Journal of the Korean Institute of Industrial Engineers
Vol. 36, No. 3, pp. 193-202, September 2010.

차원 감소 기법을 이용한 전자 상거래 추천 시스템

김용수[1†] · 염봉진[2] · Norman Kim[3]

[1]경기대학교 산업경영공학과 / [2]KAIST 산업 및 시스템공학과 /
[3]Rutgers, The State University of New Jersey, RUTCOR

Development of a Recommender System for E-Commerce Sites Using a Dimensionality Reduction Technique

Yong Soo Kim[1] · Bong-Jin Yum[2] · Norman Kim[3]

[1]Department of Industrial and Management Engineering, Kyonggi University
[2]Department of Industrial and System Engineering, KAIST
[3]RUTCOR, Rutgers, The State University of New Jersey

The recommender system is a typical software solution for personalized services which are now popular in e-commerce sites. Most of the existing recommender systems are based on customers' explicit rating data on items (e.g., ratings on movies), and it is only recently that recommender systems based on implicit ratings have been proposed as a better alternative. Implicit ratings of a customer on those items that are clicked but not purchased can be inferred from the customer's navigational and behavioral patterns. In this article, a dimensionality reduction (DR) technique is newly applied to the implicit rating-based recommender system, and its effectiveness is assessed using an experimental e-commerce site. The experimental results indicate that the performance of the proposed approach is superior or at least similar to the conventional collaborative filtering (CF)-based approach unless the number of recommended products is 'large.' In addition, the proposed approach requires less memory space and is computationally more efficient.

Keywords: Recommender System, Data mining, Dimensionality Reduction, E-commerce

1. 서 론

1.1 연구 배경 및 목적

인터넷의 급속한 보급과 더불어 전자 상거래 또한 매우 빠르게 발전하고 있다. 이에 따라 고객이 전자 상거래 업체를 선택할 수 있는 기회가 늘어나게 되면서, 전자 상거래 업체의 경쟁도 치열해지고 있다. 이렇게 치열한 경쟁에서 살아남기 위해, 전자 상거래 업체는 고객을 꾸준히 유지하고자 하는 고객관계관리(CRM) 전략을 펼쳐 나가고 있다. 추천 시스템(recommender system)은 그러한 전략을 구현하기 위한 대표적 방법이다.

추천 시스템은 각 고객에게 그 고객이 구매할 만한 상품을 추천함으로써 구매를 돕는 시스템이다. 특히, 매우 방대한 양의 상품을 가진 전자 상거래 사이트에서 고객이 모든 상품을 일일이 검색하기 어려울 때, 그 고객이 선호할 만한 상품을 추천함으로써 구매를 돕고자 하는 것이다(Berson *et al*., 2000; Lawrence *et al*., 2001; Sarwar *et al*., 2000a; Yuan and Chang, 2001).

전자 상거래를 위한 대부분의 추천 시스템은 고객의 구매 데이터만을 이용하고 있으며, 고객의 탐색 및 행동 패턴 데이터를 이용하는 추천 시스템은 찾아보기 어렵다. 예외적으로, Kim *et al*.(2005)은 고객이 클릭한 상품을 읽은 시간, 방문 회수, 자주 클릭하는 상품군, 클릭 형태(browsing을 통한 클릭과 searching

† 연락저자 : 김용수 교수, 443-760 경기도 수원시 영통구 이의동 산 94-6 경기대학교 산업경영공학과, Fax : 031) 244-3534,
E-mail : kimys@kgu.ac.kr
2010년 6월 16일 접수; 2010년 8월 15일 수정본 접수; 2010년 8월 20일 게재 확정.

을 통한 클릭), 상품 정보의 인쇄 및 bookmaking 여부 등과 같은 탐색 및 행동 패턴을 고려한 협업적 필터링 기법을 개발하였으며, 구매 데이터만을 이용했을 때 보다 더욱 우수한 성능을 나타낼 수 있다는 것을 실험적으로 보였다.

본 연구에서는 Kim *et al*.(2005)에서 제안한 방법으로 추정한 고객 선호도 데이터에 대해 협업적 필터링 대신 차원 감소 기법을 적용한 추천 시스템을 개발하고, 이 두 접근방법의 성능을 비교하고자 한다. 본 논문에서 차원 감소 기법을 고려하게 된 배경은 다음과 같다. 첫째, 고객 선호도 데이터에 협업적 필터링과 차원 감소 기법을 적용하여 두 방법의 예측력을 비교하기 위한 것이다. 둘째, 상대적으로 적은 저장 용량과 빠른 추천 속도를 갖는 차원 감소 기법의 활용 가능성을 파악하기 위한 것이다. 즉, 차원 감소 기법이 협업적 필터링보다 우수하거나 유사한 성능을 보인다면 후자를 전자로 대체하는 것이 효과적이다.

두 방법의 성능 평가를 위해, KAIST 내에 실험용 전자 상거래 사이트를 개설, 운영하였으며, 이로부터 수집한 데이터를 바탕으로 두 접근방법을 비교하여, 추천 상품 수가 지나치게 크지 않을 때에는 차원 감소 기법을 이용한 추천 시스템의 성능이 협업적 필터링을 이용한 추천 시스템의 성능과 유사하거나 더 우수하다는 것을 보였다.

1.2 문헌 연구

현재까지 알려진 많은 추천 시스템들이 가장 보편적으로 채택하고 있는 방법은 협업적 필터링이다. 협업적 필터링이란, 각 고객에 대해 가장 유사한 성향을 지닌 고객군(neighbor)을 찾아내고, 고객군이 선호하는 상품을 그 고객에게 추천하는 것이다(Hill *et al*., 1995; Resnick *et al*., 1994; Shardanand and Maes, 1995). 그러나, 협업적 필터링은 고객이 직접 선호도를 입력한 explicit rating 데이터에 대해서는 효과적이나, 전자상거래 데이터와 같은 binary data에 대해서는 적용성이 저하되는 문제점이 있다(Hayes *et al*., 2001). 이러한 문제점을 극복하기 위해, 최근에는 고객의 탐색 패턴과 행동 패턴을 이용하여 선호도를 파악하고자 하는 방법이 제안되고 있다(Claypool *et al*., 2001; Kelly and Belkin, 2001; Rafter and Smyth, 2001). 이러한 선호도 측정 기법을 implicit rating이라 한다. Implicit rating에서는 고객이 특정 항목에 대해 점수를 부여하는 것이 아니라, 고객의 탐색 및 행동 패턴을 수치화 하여 고객의 선호 정도를 나타낸다(Nichols, 1997). 또한, Lee *et al*.(2000, 2001)에 의해 다양한 전자상거래 사이트의 click stream을 분석한 사례연구도 이루어졌다. 즉, 전자 상거래 사이트에서 제품 구매 이전의 행위를 분석함으로써 그 고객이 어떤 제품에 관심이 있는지를 파악하고자 하는 것이다. 더 나아가 Kim *et al*.(2005)은 구매 이전의 행위인 '클릭한 상품을 읽은 시간', '방문 회수', '자주 클릭하는 상품군', '클릭 형태(browsing을 통한 클릭과 searching을 통한 클릭)', '상품 정보의 인쇄' 및 'bookmaking 여부' 등과 같은 고객의 탐색 및 행동 패턴을 고려한 협업적 필터링 기법을 개발하였고, 구매 데이터만을 이용했을 때 보다 더 우수한 성능을 가진다는 것을 실험적으로 보였다.

한편, 기존의 협업적 필터링은 데이터가 많지 않은 경우에는 높은 성능을 발휘하지 못한다는 것이 알려져 있다. 이를 보완하기 위해 차원감소 기법에 대한 연구가 수행되었다(Billsus, and Pazzani, 1998; Goldberg *et al*., 2001; Sarwar *et al*., 2000b; Kim and Yum, 2005). 그 중 Sarwar *et al*.(2000b)은 특이값 분해(SVD, Singular Value Decomposition)를 통해 차원을 감소시켜 구한 점수를 사용하여 추천 목록을 구성하는 시스템을 제안하였고, 그 성능을 실험적으로 검증하였다. 영화에 대한 점수를 고객으로부터 직접 입력 받은 explicit rating data에 대해서는 기존의 협업적 필터링 방법보다 더 우수한 성능을 보였으나, 전자상거래 사이트에서 얻은 implicit rating data(구매 또는 비구매)에 대해서는 낮은 성능을 보였다. 또한, Goldberg *et al*.(2001)은 주성분분석을 이용한 추천 시스템을 개발하여 explicit rating data에 대해 적용하였다. 그러나 각 고객이 일정 개수의 공통 상품을 모두 평가해야 한다는 조건을 전제로 하고 있으므로, 전자상거래에 직접 적용하기는 어렵다고 할 수 있다. 이에 반해, Kim and Yum(2005)은 Sarwar *et al*.(2000b)과 Goldberg *et al*.(2001)의 연구를 통합, 확장함으로써 데이터 형태에 제한이 없으면서 기존의 연구보다 더 높은 성능을 보이는 추천 시스템을 개발하였다. Kim and Yum(2005)은 특이값 분해 기법을 이용하여 분산의 설명력이 큰 일정수의 주성분을 찾아낸 후, 각 고객의 주성분 값을 이용하여 군집분석을 수행하였다. 그리고, 각 범주에서 고객이 한 번도 클릭하지 않은 상품에 대한 선호도를 구한 후, 가장 선호도가 높은 *Top-N* 목록을 만들어 추천을 수행하도록 하였다. 아울러, 영화 선호도 데이터, 유머 선호도 데이터와 같은 explicit rating data에 대해 적용하여 그 우수성을 보였다. 본 연구에서는 고객의 행동 및 탐색 패턴으로부터 구한 선호도 데이터에 대해 Kim and Yum(2005)이 개발한 차원 감소 기법을 적용하였다.

2. 차원 감소 기법을 적용한 추천 시스템의 개발

2.1 전자 상거래 사이트에서 수집한 데이터 형태

전자상거래 환경 하에서 고객의 탐색 패턴(browsing, searching, 상품 클릭, 장바구니 담기, 구매)과 행동 패턴(클릭한 상

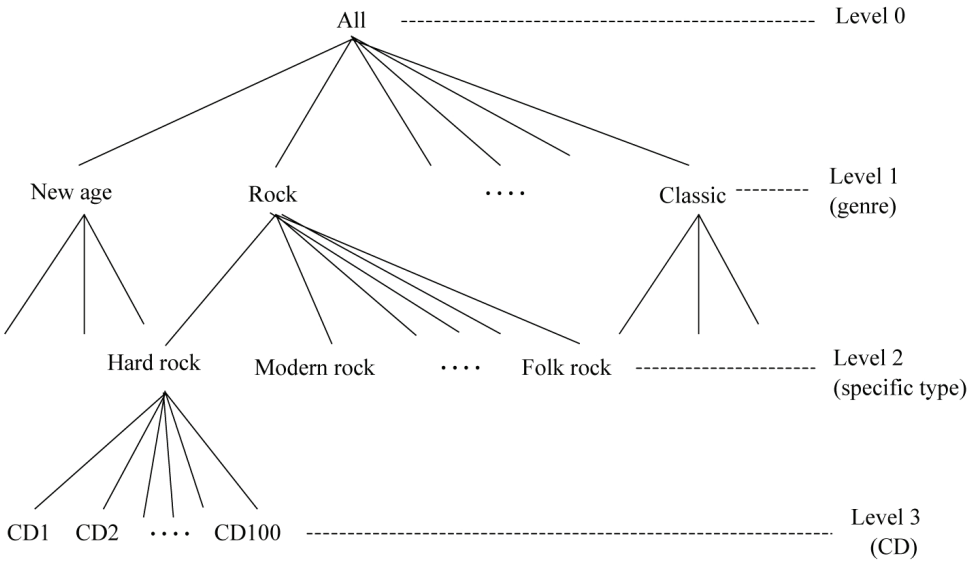

Figure 1. Product taxonomy of experimental CD e-commerce site

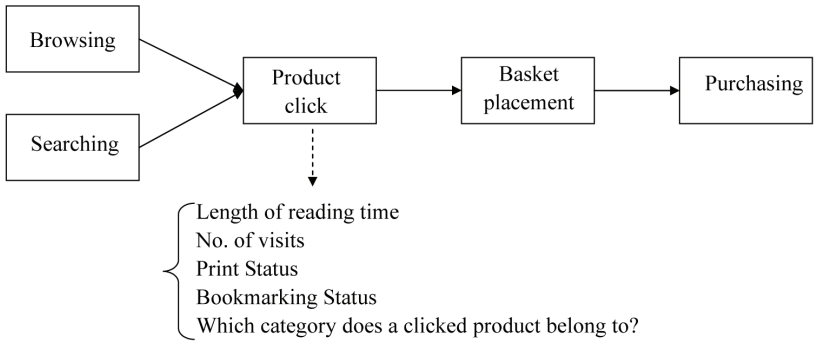

Figure 2. Possible actions that can be taken by customers in E- commerce sites and possible data that can be obtained from such actions

품을 읽은 시간, 방문회수, 상품 구조를 고려했을 때 어느 부류의 상품에 대한 클릭 비율이 높았는가 하는 정보)에 기반한 추천시스템의 개발을 위해, 본 절에서는 전자상거래 사이트에서 고객이 어떤 행동을 취할 수 있고, 그에 따라 어떤 종류의 데이터를 수집할 수 있는지에 대해 설명하고자 한다.

대부분의 전자상거래 사이트는 <Figure 1>과 같은 상품구조를 갖고 있다. 그리고, 전자상거래 사이트에서 고객이 취할 수 있는 행동과 수집 가능한 데이터는 <Figure 2>와 같다.

<Figure 2>는 전자상거래 사이트에서 고객이 로긴(log-in)에서부터 물건 구매에 이르기 까지 취할 수 있는 행동과 수집 가능한 데이터를 표현한 것이다. 먼저, 고객은 로긴 후, 자신이 선호할 만한 상품을 찾기 위해 browsing 할 수도 있고, 자신이 의도한 상품을 구매하기 위해 바로 searching을 할 수도 있을 것이다. browsing이나 searching을 하다가 관심 있는 특정한 상품을 클릭할 수도 있다. 그 때, 고객은 특정 상품에 대한 상세 정보를 볼 수 있게 되며, 그 페이지를 인쇄하거나 Bookmarking 등의 행동을 취할 수 있을 것이다. 이 작업은 당장은 아니나 추후에 구매를 하기 위해 이루어 질 수도 있고, 가격 및 사양을 비교하기 위해 이루어질 수도 있을 것이다. 그 외에 수집할 수 있는 데이터로는 '클릭한 상품을 읽은 시간', '그 상품의 방문회수', '그 상품이 포함되는 범주'와 같은 것이 있다. '클릭한 상품을 읽은 시간'이 길수록, '특정 상품의 방문 회수'가 많을수록 그 상품에 큰 관심을 갖고 있다고 유추할 수 있을 것이다. 그리고 어떤 범주의 상품에 대한 고객의 클릭 비율이 높다고 한다면, 그 범주의 상품들에 대해 많은 관심을 갖고 있다고 생각할 수 있다. 예를 들어, <Figure 1>의 Level 1에서 어떤 고객이 classic CD를 클릭한 비율이 Rock CD를 클릭한 비율보다 높다면 그 고객은 Rock 음악 보다 Classic 음악에 더 큰 관심을 가지고 있다고 유추할 수 있을 것이다.

<Table 1>은 위에서 언급한 바와 같이, 클릭이 이루어졌던 상품에 대해 수집 가능한 데이터들을 변수로 정리해 나열한 것이고, <Table 2>는 분석을 위한 데이터 구조를 표현한 것이다.

Table 1. Data collected from the experimental e-commerce site(Kim *et al*., 2005)

파라미터	설명
클릭형태	이진변수 : 검색 = 1, browsing = 0
방문회수	이산형 변수
클릭한 상품을 읽은 시간	연속형 변수(단위 : 초)
인쇄여부	이진변수 : 인쇄 = 1, 인쇄하지 않음 = 0
bookmarking 여부	이진변수 : bookmarking = 1, bookmarking하지 않음 = 0
특정 장르의 클릭 비율	연속형 변수 : 고객 i가 클릭한 상품 k에 대해 정의된 값. 상품 k가 속한 장르를 j라 했을 때, 상품 k에 대한 '특정 장르의 클릭 비율' $= \frac{\text{고객 } i\text{가 장르 } j\text{에 속한 상품을 클릭한 회수}}{\text{고객 } i\text{가 클릭한 총 상품 수}}$
특정 세부 장르의 클릭 비율	연속형 변수 : 고객 i가 클릭한 상품 k에 대해 정의된 값. 상품 k가 속한 세부 장르를 j라 했을 때, 상품 k에 대한 '특정 세부 장르의 클릭 비율' $= \frac{\text{고객 } i\text{가 세부장르 } j\text{에 속한 상품을 클릭한 회수}}{\text{고객 } i\text{가 클릭한 총 상품 수}}$
장바구니 담기 여부	이진변수 : 장바구니 담기 = 1, 장바구니 담기를 하지 않음 = 0
구매 여부	이진변수 : 구매 = 1, 구매하지 않음 = 0

Table 2. Structure of collected data(example)

고객	클릭한 CD	클릭 형태	읽은 시간	방문 회수	특정 장르의 클릭비율	특정 세부장르의 클릭비율	장바구니 담기 여부	구매 여부
1	A	0	27	2	0.67	0.33	1	1
1	C	1	31	1	0.67	0.33	1	0
1	D	1	6	1	0.33	0.33	0	0
2	A	1	9	1	0.25	0.25	0	0
2	B	0	21	3	0.75	0.50	1	0
2	C	1	12	1	0.75	0.50	1	1
2	E	0	16	1	0.75	0.25	0	0
					.			
					.			
					.			

2.2 데이터마이닝 기법을 이용한 고객의 선호도 추정 (Kim *et al*., 2005)

고객의 탐색 및 행동 패턴을 고려한 추천 시스템을 개발하기 위해서는 고객의 각 상품에 대한 선호도를 결정해야 한다. Kim *et al*.(2005)은 한 번이라도 클릭이 이루어진 상품에 대해 수치적으로 선호도를 결정하는 방법을 개발하였다. 만일, 상품을 구매했다면 선호도 값을 1로 결정하고, 구매하지 않았다면 고객의 탐색 및 행동 데이터를 이용하여 구매로 이어질 확률을 계산하였다. 이 과정은 인공 신경망 분석, 또는 로지스틱 회귀분석을 통해 수행하였다. 본 논문에서 특정 상품에 대한 고객의 선호도는 Kim *et al*.(2005)에서 개발한 방법을 사용하였으며, 자세한 절차는 다음과 같다.

Phase 1 : '장바구니 담기'에서 '구매'로 이어질 확률(p)을 구한다. 즉,

$$p = \frac{\text{구매가 이루어진 총 상품 수}}{\text{장바구니에 담겨진 총 상품 수}}$$

Phase 2 : '상품 클릭'에서 '장바구니 담기'로 이어질 확률(b)을 구한다. '장바구니 담기' 단계까지 도달하지 못한 상품의 경우에는 '클릭 형태', '방문 회수', '클릭한 상품을 읽은 시간', 'Level 1 ratio', 'Level 2 ratio', '인쇄 여부'와 'Bookmarking 여부'와 같은 변수만으로 구매에 도달할 확률을 구하는 것은 쉽지 않다. 따라서, '장바구니 담기'까지 도달하

지 않은 상품에 대해서는 '상품 클릭'에서 '장바구니 담기'까지 도달할 확률(b)'을 구한다. 이는 로지스틱 회귀분석, 인공신경망 분석 등을 통해 수행한다. 이 때 목표변수를 '장바구니 담기'로, '구매여부' 변수를 제외한 나머지 변수들을 입력변수로 하여 분석한다. '장바구니 담기'까지 도달할 확률은 목표변수의 예측값으로 주어진다.

Phase 3 : 각 고객이 클릭한 상품에 대한 선호도를 결정한다. 만일, 한 고객이 어떤 상품에 대해 '장바구니 담기'까지 도달했다가 구매하지 않았다면, 이 고객의 선호도는 p가 된다. 이는 '장바구니 담기'에서 구매에 이를 확률이 p였으므로 이를 선호도로 결정한 것이다. 그러나, '장바구니 담기'까지 도달하지 않은 상품에 대해서는 구매로 이어질 확률인 ($b \times p$)를 선호도로 결정한다(<Figure 3(b)> 참조).

	CD1	CD2	CD3	CD4
Customer 1	1	0	1	
Customer 2		1	0	0
Customer 3	1	0		
Customer 4	0		0	1
Customer 5		0		1

(a) Binary data (purchase vs no purchase)
(Sarwar, et al. 2000a)

	CD1	CD2	CD3	CD4
Customer 1	1	0.15	1	
Customer 2		1	0.82	0.44
Customer 3	1	0.15		
Customer 4	0.82		0.44	1
Customer 5		0.15		1

(b) Preference data
(Kim, et al. 2005)

Figure 3. Customer-Product Preference Level Matrix

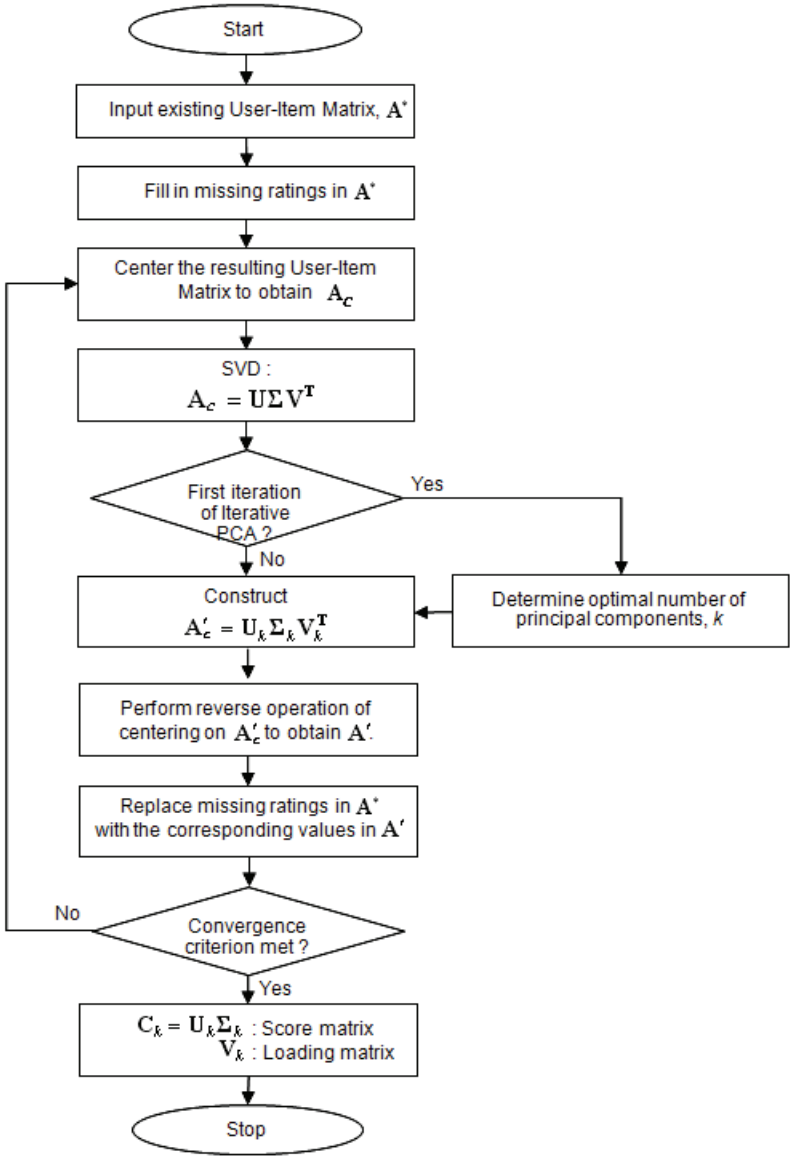

Figure 4. Phase I : Procedures for the iterative PCA(Kim and Yum, 2005)

2.3 차원감소 기법을 이용한 상품 추천

본 논문에서는 제 2.2절에서 구한 고객 선호도 데이터에 다음과 같은 Kim and Yum(2005)의 차원 감소 기법을 적용하였다.

Phase I : 반복적인 특이값 분해 기법을 이용하여 고객 선호도 데이터의 결측치를 추정하고, 동시에 분산의 설명력이 큰 일정 수의 주성분을 찾은 후 각 고객의 주성분 값을 계산한다 (<Figure 4> 참조).

Phase II : 각 고객의 주성분 값을 이용하여 군집분석을 수행함으로써 유사한 성향을 보이는 고객들을 군집화한다.

Phase III : 각 군집의 고객의 데이터를 이용하여 고객이 클릭하지 않은 상품에 대한 선호도를 구한 후, 가장 선호도가 높은 *Top_N* 목록을 각 고객에게 추천한다.

위의 절차를 자세히 기술하면 다음과 같다. 먼저 고객 선호도 데이터 행렬을 $\mathbf{A}^*$라 하자. $\mathbf{A}^*$의 결측치는 해당하는 행과 열의 평균값으로 대체한다. 그리고, 모든 원소에서 해당 열의 평균을 빼고, 그 결과 얻어진 행렬을 $\mathbf{A}_c$라 하자. $\mathbf{A}_c$에 대해 특이값 분해를 수행하면 $\mathbf{A}_c$는 다음과 같이 분해된다.

$$\mathbf{A}_c = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$$

여기서 $m \times m$행렬 $\mathbf{U}$는 $\mathbf{A}_c\mathbf{A}_c^{\mathrm{T}}$의 고유벡터(eigenvector)이며, $n \times n$ 행렬 V는 $\mathbf{A}_c^{\mathrm{T}}\mathbf{A}_c$의 고유벡터이다. 그리고, Σ은 $m \times n$ 대각행렬이 된다.

다음으로 k개의 주성분 개수를 결정하고, Σ의 $k \times k$ 부분행렬을 취한 후, 다음과 같이 $\mathbf{A}_c'$를 구한다.

$$\mathbf{A}_c' = \mathbf{U}_k\Sigma_k\mathbf{V}_k^{\mathrm{T}}$$

$\mathbf{A}_c'$의 각 원소에 각 열의 평균값을 다시 더해 구한 행렬을 $\mathbf{A}'$라고 하고, $\mathbf{A}^*$의 결측치를 $\mathbf{A}'$의 해당되는 값으로 대체한다.

위와 같은 단계를 $\mathbf{A}^*$의 결측치 값이 특정한 값으로 수렴할 때까지 반복 수행한다. 이런 과정을 통해, $\mathbf{A}^*$의 결측치를 추정하게 되고, 동시에 $m \times k$행렬 $\mathbf{C}_k(=\mathbf{U}_k\Sigma_k)$를 이용하여 주성분 값을 계산하게 된다.

그 다음, $\mathbf{C}_k(=\mathbf{U}_k\Sigma_k)$를 이용하여 군집분석을 수행한다. 군집분석 알고리즘으로는 K-means 알고리즘(McQueen, 1967)을 사용하였다. 즉, 고객들의 주성분 값을 이용하여 유사한 성향을 지닌 고객들을 군집화하는 것이다. K-means 알고리즘에서 군집의 개수는 다음 식을 바탕으로 결정한다.

$$\frac{|SS_K - SS_{K+1}|}{SS_1} < c$$

위 식에서 $\mathbf{SS}_K$는 군집의 수가 K개 일 때, 각 군집의 중심에서 군집 내의 관측치들까지의 거리의 제곱합을 의미하며, 군집의 개수를 하나 더 증가시켰을 때, 군집분석 수행 이전의 제곱합에 비해 줄어드는 비율이 c이하라면 더 이상 군집의 개수를 증가시키지 않는다는 것을 의미한다. 본 연구에서는 c를 0.005, 0.01, 0.05로 변화시켜가면서 분석을 수행하였다. c = 0.005일 때는 전체 데이터를 15개의 군집으로, c = 0.01일 때는 10개의 군집으로, c = 0.05일 때는 5개의 군집으로 나누었다.

군집분석 후, $\mathbf{A}^*$의 결측치를 다시 예측한다. 즉, 각 군집에 속한 고객들의 특정 상품에 대한 선호도의 평균을 예측값으로 한다. 모든 결측치를 예측한 후, 가장 높은 선호도를 지닌 *Top_N* 목록을 작성하여 각 고객에게 해당되는 상품을 추천하게 되는 것이다.

3. 실험 평가

3.1 추천 시스템의 성능 평가 방법

추천시스템의 성능을 평가하기 위해 본 연구에서는 KAIST 내에 실험용 전자 상거래 사이트를 구축하여, 음반 판매 쇼핑몰을 49일 간 운영하였다. 이로부터 구한 데이터에 대해 로지스틱 회귀분석, 인공 신경망 분석 등을 적용하여 계산한 고객 선호도 데이터는 <Figure 3(b)>와 같은 형태를 갖는다.

본 연구의 차원 감소 기법과 Kim *et al.*(2005)의 협업적 필터링 기법의 성능을 평가하기 위해 다음과 같은 방법을 사용한다.

단계 1) <Figure 3(b)>와 같은 선호도 데이터에서 구매가 이루어진 1의 값을 갖는 원소 중 5% 또는 10%를 랜덤하게 삭제한다. 즉, 삭제된 원소는 클릭이 이루어지지 않은 상품으로 간주하는 것이다.

단계 2) 5% 또는 10%를 삭제한 데이터에 대해 협업적 필터링과 차원 감소 기법을 적용하여 비어 있는 원소에 해당하는 선호도 값을 예측한다.

단계 3) 각 방법에 의해 생성된 데이터로부터 예측된 선호도가 높은 순으로 *Top_N* 목록을 작성한다. 본 연구에서는 N을 5, 10, 15, 20, 25, 30으로 다양하게 변화시켜 가며 *Top_N* 목록을 작성하였다.

단계 4) 작성된 *Top_N* 목록이 삭제된 상품을 어느 정도 포함하는지 파악함으로써 추천 시스템의 성능을 평가한다. 즉, *Top_N* 목록에 숨겨진 상품이 포함되어 있다는 것은 고객이 구매할 만한 상품을 추천했다는 의미이다. 따라서, *Top_N* 목록이 삭제된 상품을 많이 포함할수록 더 우수한 성능을 지닌 추천 시스템이라 할 수 있다. 이와 같은 평가 방법은 기존 연구(Sarwar

et al., 2000a)에서도 사용되었다.

성능을 평가하는 척도로는 정보 추출분야에서 널리 사용되는 'recall'과 'precision'을 사용한다. 'recall'과 'precision'의 정의는

$$recall = \frac{\sum_{i \in A} | H_i \cap Top_N_i |}{\sum_{i \in A} | H_i |}$$

$$precision = \frac{\sum_{i \in A} | H_i \cap Top_N_i |}{N \times | A |}$$

이며, H_i는 '고객 i 의 숨겨진 상품'을, N은 '각 고객 별로 추천된 상품의 수'를, Top_N은 '고객 i에게 추천된 Top_N 목록'을, A는 '하나 이상의 숨겨진 상품을 지닌 고객들'을 의미한다.

'recall'과 'precision'은 'recall'이 증가하면 'precision'은 감소하는 경향이, 'precision'이 증가하면 'recall'이 감소하는 경향이 있다. 따라서, 이 두 척도의 조화평균인 F1 값이 널리 사용된다 (Sarwar *et al*., 2000a).

$$\mathrm{F1} = \frac{2 \times \mathrm{recall} \times \mathrm{precision}}{\mathrm{recall} + \mathrm{precision}}$$

본 연구에서도 F1값을 이용하여 추천 시스템의 성능을 평가하였다.

3.2 실험결과

Kim *et al*.(2005)의 협업적 필터링을 이용한 추천 시스템과 본 논문의 차원 감소 기법을 이용한 추천 시스템의 성능 비교 결과는 <Table 3>과 같다.

차원 감소 기법을 이용하는 경우, 로지스틱 회귀분석(LR)을 사용하여 고객의 선호도를 추정하는 것이 인공신경망(ANN)을 사용하는 것 보다 일반적으로 높은 예측력을 보였다. 이러한 경향은 '추첨 상품 수(N)', '군집분석 기준값(C)' 또는 '숨겨진 상품 수의 비율' 등과 무관하다. 그리고 군집분석을 수행할 때 기준값으로 0.05를 사용하는 것이 0.005 또는 0.01을 사용하는 것보다 일반적으로 우수한 예측력을 보였다. 즉, 군집의 수를 적게 할수록 우수한 성능을 나타낸 것이다. 이 결과 또한 '추첨 상품 수', '숨겨진 상품 수의 비율' 또는 '선호도 추정 방법' 등에 의존하지 않는다.

협업적 필터링을 이용한 추천 시스템과 차원 감소 기법을 이용한 추천 시스템의 성능을 비교하는데 있어, 차원 감소 기법의 선호도 추정 방법은 로지스틱 회귀분석을, 군집분석의 기준값은 0.05를 사용하는 것으로 하였다. 그 결과, '추천 상품 수'가 5개 또는 10개일 때는 차원 감소 기법을 이용하는 것이 우수한 성능을 나타냈고, 25개 또는 30개일 때는 협업적 필터링을 이용한 추천 시스템이 더 우수한 성능을 보였다. 그리고 '추천 상품 수'가 15개 또는 20개인 경우에는 두 방법의 성능 차이는 나타나지 않았다. 이 결과는 '숨겨진 상품 수의 비율'에 의존하지 않는다.

추천 상품수가 적을 때, 차원감소 기법의 성능이 우수하게

Table 3. Experiment results : F1 values for proposed and conventional approaches

N	*H*	*C*	차원감소 기법 이용 *P*		협업적 필터링 기법 이용(Kim *et al*., 2005)
			ANN	LR	
5	5	0.005	0.042	0.042	0.042
		0.01	0.042	**0.083**	
		0.05	0.063	**0.083**	
	10	0.005	0.031	0.042	0.052
		0.01	0.021	0.063	
		0.05	0.052	**0.083**	
10	5	0.005	0.023	0.023	0.034
		0.01	0.023	**0.045**	
		0.05	0.034	**0.045**	
	10	0.005	0.023	0.023	0.028
		0.01	0.011	0.034	
		0.05	0.028	**0.045**	
15	5	0.005	0.023	0.016	**0.031**
		0.01	0.016	**0.031**	
		0.05	0.023	**0.031**	
	10	0.005	0.016	0.016	**0.031**
		0.01	0.008	0.023	
		0.05	0.020	**0.031**	
20	5	0.005	0.018	0.012	**0.024**
		0.01	0.012	**0.024**	
		0.05	0.018	**0.024**	
	10	0.005	0.012	0.015	**0.024**
		0.01	0.006	0.018	
		0.05	0.015	**0.024**	
25	5	0.005	0.014	0.010	**0.024**
		0.01	0.010	0.019	
		0.05	0.014	0.019	
	10	0.005	0.010	0.014	**0.022**
		0.01	0.005	0.017	
		0.05	0.012	0.019	
30	5	0.005	0.012	0.008	**0.028**
		0.01	0.012	0.016	
		0.05	0.012	0.016	
	10	0.005	0.012	0.012	**0.020**
		0.01	0.008	0.014	
		0.05	0.010	0.016	

나타나는 근거는 다음과 같이 추정된다. 차원 감소 기법을 적용할 경우, 원시 데이터의 정보량이 줄어드는 반면 잡음이 제거되므로, 사용자의 선호도가 높은 소수의 상품에 대해 더욱 우수한 예측력을 지닌 것으로 판단된다. 그러나, 추천 상품의 개수가 증가할수록 데이터의 정보량이 부족한 차원 감소기법의 예측력은 모든 데이터를 사용하는 기존의 협업적 필터링에 비해 저하되는 것으로 보인다.

여러 파라미터가 F1 값에 미치는 효과를 파악하기 위해 본 연구에서 개발한 방법에 대해 분산 분석을 수행하였고, 실험 설계는 완전 요인 배치법(full factorial design)을 사용하였다 (Montgomery, 2000). 본 연구에서 개발한 방법에 포함된 인자는 선호도 예측 방법(P), 추천 상품 수(N), 삭제된 데이터 비율(H), 군집 수를 결정하는 기준값 (C) 등이다. P는 인공신경망 분석(ANN), 로지스틱 회귀분석(LR)의 2수준을, N은 5, 10, 15, 20, 25, 30의 6수준을, H는 5, 10의 2수준을, C는 0.005, 0.01, 0.05의 3수준을 갖는다.

<Table 4>는 차원 감소 기법을 이용한 방법의 분산분석 결과표로서, 모든 주효과와 모든 교호작용이 유의수준 0.05에서 통계적으로 유의함을 나타낸다. 주효과 P와 교호작용 $P \times C$, $P \times H$를 도시한 <Figure 5>~<Figure 7>로부터 예측 방법의 성능을 비교할 수 있다. C가 0.05 또는 0.01을 가질 때 LR이 ANN 보다 우수한 성능을 나타냈으나, 0.005인 경우에는 우수하지 못한 성능을 나타냈으며 (<Figure 6> 참조), H의 수준과는 상관없이 LR이 ANN보다 우수한 성능을 보였다. 그리고, <Figure 5>, <Figure 8>~<Figure 10>으로부터 $N = 5$일 때 가장 높은 F1값을 가짐을 알 수 있다. 또한 C가 0.05를 가질 때 가장 우수한 예측력을 보였다(<Figure 5>, <Figure 10>, <Figure 11> 참조). 즉, 군집의 수가 작을수록 우수한 성능을 나타낸 것이다.

Table 4. ANOVA table for the proposed approach

요인	자유도	제곱합 ($\times 10^4$)	평균 제곱합 ($\times 10^4$)	F	p 값
P	1	16.836	16.836	142.91	0.000
N	5	147.510	29.502	250.42	0.000
H	1	3.115	3.115	26.44	0.000
C	2	16.074	8.037	68.22	0.000
$P \times N$	5	9.196	1.839	15.61	0.000
$P \times H$	1	0.733	0.733	6.22	0.017
$P \times C$	2	10.355	5.177	43.95	0.000
$N \times H$	5	1.792	0.358	3.04	0.021
$N \times C$	10	11.730	1.173	9.96	0.000
$H \times C$	2	1.931	0.966	8.19	0.001
오차	37	4.359	0.118		
Total	71	223.631			

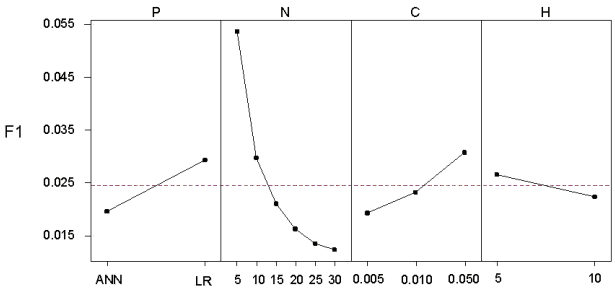

Figure 5. Main effect plots for the proposed approach

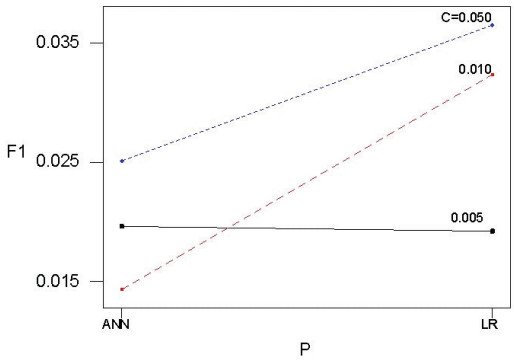

Figure 6. Interaction plot for P and C for the proposed approach

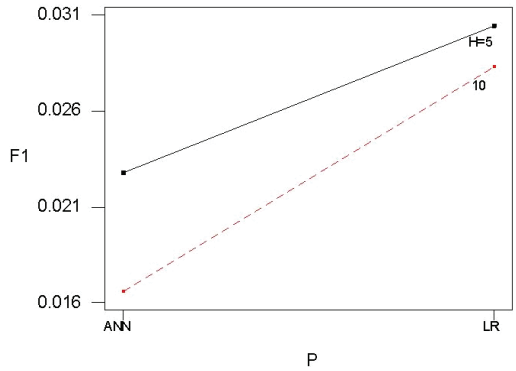

Figure 7. Interaction plot for P and H for the proposed approach

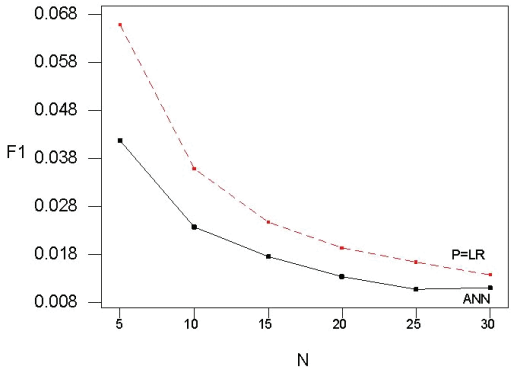

Figure 8. Interaction plot for P and N for the proposed approach

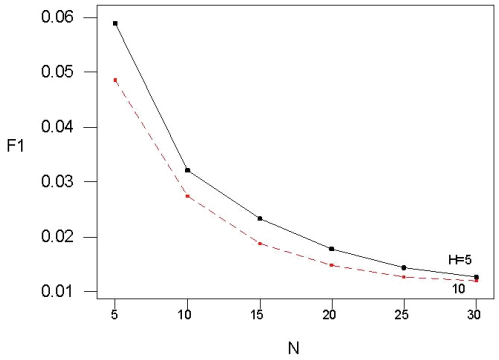

Figure 9. Interaction plot for *H* and *N* for the proposed approach

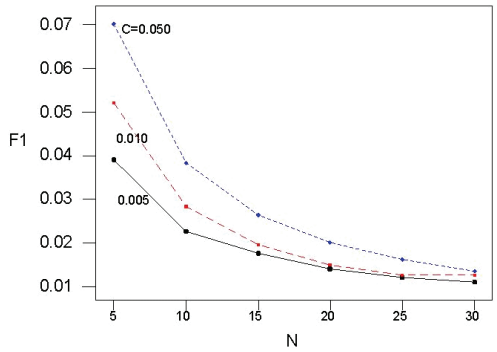

Figure 10. Interaction plot for *C* and *N* for the proposed approach

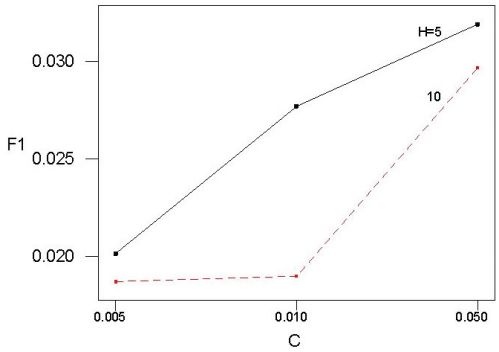

Figure 11. Interaction plot for *C* and *H* for the proposed approach

4. 결 론

본 연구에서는 고객의 행동 및 탐색 패턴으로부터 추정한 선호도, 즉 implicit rating 데이터에 대해 차원 감소 기법을 적용함으로써, 이 기법의 응용범위가 다양한 경우로 확대될 수 있다는 것을 보였다. 차원 감소 기법으로는 특이값 분해와 주성분값을 대상으로 한 군집분석을 사용하였다. 그리고 성능 평가를 위해 차원 감소 기법을 적용했을 때와 기존의 협업적 필터링을 적용했을 때의 예측력을 비교하였다.

실험용 음반 사이트로부터 수집한 데이터를 기준으로 두 접근방법을 비교한 결과, 추천 상품의 수가 지나치게 클 때를 제외하고는 차원 감소 기법을 이용한 추천 시스템의 성능이 Kim *et al.*(2005)에서 이용한 협업적 필터링 기법에 비해 우수한 것으로 나타났다. 차원 감소 기법이 상대적으로 적은 저장 용량을 필요로 하고 높은 계산 속도를 갖는다는 점을 고려할 때, 이는 매우 고무적인 결과라고 판단된다. 그 밖에 본 연구를 통해 확인한 결과는 다음과 같다. 첫째, 차원 감소 기법을 이용한 추천시스템에서, 고객 선호도 데이터를 구성할 때 로지스틱 회귀분석을 이용하는 것이 인공 신경망 분석을 이용하는 것 보다 높은 예측력을 나타냈다. 둘째, 군집분석 시, 전체 데이터를 적은 수의 군집으로 나눌 때 더욱 우수한 성능을 보였다.

본 논문에서는 다소 작은 규모의 실험용 전자 상거래 사이트를 대상으로 추천 시스템의 성능을 비교하였다. 따라서 다양한 형태의 실제 전자 상거래 사이트를 대상으로 한 비교 연구를 통해 본 연구에서 얻은 결론의 일반성을 확인할 필요가 있다.

본 연구에서 정립한 개념과 방법론은 다양한 환경에서도 적용 및 확장이 가능하다. 예를 들어, 기존의 전자 상거래 사이트뿐만 아니라, 고객이 판매자와 소비자로서 참여하는 경매 사이트에도 적용할 수 있으며, 신문과 같은 다양한 형태의 컨텐츠 사이트에도 적용이 가능하다. 아울러, 본 연구의 확장을 통해 시간과 공간을 고려한 고객의 탐색 및 행동 패턴을 바탕으로 모바일 추천 시스템의 개발에 관한 연구도 이루어질 수 있을 것이다.

참고문헌

Berson, A., Smith, K., and Thearing, K. (2000), Building data mining applications for CRM, *New York* : *McGraw-Hill.*

Billsus, D. and Pazzani, M. J. (1998), Learning collaborative information filters, *Proceedings on the Fifteenth International Conference on Machine Learning*, 46-54, Madison, WI.

Claypool, M., Le, P., Wased, M., and Brown, D. (2001), Implicit interest indicators. *Proceedings on the International Conference on Intelligent User Interfaces*, 33-40, Santa Fe, New Mexico.

Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. (2001), Eigentaste : A Constant Time Collaborative Filtering Algorithm. *Information Retrieval Journal*, **4**(2), 133-151.

Hayes, C., Cunningham, P., and Smyth, B. (2001), A case-based reasoning view of automated collaborative filtering, *Proceedingsof the Fourth International Conference on Case-Based Reasoning*, 243-248, Vancouver.

Hill, W., Stead, L., Rosenstein, M., and Furnas, G. (1995), Recommending and evaluating choices in a virtual community of use, *Proceedings of the 1995 ACM Conference on Factors in Computing Systems*, 194-201, New York.

Kelly, D. and Belkin, N. J. (2001), Reading time, scrolling, and interaction : exploring implicit sources of user preferences for relevance feedback, *Proceedings of the Twenty Fourth Annual ACM SIGIR Conference on Research and Development in Information*

Retrieval, 408-409, New Orleans, LA.

Kim, D. and Yum, B.-J. (2005), Collaborative filtering based on iterative principal component analysis. *Expert Systems with Applications*, **28**(4), 823-830.

Kim, Y. S., Yum, B.-J., Song, J., and Kim, S.-M. (2005), Development of a recommender system based on navigational and behavioral patterns of customers in e-commerce sites, *Expert Systems with Applications*, **28**(2), 381-393.

Lawrence, R. D., Almasi, G. S., Korlyar, V., Viveros, M. S., and Duri, S. S. (2001), Personalization of supermarket product recommendations, *Data Mining and Knowledge Discovery*, **5**(1), 11-32.

Lee, J., Podlaeck, M., Schonberg, E., and Hoch, R. (2001), Visualization and analysis of click stream data of online stores for understanding web merchandising, *Data Mining and Knowledge Discovery*, **5**(1/2), 59-84.

Lee, J., Podlaeck, M., Schonberg, E. m Hoch, R., and Gomory, S. (2000), Understanding merchandising effectiveness of online stores, *Electronic Markets*, **10**(1), 1-9.

Montgomery, D. C. (2000), Design and analysis of experiments, New York : Wiley.

Nichols, D. M. (1997), Implicit rating and filtering, *Proceedings of the Fifth Workshop on Filtering and Collaborative Filtering*, 31-36, Budapest.

Rafter, R. and Smyth, B. (2001), Passive profiling from server logs in an online recruitment environment, *IJCAI's Workshop on Intelligent Techniques for Web Personalisation*, 35-41, Seattle, WA.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedle, J. (1994), Grouplens : An open architecture for collaborative filtering of netnews, *Proceedings of the ACM 1994 Conference on Computer Supported Cooperative Work*, 175-186, Chapel Hill.

Sarwar, B., Karypis, G., Konstan J. A., and Riedl, J. (2000a), Analysis of recommendation algorithms for e-commerce, *Proceedings of ACM E-Commerce 2000 Conference*, 158-167, Minneapolis, MN.

Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. (2000b), Application of Dimensionality Reduction In Recommender System-A Case Study, *Proceedings of ACM WebKDD Workshop*, Boston, MA.

Shardanand, U. and Maes, P. (1995), Social information filtering : algorithms for automating word of mouth, *Proceedings of Conference on Human Factors in Computing Systems*, 210-217, Denver, CO.

Yuan, S. and Chang, W. (2001), Mixed-initiative synthesized learning approach for web-based CRM, *Expert Systems with Applications*, **20**(2), 187-200.