구성

고객들의 평가치를 이용하여 고객들간의 유사도를 계산한 후 최근접이웃을 구성한다. 유사도를 계산하기 위한 측정지수로는 일반적으로 Pearson Correlation Coefficient와 Cosine이 사용된다.

Pearson Correlation Coefficient $w(A,B)$는 두 고객 A, B에 의해 공통적으로 평가된 아이템들의 평가치를 이용하여 식 1과 같이 계산한다(Resnick et al. 1994).

$$w(A,B) = \frac{\sum_{i=1}^{q}(R_{A,i} - \overline{R_A})(R_{B,i} - \overline{R_B})}{\sqrt{\sum_{i=1}^{q}(R_{A,i} - \overline{R_A})^2 \sum_{i=1}^{q}(R_{B,i} - \overline{R_B})^2}} \quad (1)$$

여기서 $R_{A,i}$와 $R_{B,i}$는 고객 A와 B가 공통으로 평가한 아이템 i의 평가치를 뜻한다. 그리고 $\overline{R}_A$, $\overline{R}_B$는 고객 A와 B의 이용 가능한 평가치들의 평균값을 뜻한다.

Cosine $\cos(\vec{A}, \vec{B})$은 고객 A와 B가 공통으로 평가한 아이템들의 평가치를 q차원 공간에 벡터화한 후 두 벡터 사이 각의 Cosine값을 식 (2)와 같이 계산한다.

$$\cos(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\| \vec{A} \|^2 \times \| \vec{B} \|^2} \quad (2)$$

Herlocker et al.[1999]은 고객들간의 유사도를 계산할 때, Pearson Correlation Coefficient를 사용하는 것이 Cosine을 사용하는 것보다 높은 추천 성과를 보인다고 발표하였다. 따라서 본 연구에서는 최

근접이웃을 구성하기 위해 Pearson Correlation Coefficient을 사용하였다.

목표고객에 대한 다른 고객들의 유사도를 계산한 후 목표고객의 최근접이웃을 다음의 두 가지 방법으로 구성할 수 있다. 첫 번째는 Threshold based Selection(Shardanand and Maes, 1995)으로서 목표고객과의 유사도가 미리 설정한 Threshold값 이상인 고객들로 최근접이웃을 구성하는 것이고, 두 번째는 k-NN(k-Nearest Neighbors) 기법으로서 목표고객과의 유사도가 높은 상위 k명의 고객들로 최근접이웃을 구성하는 것이다(Resnick et al., 1994).

제 3단계 : 추천목록의 생성

최근접이웃의 평가치를 이용하여 목표고객이 평가하지 않은 아이템의 평가치를 예측한 후 추천목록을 생성한다. 평가치 $R_{A,i}$는 고객 A의 아이템 i에 대한 평가치로 고객 A의 최근접이웃의 평가치들을 가중 평균하여 식 (3)과 같이 예측한다.

$$R_{A,i} = \overline{R}_A + \frac{\sum_{j=1}^{k} w(A,j)(R_{j,i} - \overline{R}_j)}{\sum_{j=1}^{k} |w(A,J)|} \quad (3)$$

여기서 $\overline{R}_j$는 고객 A의 최근접이웃인 고객 j의 이용 가능한 평가치들의 평균값을 뜻한다.

목표고객이 평가하지 않은 아이템들의 평가치를 예측한 후 Top-N 기법으로 추천목록을 생성한다. 즉, 목표고객의 예측된 평가치 중에서 수치가 높은 상위 N개의 아이템을 목표고객에 대한 추천목록으로 생성한다.

2.2 협업 필터링의 성능 측정

추천시스템의 성능을 측정하는 방법은 여러 가지가 있다(Sarwar et al., 2000). 최근접이웃의 질과 같이 중간단계를 평가하는 것도 가능하나 본 연구에서는 추천시스템의 결과에 관심이 있기 때문에 예측 평가 지표와 추천 평가 지표만을 고려하였다.

2.2.1 예측 평가 지표

고객이 실제로 부여한 평가치와 추천 알고리즘에 의해 예측된 평가치의 차이로 추천시스템의 성능을 측정한다. 측정지표로는 MAE(Mean Absolute Error)가 사용된다. MAE는 실제 고객의 평가치와 추천 알고리즘에 의해 예측된 평가치와의 차이를 나타내는 지표로 식 (4)와 같이 계산된다.

$$MAE = \frac{\sum_{i=1}^{q} |\text{실제고객평가치 } i - \text{예측된 평가치 } i|}{q} \quad (4)$$

MAE 지표는 예측된 평가치들이 실제 고객의 평가치들과 평균적으로 얼마나 흡사하냐를 나타내는 지표로서, MAE가 낮게 나오는 추천시스템을 우수한 성능의 추천시스템으로 평가한다. 하지만, MAE가 낮다는 것이 추천의 적중률이 높다는 것을 보장해주지는 않는다는 점에서 MAE 지표를 이용한 추천시스템의 성능 평가에 문제점이 있다.

추천시스템의 목적은 고객들에게 아이템을 선정하여 추천하는 것이다. 즉, 여러 개의 아이템들 중에서 예측된 평가치가 높은 상위 1개, 5개 또는 10개를 선정하여 추천하는 것이다. 이것들 중에 실제로 고객이 구매한 아이템이 속해 있으면 추천은 적중한 것이다. 하지만 MAE 지표를 이용한 평가에서는 추천시스템을 이와 같은 방법으로 평가하지 않는다. 그러므로 MAE 지표가 아무리 낮게 나오는 추천시스템이라고 해도 상위로 예측된 평가치의 아이템들이 고객이 실제로 높게 평가한 아이템들과 일치한다는 보장이 없다. 예를 들어, 고객이 실제로

가장 높게 평가한 상위 5개의 아이템이 추천시스템에 의해서는 가장 낮은 하위 5개의 아이템으로 예측되어 결국 추천이 안되어도 전체적으로 MAE만 낮게 나오면 그 추천시스템은 우수한 추천시스템으로 평가된다. 이는, MAE가 단지 예측된 평가치들과 실제 고객의 평가치들이 평균적으로 흡사하다는 것을 나타내는 것이지, 각 아이템별로 평가치를 비교하는 것은 아니기 때문이다.

2.2.2 추천 평가 지표

Top-N 기법에 의해 생성된 추천목록의 평가에는 정보검색분야에서 사용되는 두 가지 측정지표인 Recall과 Precision이 사용된다. T집합을 '고객이 실제로 경험하여 평가한 아이템들 중에서 추천시스템의 성능측정을 위해 따로 모아둔 아이템들의 집합'이라고 하고, R집합을 '추천시스템에 의해 생성된 추천목록의 집합'이라고 하자. 한 아이템은 T집합 또는 R집합 또는 두 집합 모두에 속할 수 있다.

Recall은 목표고객이 실제로 경험하여 평가한 아이템들 중에서 추천시스템이 생성한 추천목록에 속하게 된 아이템의 비율로서 식 (5)와 같이 계산된다.

$$Recall = \frac{|T \cap R|}{|T|} \tag{5}$$

Precision은 추천시스템이 생성한 추천목록의 아이템들 중에서 목표고객이 실제로 경험하여 평가한 아이템의 비율로서 식 (6)과 같이 계산된다.

$$\mathrm{Precision} = \frac{|T \cap R|}{|R|} \tag{6}$$

일반적으로 Recall과 Precision은 추천목록의 개수에 따라 상충관계에 있다. 즉, 추천목록의 개수를 증가시키면, Recall은 분자가 증가하여 그 값이 커지게 되고 Precision은 분모가 증가하여 그 값이 작아지게 된다. 이러한 상충관계 때문에 두 성능지표를 동시에 고려하는 F1이 식 (7)과 같이 계산되어 사용된다.

$$F1 = \frac{2 \times Recall \times \mathrm{Precision}}{Recall + \mathrm{Precision}} \tag{7}$$

2.3 협업 필터링의 과제들

전통적인 협업 필터링은 희박성(Sparsity), 확장성(Scalability) 그리고 투명성(Transparency) 등의 한계점을 가지고 있다(Sarwar, 2001 ; Li and Yamada, 2004 ; Yang et al., 2004 ; Adomavicius and Tuzhilin, 2005).

희박성(Sparsity)은 고객이 아이템에 부여한 평가치의 개수가 부족하여 추천의 성과가 떨어지는 문제점이다. 취급하는 아이템의 종류가 많은 Amazon이나 e-Bay 같은 대형 인터넷 쇼핑몰에서는 고객들이 실제로 경험하고 평가하는 아이템들의 개수가 전체의 1%도 안될 뿐만 아니라 평가를 하지 않는 고객들도 많다. 즉, 협업 필터링 과정 중 제 1단계에서 준비되어야 하는 고객 × 아이템 매트릭스의 대부분이 비어 있게 된다. 그러므로 협업필터링이 제대로 작동할 수 없게 되는 것이다.

희박성의 문제를 완화하기 많은 연구들이 진행되고 있다. Breese(1998)는 평가치가 비어 있는 칸에 Default 평가치를 부여하는 방법인 Default Voting을 사용하였고, Sarwar et al.(2001)은 고객들 간이 아닌 아이템들 간의 유사도를 계산하여 추천에 활용하는 방법을 제안하였다. 다른 방법으로는 차원감소법이 있다. 이는 직접적으로 고객 × 아이템 매트릭스의 차원을 감소시키는 방법으로서 고객이나

아이템을 군집화하여 군집화된 그룹을 기본단위로 하여 협업 필터링 기법을 적용하는 방법이다 (Billsus and Pazzani, 1998 ; Sarwar et al., 2000 ; Goldberg et al., 2001). 그리고 협업 필터링과 내용기반 필터링을 결합한 하이브리드 기법으로 희박성을 완화시키려는 연구가 많이 진행되고 있다(Balabanovic and Shoham, 1997 ; Basu et al., 1998 ; Condiff et al., 1999 ; Good et al., 1999 ; Huang et al., 2002 ; Pazzani, 1999). 이 기법은 데이터가 희박하여 협업필터링기법으로는 추천이 힘들 때, 고객들의 인구통계정보나 거래내역기록, 아이템의 속성정보를 활용하여 추천을 하는 기법이다.

확장성(Scalability)은 고객의 수와 거래데이터의 개수가 늘어남에 따라 목표고객의 최근접이웃을 찾기 위한 연산이 기하급수적으로 늘어난다는 문제점이다. 예를 들어, 협업 필터링이 만 명의 목표고객들에게 추천을 하고자 한다면, 만 명의 목표고객들의 최근접이웃을 찾기 위해 수십 만 명의 고객데이터와 이들이 부여한 수백 만개의 평가치를 검색해야 한다. 추천시스템이 상대하는 고객의 수나 취급하는 아이템의 개수가 늘어나면 최근접이웃을 구성하기 위한 계산의 양이 기하급수적으로 늘어나서 추천목록을 생성하기까지 오랜 시간이 걸려서 시스템의 효율성이 떨어지게 된다. 확장성의 문제를 완화하기 위해 개별 고객들간의 유사도를 계산하는 대신, 아이템들간의 유사도를 계산하거나(Sarwar et al., 2001) 고객을 군집화한 후 군집들간의 유사도를 계산하는 방법들이(Li and Yamada, 2004 ; Xue et al., 2005) 제안되었다.

투명성(Transparency)은 추천 결과와 고객의 선호도와의 관계가 불명확하다는 문제점이다(Li and Yamada, 2004). 협업필터링은 수식계산을 포함한 블랙박스를 통하여 고객에게 추천목록을 제시하기 때문에, 고객이 아이템들이 어떻게 추천되었는지를 이해하는 것이 어려울 수 있다. 투명성 문제를 완화하기 위해 김재경 등(2006)은 추천과정에서 발생한 고객의 프로파일 정보, 최근접이웃 고객의 정보, 고객의 평가치 정보, 추천목록의 정보 등과 웹로그 데이터, 아이템 데이터, 시스템이 생성한 추천목록 데이터 등을 이용하여 아이템 조회비율, 장바구니에 담은 비율, 구매비율, 유명 매체 추천 등 20가지 유형의 설명기능을 추가한 추천시스템을 구현하였다.

본 연구에서 제안하는 장르별 협업필터링 방법은 이러한 문제점들 중에서 희박성과 확장성 문제를 극복하고자 하는 것이다. 장르별 협업필터링 방법에서는 먼저 아이템에 대한 평가치 정보들을 아이템의 상위 카테고리인 장르에 대한 정보로 집적한다. 이렇게 함으로써, 협업필터링에서 사용할 매트릭스의 차원을 축소할 수 있고, 따라서 희박성이 완화될 수 있다. 또한 차원이 축소됨으로써, 그만큼 계산의 양이 줄어들게 되므로 확장성 문제에도 대처할 수 있게 되는 것이다.

3. 데이터 준비와 전통적 협업 필터링 추천시스템

3.1. 원본 데이터

본 연구의 대상 영역은 영화로서, 미네소타 대학의 GroupLens Research Project에서 수집된 MovieLens 데이터를 사용하였다(Resnick et al., 1994). MovieLens 데이터는 1997년 9월에서 1998년 4월까지의 조사기간 동안 수집된 1682편의 영화에 대한 943명의 고객들의 평가치로 구성 되어 있다. 조사대상인 1682편의 영화의 배포연도는 1922년부터 1998년까지이고 배포연도의 정보가 없는 1편을 제외한 각 연대별 분포는 <표 2>와 같다.

<표 2> 조사대상 1681편의 연대별 편수

연대	1920	1930	1940	1950	1960	1970	1980	1990
편수	2	29	45	57	46	55	110	1337

MovieLens 데이터는 User ID와 Movie ID, 각 고객들이 관람한 영화에 대해 1점에서 5점까지의 5점 척도로 부여한 평가치, 평가한 시점의 시간으로 구성되어있다. 고객 × 영화 매트릭스의 칸의 총 개수는 1586126(= 943 × 1682)개인데, 이 중에서 100000개의 칸에 평가치가 부여되어 있다. 즉, 평가치가 부여된 칸의 수가 6.3%에 불과하여 MovieLens 데이터의 희박성은 93.7%이다. 각 영화는 미국의 최대 영화 데이터베이스인 IMDB(Internet Movie Database, http://us.imdb.com)의 기준에 따라 18개의 장르에 속해있다. 장르에 대한 정보가 없는 2편을 제외한 장르별 영화 수는 <표 3>과 같다. 한 영화는 최소 1개에서 최대 5개까지의 장르에 속해있다.

<표 3> 장르별 영화 수

장르	편수	장르	편수	장르	편수
Action	251	Documentary	50	Mystery	61
Adventure	135	Drama	725	Romance	247
Animation	42	Fantasy	22	Sci-Fi	101
Children	122	Film-Noir	24	Thriller	251
Comedy	505	Horror	92	War	71
Crime	109	Musical	56	Western	27

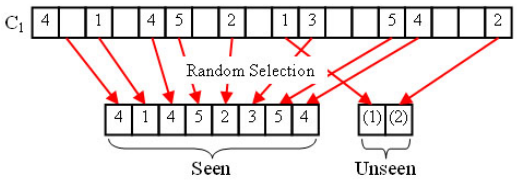
3.2 데이터 준비 및 k의 설정

본 연구에서는, 먼저 고객×영화의 매트릭스 준비하고 다음으로 최근접이웃의 수를 설정하였다.

제 1단계 : 고객 × 영화의 매트릭스 준비

1682편의 영화들에 대해 943명의 고객들이 부여한 평가치 매트릭스를 준비하였다. 943명의 고객들로부터 무작위로 70%인 643명의 고객을 추출하여 학습데이터집합(Training Data Set)으로 사용하고, 20%인 200명의 고객을 검증데이터집합(Validation Data Set)으로 사용하고, 나머지 10%인 100명을 평가데이터집합(Test Data Set)으로 사용하였다. 또한 샘플링에 따른 Bias를 없애기 위해 10-fold Cross Validation을 수행하였다.

검증데이터집합과 평가데이터집합에 속하는 각 고객의 평가치들은 [그림 2]와 같이 80 : 20으로 나누었다. 모든 평가치는 이미 주어져 있는 상태이다. 구현하는 추천시스템의 성능측정을 위해서 일부 아이템의 평가치를 모른다고 가정하고, 그 아이템의 평가치를 예측한 후에 그것을 원래 주어졌던 평가치와 비교해야 한다. 즉, 각 고객의 1682편의 영화에 대한 평가치들을 무작위로 80%와 20%로 나눈 후에 80%는 알려진 평가치로 사용하고 20%는 모른다고 가정하고 추천시스템이 예측해야 하는 평가치의 대상으로 사용하는 것이다. 80% 부분을 'Seen'으로, 20% 부분을 'Unseen'으로 명명한다.



[그림 2] 추천시스템의 성능평가를 위한 각 고객들의 평가치의 분할

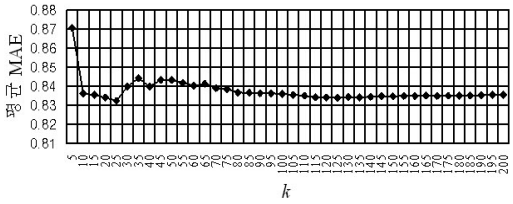
제 2단계 : 목표고객의 최근접이웃의 수 설정

최근접이웃의 수인 k를 설정하기 위하여 아래의 과정에 따라 k를 증가시켜가면서 MAE의 변화를 관찰하였다. Validation Data Set의 고객들을 목표고객으로 설정하고 Training Data Set의 고객들 중에

서 최근접이웃을 찾았다.

① n을1로 설정한다.
② k를 5 ×n으로 설정한다
③ 목표고객의 Seen부분의 평가치들을 이용하여 Training Data Set의 고객들과의 유사도를 계산한다.
④ 유사도가 높은 상위 k명으로 목표고객의 최근접이웃을 구성한다.
⑤ 구성된 최근접이웃의 평가치를 기반으로 목표고객의 Seen부분에 속하는 않는 아이템에 대한 평가치를 예측한다.
⑥ 목표고객의 Unseen부분의 실제 평가치와 ⑤에서 예측된 평가치로 MAE를 계산한다.
⑦ Validation Data Set에 다른 목표고객이 남아있으면 ③으로 돌아간다.
⑧ Validation Data Set의 모든 목표고객들의 MAE가 구해졌으면 k값에 대한 MAE의 평균을 계산한다.
⑨ k값이 200보다 작으면 n을 1증가시켜 ②로 돌아간다.

본 연구에서는 k를 5에서부터 200까지 5씩 증가시키면서 최근접이웃의 수의 변화에 따른 평균 MAE의 변화를 관찰하여 [그림 3]과 같이 정리하였다.

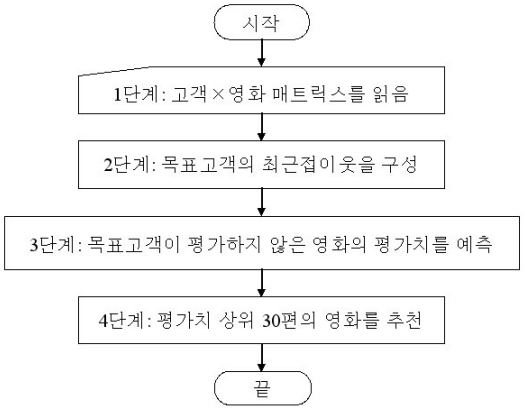


[그림 3] k의 변화에 따른 평균 MAE의 변화

[그림 3]에서 보는 바와 같이 최근접이웃의 수가 25명일 때, 평균 MAE가 제일 작았다. 따라서 본 연구에서는 최적의 최근접이웃의 수를 25명으로 설정하였다.

3.3 전통적 협업 필터링 시스템

전통적인 협업 필터링(Sarwar et al., 2001)을 적용하여 구현한 추천시스템을 Basic_CF로 명명하였다. Basic_CF는 [그림 4]와 같이 진행된다.



[그림 4] Basic_CF의 순서도

<표 4>는 Test Data Set에 대한 추천 결과의 일부를 보여주고 있다.

<표 4> Basic_CF의 고객별 추천 목록(일부)

User	추천목록의 Movie ID								
	1	2	3	4	…	27	28	29	30
2	174	318	172	135	…	659	183	176	69
5	313	515	475	15	…	275	227	751	329
28	269	302	515	242	…	902	272	1006	273
47	98	50	56	100	…	313	114	195	132
50	300	313	50	258	…	311	896	79	434
	⋮	⋮	⋮	⋮	…	⋮	⋮	⋮	⋮
903	313	258	316	286	…	309	83	344	905
915	100	50	1	56	…	631	318	69	223
918	50	181	100	866	…	275	257	762	313
939	313	50	172	302	…	192	179	187	178
943	515	257	285	898	…	272	1152	70	514

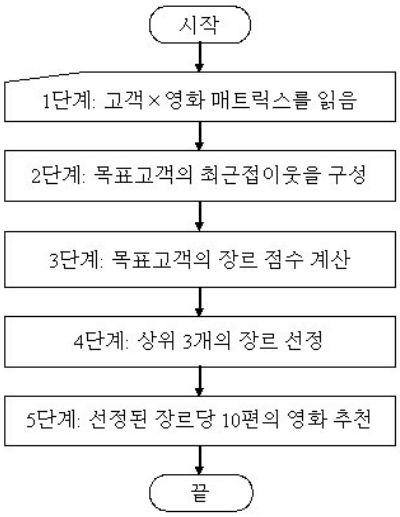
<표 4>에서 음영으로 처리된 부분은 추천목록과 Test Data Set의 Unseen부분 양쪽에 속해있는 Movie ID를 나타낸다. 즉, 추천이 적중한 Movie ID들이다. Test Data Set에 대한 Basic_CF의 10 Fold 성능 평균은 <표 5>와 같다.

<표 5> Basic_CF의 성능(10 Folds 평균)

	Recall	Precision	F1
측정값	0.072	0.043	0.046

4. 장르별 협업 필터링 추천시스템

본 연구에서 제안하는 장르별 협업 필터링 추천시스템 GenreWise_CF(Genre wise Collaborative Filtering)는 최종적으로 영화를 추천하기 전에, 중간 단계에서 영화를 분류하는 기준인 장르의 점수를 계산하여 영화추천에 이용한다. 장르점수를 기반으로 선정된 장르에 속하는 영화를 추천하기 위해 최근접이웃을 새롭게 구성하여 영화를 추천하는 방법으로서, GenreWise_CF의 과정은 [그림 5]와 같다.



[그림 5] GenreWise_CF의 순서도

[그림 5]의 제 3단계에서 목표고객의 장르점수는 최근접이웃의 평가치를 이용하여 계산되는데, 먼저 최근접이웃 j의 장르 l의 장르점수 $G_{j,l}$은 식 (8)에 의해 계산된다.

$$G_{j,l} = \begin{cases} \dfrac{\sum_{i \in E_l^j} R_{j,i}}{|E_l^j|} & if\ |E_l^j| \neq \phi \\ 0 & if\ |E_l^j| = \phi \end{cases} \quad (8)$$

여기서 E_l^j은 최근접이웃 j가 본 장르 l에 속하는 영화의 집합을 뜻한다.

장르의 개수를 L이라 할 때, $\overline{G}_j$는 최근접이웃 j의 장르점수들을 평균한 값으로서 식 (9)와 같이 계산된다.

$$\overline{G}_j = \frac{\sum_{l=1}^{L} G_{j,l}}{L} \quad (9)$$

목표고객 A의 장르 l의 장르점수 $GS_{A,l}$은 식 (10)과 같이 계산된다.

$$GS_{A,l} = \frac{\sum_{j=1}^{k} w(A,j) \dfrac{(G_{j,l} - \overline{G}_j)}{\sigma}}{\sum_{j=1}^{k} |w(A,j)|} \quad (10)$$

여기서 k는 최근접이웃의 수이고, σ는 $G_{j,l}$의 표준편차를 뜻한다.

장르 선정의 정확성은 Test Data Set 고객의 Unseen 부분의 아이템 중에서 평가치가 제일 높은 영화의 장르가 선정되었는가로 측정하였다. 선정된 장르들과 적중된 장르의 일부는 <표 6>과 같다.

<표 6> 최근접이웃의 평가치기반의 장르 선정(일부)

User ID	목표 고객의 장르별 장르점수																		선정된 장르 ID		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18			
2	0.83	0.34	-1.16	-0.35	0.65	0.50	-2.29	0.98	-2.16	-0.19	-0.16	-0.26	0.84	0.97	0.40	1.10	0.74	-0.79	16	8	14
5	1.04	0.82	-0.78	-0.43	0.84	0.60	-1.53	1.19	-1.92	-1.08	-0.68	-0.49	0.70	1.13	0.85	0.68	1.25	-2.19	17	8	14
28	0.83	0.65	-1.03	0.29	0.73	0.90	-2.31	1.20	-1.54	-0.80	-0.35	-0.46	0.26	0.92	1.15	0.68	0.67	-1.80	8	15	14
47	0.37	0.43	-0.22	-0.28	-0.01	0.78	-2.63	0.88	-1.40	0.53	-0.17	-0.33	0.40	0.60	0.50	0.43	0.96	-0.82	17	8	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
915	0.69	0.65	0.13	0.12	0.49	0.39	-2.02	0.85	-0.98	-1.37	-0.39	-0.58	-0.08	0.87	0.74	0.55	0.90	-0.98	17	14	8
918	1.04	0.89	0.10	0.36	0.76	0.82	-1.87	1.11	-1.71	-1.67	-0.97	-0.75	0.31	1.17	0.95	0.74	1.26	-2.53	17	14	8
939	0.65	0.42	-0.40	-0.47	0.32	0.72	-3.10	0.92	-2.08	0.68	-0.17	-0.20	0.45	0.75	0.69	0.72	0.96	-0.85	17	8	14
943	0.99	0.80	-0.31	0.00	0.93	1.09	-1.71	1.16	-1.62	-1.45	-0.98	-0.82	0.05	1.06	0.95	0.81	1.16	-2.11	8	17	6

<표 6>의 '선정된 장르 ID'열에서 음영으로 처리된 부분은 선정된 장르와 Test Data Set 고객의 Unseen 부분의 영화 중에서 평가치가 제일 높은 영화의 장르가 일치하는 것을 뜻한다. 이 일치여부를 Test Data Set 고객 모두에 대하여 구한 후에, 이것들의 평균을 계산하여 장르 선정의 정확성을 측정하기 위한 지표로 사용하였다. 추천장르를 선정한 후, 각 장르에 속하는 영화를 10편씩 총 30편을 추천한 GenreWise_CF의 결과의 일부는 <표 7>과 같다.

<표 7>의 '추천목록의 Movie ID'중에서 음영으로 처리된 부분은 Test Data Set 고객의 Unseen 부분의 영화와 추천목록의 영화가 일치하는 것을 뜻한다.

4.2 GenreWise_CF 시스템의 성능 평가

Basic_CF와 GenreWise_CF의 성능을 F1 지표로 비교한 결과는 <표 8>과 같다.

<표 7> GenreWise_CF의 고객별 추천목록(일부)

User ID	추천 장르	추천목록의 Movie ID									
		1	2	3	4	5	6	7	8	9	10
2	16	135	98	302	185	12	480	654	23	195	603
	8	318	172	135	98	427	191	187	197	483	64
	14	172	197	181	185	483	781	69	402	568	794
5	17	515	326	172	205	318	22	971	744	474	483
	8	313	515	475	15	326	272	124	9	316	172
	14	313	172	781	875	955	275	751	170	483	283
28	8	515	15	9	475	127	313	955	875	887	270
	15	270	121	1006	411	172	181	179	257	183	204
	14	313	955	875	268	319	781	275	421	283	411
47	17	50	181	176	483	172	190	528	474	318	651
	8	98	56	100	275	64	483	475	127	172	87
	6	56	100	12	127	182	156	194	656	248	239
50	1	300	313	50	515	172	183	181	195	265	96
	14	313	50	172	181	173	311	485	498	648	237
	8	313	258	272	515	172	347	655	510	98	311

<표 8> Basic_CF와 GenreWise_CF의 성능 비교

Fold	Basic_CF F1	GenreWise_CF	
		장르적중률	F1
01	0.048	0.90	0.053
02	0.044	0.89	0.053
03	0.050	0.89	0.059
04	0.052	0.83	0.056
05	0.047	0.89	0.059
06	0.037	0.90	0.050
07	0.048	0.88	0.065
08	0.043	0.87	0.053
09	0.044	0.86	0.053
10	0.043	0.86	0.056
평균	**0.046**	**0.88**	**0.056**

<표 81>에서 보는 바와 같이, 영화추천의 중간 단계인 장르선정의 정확성을 나타내는 '장르적중률'은 0.88로 높게 나타났다. F1 성능지표에 있어서는, GenreWise_CF가 0.056으로서 Basic_CF의 0.046보다 0.010만큼 향상되었다. 이같은 결과가 통계적으로 신뢰할 만한 것인지를 평가하기 위해 Pairwise t-Test를 수행하였고 그 결과는 <표 9>와 같다.

<표 9> 성능차이의 유의성 검증

대응쌍	대 응 차							유의확률(양쪽)
	평균차	표준편차	차이의 표준오차	차이의 95% 신뢰구간		t-Value	자유도	
				상한	하한			
GenreWise_CF - Basic_CF	0.010	0.004	0.001	0.007	0.013	8.250	9	0.000

<표 9>에서 보는 바와 같이 t값이 8.250으로서 유의수준 1%일 때의 t값인 2.821보다 크게 나타났다. 그러므로, 유의수준 1%에서 GenreWise_CF의 성능이 Basic_CF의 성능보다 우수하다고 할 수 있다.

5. 결론 및 향후 연구

경쟁이 점점 치열해지는 기업환경에서 제품이나 서비스 제공자들이 고객 개개인에게 맞춤 아이템이나 정보를 제공하는 것은 매출신장에 있어 중요한 요인 중의 하나이다. 이런 개인화 서비스를 가능하게 하는 방법 중의 하나가 추천시스템이다. 추천시스템에서 사용되는 여러 기법들 중에서 전자상거래에서 성공으로 적용된 대표적인 기법은 협업 필터링이다. 협업 필터링은 명시적인 속성만으로 규정짓기 힘든 동영상이나 음악 같은 아이템들에도 효과적인 성능을 발휘한다는 장점이 있다.

본 연구에서는 전통적인 협업필터링의 한계점인 희박성과 확장성에 대처하는 방안으로 장르별 협업 필터링을 제안하였다. 장르별 협업 필터링은 아이템을 최종적으로 추천하기 전에 아이템의 상위 카테고리인 장르에 대한 정보를 활용하는 방법이다. 즉, 아이템에 대한 평가치 정보를 아이템의 상위 개념인 장르에 대한 정보로 집적함으로써, 협업 필터링의 대상이 되는 고객×아이템 매트릭스의 차원을 축소시키는 것이다.

본 연구에서는 영화 데이터를 사용하여 장르별 협업 필터링을 적용한 추천시스템인 GenreWise_CF를 개발하였다. 실험 결과, 본 연구에서 제안한 GenreWise_CF의 장르추천 적중률은 88%로서 매우 높은 적중률을 보였으며, 개별 영화의 추천에 있어서도 전통적인 협업 필터링을 적용하여 개발한 추천시스템인 Basic_CF보다 향상된 성능을 보였다. 다시 말하면, 장르별 협업 필터링 방법은 전통적 협업 필터링 방법의 한계점에 대처할 수 있을 뿐만 아니라, 추천 성능을 향상시키는 결과를 가져왔다

향후 연구로는 GenreWise_CF 시스템의 활용에 대한 구상을 기술하고자 한다. 유비쿼터스 컴퓨팅 환경에 대비하여, 영화 추천도 고객에게 모바일 환경에서 제공되어야 한다. 장르별 협업 필터링 방법은 전통적인 협업 필터링 방법과는 달리 중간에 장르를 선정하는 단계가 있다. 그러므로 장르별 협업 필터링 방법은 유비쿼터스 컴퓨팅 환경에 적합한 추천시스템의 개발에 적합하게 활용될 수 있다. 즉, 고객과의 상호작용을 통하여 장르를 선정하게 함으로써 추천성능뿐만 아니라, 추천서비스에 대한 고객의 만족도도 향상시킬 수 있을 것이다.

참고문헌

[1] 김재경, 이희애, 안도현, 조윤호, "설명기능을 추가한 협업 필터링 기반 개인별 상품추천시스템 : WebCF-Exp", 경영학연구, Vol.35, No.2(2006), 493~519.

[2] Adomavicius, G. and A. Tuzhihin, "Toward the Next Generation of Recommenders Systems : A Survey of the State-of-the-art and Possible Extensions", *IEEE Transactions on Knowledge and Data Engineering*, Vol.17, No.6(2005), 734~749.

[3] Ansari, A., S. Essegaier and R. Kohli, "Internet Recommender Systems", *Journal of Marketing Research*, Vol.37, No.3(2000), 363~375.

[4] Balabanovic, M. and Y. Shoham, "FAB: Content-based, Collaborative Recommendation", *Communications of the ACM*, Vol.40, No.3 (1997), 66~72.

[5] Basu, C., H. Hirsh and W. Cohen, "Recommendation as Classification : Using Social and Content based Information in Recommendation", *Proceedings of the 15th National Conference on Artificial Intelligence* (1998), 714~720.

[6] Billsus, D. and M. J. Pazzani, "Learning Collaborative Information Filters", *Proceedings of the 15th International Conference on Machine Learning*(1998), 46~54.

[7] Breese, J. S., D. Heckerman and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence*(1998), 43~52.

[8] Canny, J., "Collaborative Filtering with Privacy," *Proceedings of IEEE Symposium on Security and Privacy*(2002). 45~57.

[9] Condliff, M. K., D. D. Lewis, D. Madigan and C. Posse, "Bayesian Mixed-effects Models for Recommender Systems", *Proceedings of the ACM SIGIR Workshop on Recommender Systems*(1999).

[10] Goldberg, K., T. Roeder, D. Gupta, and C. Perkins, "Eigentaste : A Constant Time Collaborative Filtering Algorithm", *Information Retrieval J.*, Vol.4, No.2(2001), 133~151.

[11] Good, N., J. Schafer, J. Konstan, A. Borchers, B. Sarwar, J. Herlocker and J. Riedl, "Combining Collaborative Filtering with Personal Agents for Better Recommendations", *Proceedings of the 16th National Conference on Artificial Intelligence*, (1999), 439~446.

[12] Herlocker, J., J. A. Konstan, R. Borschers and J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering", *Proceedings of the 22nd ACM SIGIR Conf. on Research and Development in Information Retrieval*(1999), 230~237.

[13] Hill, W., L. Stead, M. Rosenstein, and G. Furnas, "Recommending and Evaluating Choices in a Virtual Community of Use", *Proceedings of Conf. Human Factors in Computing Systems*(1995).

[14] Huang, Z., H. Chen, and D. Zeng, "Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering", *ACM Trans. Information Systems*, Vol 22, No.1(2004), 116~142.

[15] Huang, Z., W. T. Chung, T.-H. Ong and H. Chen, "A Graph-based Recommender System for Digital Library", *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*(2002), 65~73.

[16] Lekakos, G. and G. M. Giaglis, "Improving the Prediction Accuracy of Recommendation Algorithms : Approaches Anchored on Human Factors", *Interacting with Computers*, Vol. 18(2006), 410~431.

[17] Li, P. and S. Yamada, "A Movie Recommender System Based on Inductive Learning", *IEEE Conf. on Cybernetics and Intelligent Systems*,(2004), 318~323.

[18] Mulvenna, M., S. S. Anand and A. G. Büchner, "Personalization on the Net Using Web Mining : Introduction", *Communications of the ACM,* Vol.43, No.8(2000), 122~125.

[19] Pazzani, M. "A Framework for Collaborative, Content based and Demographic Filtering", *Artificial Intelligence Rev*(1999), 393~408.

[20] Resnick, P., N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl, "Grouplens : An Open Architecture for Collaborative Filtering of Netnews", *Proceedings of the ACM Conf. on Computer Supported Cooperative Work* (1994), 175~186.

[21] Sarwar, B., *Sparsity, Scalability, and Distribution in Recommender Systems*, Ph.D. Diss., Dept. of Computer and Information Sciences, Univ. of Minnesota(2001).

[22] Sarwar, B., G. Karypis, J. Konstan and J. Riedl, "Application of Dimensionality Reduction in Recommender System : a Case Study", *Proceedings of the ACM WebKDD-2000 Workshop*(2000).

[23] Sarwar, B., G. Karypis, J. Konstan and J. Riedl, "Item based Collaborative Filtering Recommendation Algorithms", *Proceedings of the 10th International World Wide Web Conference*(2001), 285~295.

[24] Shardanand, U. and P. Maes, "Social Information Filtering : Algorithms for Automating 'Word of Mouth'", *Proceedings of the ACM CHI'95 Conf. on Human Factors in Computing Systems*(1995), 210~217.

[25] Terveen, L., W. Hill, B. Amento, D. McDonald, and J. Creter, "PHOAKS : A System for Sharing Recommendations", *Communications of the ACM*, Vol.40, No.3(1997), 59~62.

[26] Xue, G. -R., C. Lin, Q. Yang, W. Xi, H. -J. Zeng, Y. Yu and Z. Chen, "Scalable Collaborative Filtering using Cluster based Smoothing", *Proceedings of the 2005 ACM SIGIR Conference*(2005), 114~121.

[27] Yang, W., Z. Weng and M. You, "An Improved Collaborative Filtering Method for Recommendations' Generation", *IEEE Int'l Conf. on Systems, Man and Cybernetics* (2004), 4135~4139

Abstract

Performance Improvement of a Movie Recommendation System using Genre-wise Collaborative Filtering

Jae Sik Lee* · Seog Du Park**

This paper proposes a new method of weighted template matching for machine-printed numeral recognition. The proposed weighted template matching, which emphasizes the feature of a pattern using adaptive Hamming distance on local feature areas, improves the recognition rate while template matching processes an input image as one global feature. Template matching is vulnerable to random noises that generate ragged outlines of a pattern when it is binarized. This paper offers a method of chain code trimming in order to remove ragged outlines. The method corrects specific chain codes within the chain codes of the inner and the outer contour of a pattern. The experiment compares confusion matrices of both the template matching and the proposed weighted template matching with chain code trimming. The result shows that the proposed method improves fairly the recognition rate of the machine-printed numerals.

Key Words : Template Matching, Numeral Recognition, Chain Code, Hamming Distance

* Division of e-Business, School of Business, Ajou University
** Ubiquitous Convergence Research Institute