한국지능정보시스템학회논문지 제13권 제2호 2007년 6월(pp.69~80)

구매순서를 고려한 개선된 협업필터링 방법론*

조영빈
건국대학교 사회과학대학 경영학과
(ybcho111@kku.ac.kr)

조윤호
국민대학교 경영학부
(www4u@kookmin.ac.kr)

고객의 선호도는 시간에 따라 변화하지만 기존 협업필터링기법(Collaborative Filtering : CF)은 정적인 데이터만을 다룬다. 이는 기존 CF 기법이 특정 기간 동안 고객의 구매 여부만 고려할 뿐 고객의 구매순서를 사용하지 않기 때문이다. 따라서 기존 CF 기법은 고객의 동적인 데이터인 구매순서를 고려함으로써 추천의 품질을 높일 가능성이 있다.

본 연구에서는 고객의 구매순서를 활용함으로써 CF 기법의 추천 품질을 향상시키는 새로운 상품추천 방법론을 제안한다. 즉, 군집분석기법인 자기조직화지도(Self-Organizing Map : SOM)를 활용하여 고객의 구매순서를 파악한 후 연관규칙탐사(Association Rule Mining : ARM)를 사용하여 고객들의 구매순서 중 일정 정도의 통계적인 타당성을 갖는 구매순서 패턴을 찾아내어 이를 추천 시에 활용한다.

대형 백화점의 구매자료에 적용하여 제안한 방법론의 효과성을 실험한 결과 제안한 방법론이 기존 CF 기법보다 우수한 추천품질을 가지고 있음이 실증적으로 확인되었다.

논문접수일 : 2007년 01월 게재확정일 : 2007년 05월 교신저자 : 조윤호

1. 서론

상품추천시스템은 고객관계관리분야의 연구자나 실무자들이 지속적으로 주목하고 있는 연구분야이다. 상품추천시스템이 집중적으로 채택되고 있는 산업은 소매업이라 할 수 있는데, 이는 소매시장이 특정한 기간 동안 반복구매가 일어나고, 많은 고객이 있으며, 과거 고객의 구매기록으로부터 많은 정보를 얻을 수 있기 때문이다(Schmittlein and Peterson, 1994).

최근까지 다양한 추천시스템이 개발되었다(Balabanović and Shoham, 1997; Basu, Hirsh and Cohen, 1998; Hill et al., 1995; Lawrence et al., 2001; Resnick et al., 1994; Sarwar et al., 2001; Shardanand and Maes, 1995). 이중 협업필터링 기법(CF: Collaborative Filtering)은 가장 성공적인 방법으로 알려져 있고, 웹페이지, 영화, 논문, 신문기사 추천 등의 다양한 적용사례를 갖고 있다(Hill et al., 1995; Resnick et al., 1994; Shardanand and Maes, 1995; 김재경 외, 2006, 2005; 신택수 외, 2006; 김룡 외, 2005). 협업필터링 기법은 목표고객과 가장 비슷한 구매 선호도를 가진 고객을

* 본 논문은 2006년도 국민대학교 교내연구비를 지원받아 수행한 연구임.

찾고 그 고객의 구매 제품 중 목표고객이 구매할 가능성이 가장 큰 제품을 추천하는 방식으로 작동한다. 소매업에서 협업필터링을 기반으로 하는 추천시스템은 다음과 같은 방법으로 제품을 추천하게 된다(Sarwar et al., 2000, 2001).

1) 고객 프로파일 생성(Customer profile construction)

특정 기간 동안 임의의 고객의 구매기록은 해당 고객의 상품 선호도를 나타내는 고객 프로파일을 구성하는데 사용된다. 고객 프로파일은 통상적으로 **P**로 표현되는데 **P의** 원소 p_{ij}는 고객 i가 제품 j를 구매하면 1, 그렇지 않으면 0으로 표시된다.

2) 유사 선호고객(Neighborhood) 형성

이 과정은 협업필터링 기반의 추천시스템에서 가장 중요한 부분이다. 이 과정에서 목표고객과 유사한 과거 구매 행태를 보인 유사 선호고객(neighborhood)들을 찾는다. 목표고객의 프로파일과 개개 고객의 프로파일간의 상관관계를 계산하여 유사 선호고객을 찾게 된다. 일반적으로 목표고객과 유사 선호고객간의 유사도에 따라 K개의 유사 선호고객을 찾게 되는데, 이를 K크기의 유사선호고객의 집합이라고 한다.

3) 추천 제품 도출(Recommendation generation)

일단 목표고객에 대한 유사 선호고객이 결정되면, 추천 시스템은 유사 선호고객들이 구매한 제품이면서, 목표고객이 아직 구매하지 않은 상품 중 목표고객이 구매할 가능성이 높은 N개의 상품을 추출하여 추천하게 된다.

그렇지만 기존의 통상적인 협업필터링 기법은 목표 고객의 유사 선호고객을 결정하는데 있어, 특정 기간 동안 고객들의 구매여부에 대한 정보만을 이용할 뿐, 고객들의 구매 순서에 관한 정보를 활용하지는 않는다. 따라서 기존 방법은 정태적이라 할 수 있다. 그러나 실제적으로 소매 시장의 고객들은 정태적이라고 하기 어렵고, 시간의 흐름에 따라 구매 선호도가 바뀐다. 고객들의 구매순서를 활용한다면 기존 협업필터링 방법보다 추천 정확도를 제고할 수 있게 될 것이다.

고객 구매순서를 추천에 반영하기 위해서는 고객의 프로파일을 시간에 따라 재 분류해야 하는데, 이렇게 하면 데이터의 희박성(spasity)을 악화시키는 심각한 문제에 봉착하게 된다. 왜냐하면 구매순서를 반영한 고객 프로파일은 시간차원을 추가하는 형태로 이루어지기 때문이다. 데이터 희박성 문제는 상품추천 분야 연구자에게 잘 알려진 문제로, 데이터의 희박성이란 0이 아닌 데이터가 거의 없는 데이터 집합으로 상품 추천의 정확성을 떨어뜨리는 주요인으로 알려져 있다(Mobasher et al., 2001). 따라서 데이터 희박성을 완화시킬 새로운 방법론이 필요하게 된다.

데이터 희박성을 완화하는 방법으로 차원감소(dimensionality reduction) 기법이 많이 사용되었다(Cho et al., 2004; Sarwar et al., 2000). 본 연구에서는 고객의 유사한 거래를 서로 묶는 군집화 기법으로서, 최근 들어 빈번히 사용되는 SOM(Self-Organizing Map) 기법을 사용한다. 연구의 기본적인 아이디어는 SOM 기법을 사용하여 고객들의 모든 거래정보를 클러스터에 배정하고, 거래정보를 해당 클러스터의 번호로 변환한다. 이렇게 하면 고객의 모든 거래정보에 있는 구매순서 정보는 부여된 클러스터 번호 순서로 변환되어 나타내게 된다. 이러한 클러스터 번호의 변화를 관찰하면 고객의 구매순서를 파악할 수 있게 되고, 이 정보는 목표고객의 미래 구매를 예측할 수 있는 잠재정보가 된다. 그러나 모든 구매 순서가 통계적인 타당성을 갖지 않기 때문에 연관규칙탐사(Association Rule Mining) 기법을 사용하여 고객들의 구매순서 중 일정 정도의 통계적인 타당성을 갖는 순서패턴을

찾아낸다. 그런 다음 목표고객이 가장 구매할 가능성이 높은 제품을 찾아내게 된다.

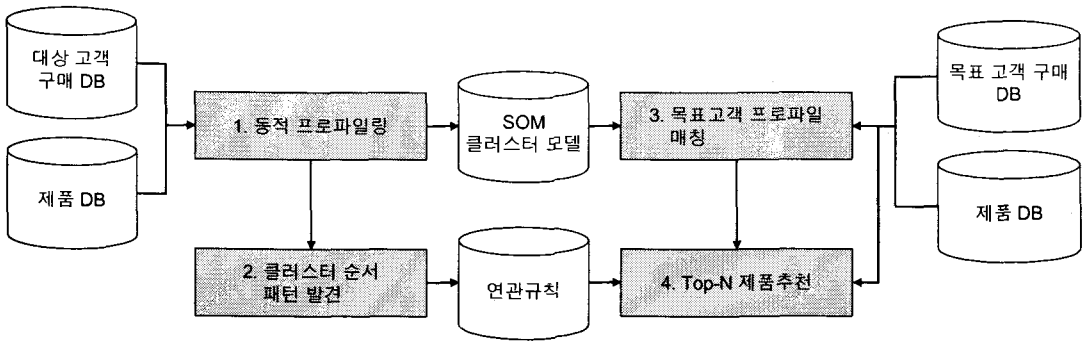
본 논문은 다음과 같이 구성된다. 제 2장에서는 본 연구에서 제안한 방법론의 절차와 내용을 설명한다. 제 3장에서는 제안한 방법론을 실제 현장에서의 데이터에 적용하고 기존 협업 필터링 기법과 비교하여 제안 방법론의 효과를 제시한다. 제 4장에서는 연구의 결론과 향후 연구방향 및 내용을 기술한다.

2. 제안 방법론

2.1 전체적인 절차

본 연구의 문제는 소매업의 고객을 대상으로 상품을 추천하는데 있어 고객의 구매순서를 반영한 방법론을 개발하는 것이다. 일반적으로 소매업에서의 상품 추천은 사전에 정해진 기간 동안(예를 들어 3개월, 혹은 6주 등)의 거래정보를 분석하여 이루어지게 된다. 본 연구에서도 고객들의 구매순서를 파악하기 위해 T시점을 포함한 l 기간 동안의 거래정보를 바탕으로 하고, T시점에 목표고객이 구매할 제품을 추천하는 것으로 가정한다. 그러면 본 연구에서의 문제는 "목표고객의 T시점 이전 l-1기간 동안의 구매정보와 구매순서 정보가 있을 때, T시점에 목표고객이 구매할 가능성이 높은 제품이 무엇인지 찾는 것"으로 정리된다.

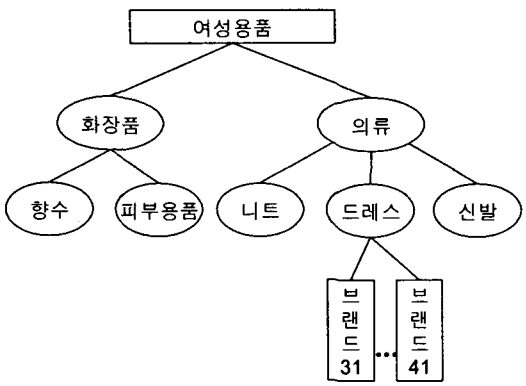
이 문제를 풀기 위하여 제안한 방법론은 [그림 1]에서와 같이 동적 프로파일링, 클러스터 순서 패턴 발견, 목표고객 프로파일 매칭, Top-N 제품 추천 등 크게 네 가지 단계로 구분된다. 동적 프로파일링, 클러스터 순서 패턴 발견에서는 고객의 거래 DB에서 신뢰할만한 상품추천 모델을 만들게 된다. 상품 추천모델은 동적 프로파일링을 통하여 산출된 SOM 클러스터 모델과 클러스터 순서패턴 발견에서 나온 연관규칙으로 구성된다. 여기서의 상품추천모델은 목표고객의 구매자료를 이용하여 추천한 상품을 찾아내는데 사용된다. 목표고객 프로파일 매칭, Top-N 제품 추천에서는 SOM 모델을 이용하여 목표고객의 거래정보를 군집화하면서 시작된다. 다음으로는 목표고객의 구매순서를 연관규칙 DB에 저장된 구매 순서 패턴과 비교하게 되고, 마지막으로 가장 유사한 구매패턴을 파악한 후 목표 고객이 구매할 가능성이 가장 높은 N개의 제품을 찾게 된다.



[그림 1] 전체적인 추천절차

2.2 동적 프로파일링(dynamic profiling)

이 절에서 기술할 동적 프로파일링은 상품추천 모델을 만드는데 필요한 거래정보 군집화(transaction clustering)에 관한 내용이다. 거래정보를 군집화하는 이유는 데이터 집합의 차원을 줄여 데이터 희박성 문제를 완화하기 위해서이다. 그런데 데이터 희박성을 완화하기 위하여 사용할 수 있는 실무적인 방법이 있는데, 이는 소매업에서 통상적으로 운영하고 있는 제품 계층구조(product hierarchy)이다. [그림 2]는 대규모 백화점의 여성용품 매장에서 운용하고 있는 제품 계층구조의 예를 보여주고 있다. 이러한 제품의 그룹화는 상품추천분야의 데이터 희박성(sparsity) 문제를 해결하는 좋은 방법으로 알려져 있다(Cho et al., 2002, 2004; Lawrence et al., 2001).



[그림 2] 제품 계층구조의 예

제품 종류의 집합을 P라고 할 때 n개의 서로 다른 세부 종류가 있고, 개개의 세부 종류는 하위 계층의 제품군으로 구성되게 된다. 이 과정을 계속하면 결국은 개개의 제품 브랜드가 된다. 이러한 점을 반영하여 제품 종류의 집합은 다음과 같이 표시될 수 있다.

$$\mathrm{P}=\{P_1, P_2, \cdots, P_n\}. \quad (1)$$

A를 m명의 고객들이 T시점 이전 l기간 동안의 거래정보로 이루어진 집합이라고 할 때, 집합 A는 다음과 같이 구성될 수 있다.

$$\mathrm{A}=\{A_{1,T-k}, A_{2mT-k}, \cdots, A_{m,T-k}\},\quad k=0, 1, \cdots, l-1, l \le 2, \quad (2)$$

여기서 $A_{j,T-k} \in$ A이고, P의 공집합이 아닌 부분집합이다.

각각의 $A_{j,T-k}$는 고객 j가 T$-k$ 시점에 구매한 제품 혹은 제품 종류를 나타낸다. 또한 $A_{j,T-k}$는 상품 추천에 적당한 입력형식을 위하여 0-1의 비트 벡터로 변환된다. 이렇게 시간 흐름에 따라 변환된 고객의 거래정보는 해당 고객의 구매이력이 되며, 이는 고객의 동적 고객프로파일(dynamic customer profile)로 정의된다.

[정의 1] 동적 고객 프로파일(Dynamic customer profile)

$\overline{\mathrm{A}}$를 동적 고객 프로파일 이라 하자. 그러면 $\overline{\mathrm{A}}$는 l 기간 동안 n개의 제품 종류와 m명의 고객들로 구성된, 다음과 같은 서열화 매트릭스로 정의된다:

$$\overline{A}_{j,T-k}=\left\langle P_1^{A_{j,T-k}}, P_2^{A_{j,T-k}}, \cdots, P_n^{A_{j,T-k}}\right\rangle$$
$$j=1, 2, \cdots, m,\ k=1, 2, \cdots, l-1,\ l \ge 2, \quad (3)$$

여기서 $P_i^{A_{j,T-k}}=\begin{cases}1, \text{if } P_i \in A_{j,T-k}\\ 0, \text{otherwise}\end{cases}$

SOM 모델을 이용하여 고객의 거래정보를 군집화 하게 되면 다음과 같은 q개의 클러스터로 이루어진 집합을 이루게 된다.

$$\mathrm{C}=\{C_1, C_2, \cdots, C_q\}, \quad (4)$$

여기서 각각의 C_i는 식 (3)의 $\overline{A}$의 부분집합.

각각의 클러스터는 오로지 비슷한 패턴을 가진 거래기록들의 집합을 나타낸다. 각 고객별, 시간별로 클러스터를 분류하면 개별고객의 동적 행태를 나타낼 수 있다. 이를 개념화하기 위하여 다음과 같은 정의가 요구된다.

[정의 2] 고객행태궤적(Customer Behavior Locus) L_i를 고객 i의 구매행태 궤적이라 하자. 그러면 구매행태궤적 L_i는 l기간 동안 고객 i가 옮겨 다닌 클러스터 숫자와 같게 되고, 다음과 같이 정의된다.

$$L_i = \{C_{i,T-l+1}, \cdots, C_{i,T-1}, C_{i,T}\}, \quad i = 1, 2, \cdots, m, \tag{5}$$

여기서 $C_{i,T-k} \in C,\ k = 0, 1, 2, \cdots, l-1,\ l \geq 2.$

구매행태 궤적을 찾는 프로세스는 거래 클러스터링을 통하여 손쉽게 행해질 수 있다. 다음 예제는 고객의 구매행태 궤적의 예를 보여준다.

2.3 클러스터 순서 패턴 도출

모든 고객은 이전 거래에 기반한 구매행태 궤적을 갖는다. 그러나 통계적인 타당성이 떨어지는 구매행위 궤적은 목표고객의 구매행태를 예측하는데 사용되는 규칙을 도출하는데 부적절할 수 있다. 왜냐하면 그러한 구매행위 궤적은 특정 고객의 구매행위만을 반영하기 때문이다. 연관규칙기법(association rule)은 자동화된 필터링 기능을 갖고 있으면서 신뢰성이 높은 규칙을 도출할 수 있기 때문에 이러한 상황에서 사용하기에 적절한 기법이라 할 수 있다(Agrawal et al., 1993).

원래 연관규칙 기법과는 달리 우리는 조건부와 결과부를 구분하여 사용한다. 연관규칙의 조건부는 공식 (5)의 왼쪽부분으로 $T-1$시점까지의 구매행위 궤적으로 $\{C_{i,T-l+1}, \cdots, C_{i,T-1}\}$, 결과부는 오른쪽 부분으로 T시점의 구매행위 클러스터로 $C_{i,T}$와 같이 표현된다. 이렇게 구분하는 이유는 목표고객의 과거행적에 비추어 볼 때 시간 T에 어떤 클러스터에 있을지를 예측하려 하기 때문이다. 더 나아가 기존 연관규칙 기법은 특정 사용자에게 부적절한 규칙을 너무 많이 제공하는 단점이 있다(Lin et al., 2002). 따라서 본 연구에서는 이러한 단점을 없애기 위하여 Wang et al.(2003)의 연구에서 제시한 바 있는 목적 지향형 연관규칙(goal-oriented association rule)을 사용한다.

최소 지지도와 신뢰도이상의 연관규칙을 R_j라 하면 R_j는 다음과 같이 나타낼 수 있다.

$$R_j : r_{j,T-l+1}, \cdots, r_{j,T-1} \Rightarrow r_{j,T}(\text{지지도}_j\ \text{신뢰도}_j), \tag{6}$$

여기서 $r_{j,T-k} \in C$ or ϕ, and $r_{j,T} \in C$.

규칙 R_j는 고객의 과거 구매행위 궤적이 $r_{j,T-l}$, $r_{j,T-l+1}$, $\cdots$, $r_{j,T-1}$라면 T시간에는 그 고객의 클러스터가 $r_{j,T}$임을 의미한다.

2.4 목표고객 프로파일 매칭

이 단계에서 주어진 목표고객에 대해서 우리는 그들의 동적 구매행위에 가장 잘 맞는 제품을 찾게 된다. 목표고객들의 거래들은 이전 단계에서 제시한 바와 같이 SOM 모델을 사용한 구매행위 궤적으로 변환된다. 변환 후 연관규칙 베이스에 저장된 최적 구매행위 궤적을 찾게 되고, 목표고객이 구매할 가능성이 가장 높은 제품을 찾는다.

구매행위 궤적의 예측은 목표고객 거래들을 SOM 모델에 입력하면서 시작된다. 임의의 목표고객의 T시간 이전 l기간 동안의 구매행위 궤적은 SOM 모델에 목표고객의 거래를 입력하면 알 수

있게 된다. 목표고객의 클러스터 궤적은 다른 고객들의 궤적으로부터 도출된 순차패턴(sequential pattern)과 비교되게 되고, 최적의 궤적을 찾게 된다. 이러한 프로세스를 수행하기 위해서는 목표고객의 구매행위궤적과 모델베이스의 순차규칙간의 유사도를 측정하기 위한 새로운 척도가 필요할 것이다.

이들 둘 사이의 유사도는 목표고객의 구매행위 궤적이 모델베이스의 조건부와 얼마나 비슷한 가를 나타낸다. 여기서 우리는 유사도 척도를 다음과 같이 정의한다.

[정의 3] 유사도 척도

L_i^C를 목표고객 i의 $l-1$기간 동안의 구매행위 궤적이라 하고, R_j^C를 모델베이스에 저장되어 있는 연관규칙 j의 조건부라고 하자. 그러면 구매행위 궤적과 연관규칙의 조건부는 각각 $L_i^C=\{C_{i,T-l+1}, \cdots, C_{i,T-1}\}$, $R_j^C=\{r_{i,T-l+1}, \cdots, r_{i,T-1}\}$와 같이 표현될 수 있다. SM_i^j를 L_i^C와 R_i^C의 유사도 척도라 하면 SM_i^j는 다음과 같이 정의된다.

$$SM_i^j=\sum_{k=1}^{l-1} S_{i,T-l}^j,\ l=1, 2, \cdots, l-1, \qquad (7)$$

여기서 $S_{i,T-k}^j=\begin{cases}1, \text{ if } C_{i,T-k}=r_{j,T-k}\\ 0, \text{ otherwise}\end{cases}$.

위의 정의는 목표고객 i의 구매행위 궤적이 연관규칙 j의 조건부와 같은 시기에 같다면, $S_{i,T-k}^j$는 1이고 그렇지 않으면 0임을 나타낸다. SM_i^j가 크면 클수록 목표고객의 구매행위 궤적과 연관규칙의 조건부의 유사도가 높아짐을 알 수 있다.

그러나 유사도 척도가 크다고 할지라도 목표고객이 다음에 어떤 클러스터에 있을지를 예측하는 것은 쉽지 않다, 왜냐하면 순차규칙의 지지도와 신뢰도가 매우 낮을 경우 그러한 규칙의 일반화가 어렵기 때문이다. 따라서 목표고객의 궤적과 순차규칙의 조건부사이의 적합도(fitness)를 측정하는 것이 바람직하다. 적합도 척도는 다음과 같이 정의된다.

[정의 4] 적합도 척도

FM_i^j를 목표고객 i의 구매행위 궤적과 모델베이스의 연관규칙 j사이의 적합도 척도라 하자. 그러면 FM_i^j는 다음과 같이 정의된다.

$$FM_i^j=SM_i^j\times \text{지지도}_j\times \text{신뢰도}_j \qquad (8)$$

이와 같은 정의를 이용하여 우리는 T 시간에 목표고객의 클러스터를 최대 FM_i^j를 가진 순차규칙 j의 결과부 $r_{j,T}$로 찾을 수 있다.

2.5 Top N 제품의 추천

마지막 단계는 시간 T에 목표고객이 갈 것으로 예측되는 클러스터에서 목표고객이 구매할 가능성이 높은 N개의 제품을 추천하는 Top N 추천이다. 각각의 목표고객에 대하여 우리는 그 목표고객이 구매할 가능성이 높은 N개의 제품 리스트를 찾아내게 된다. 특정한 목표고객을 위한 추천은 목표고객의 구매 데이터베이스로부터 도출하고, 시간 T에 목표고객의 클러스터에 속한 거래품목 중 가장 매출량이 많은 품목 중 상위 N개의 제품을 추천한다.

C*를 시간 T의 목표고객의 예측 클러스터라 하자. 그러면 C*는 시간 T이전 l기간 동안에 발생한 거래를 포괄하게 된다. 또한 목표고객이 이전에 구매한 제품을 포함할지도 모른다. 여기서의 제품은 개개 제품뿐만 아니라 제품 카테고리 수준을 포함한다. 다시 말하여 제품 계층도에서의 말단노드를 의미한다. 따라서 목표고객에게 Top N개의 제품을 추천하는데 있어 적합한 거래들을 골라내

어야 한다. 우리는 시간 T의 클러스터에 속한 거래만을 선택했다. 왜냐하면 목표고객이 구매할 가능성이 높은 제품은 시간 T에 C*에 속한 다른 고객이 구매한 상품이기 때문이다. 마지막으로 이전에 구매한 상품은 추천리스트에서 제외했는데, 이는 목표고객의 상품구매패턴을 좀 더 확대시키려는 의도에서였다.

우리는 해당 클러스터에 속한 제품 중 가장 구매빈도가 높은 제품을 목표고객에 대한 TOP N 추천리스트로 결정한다.

[정의 5] 목표고객에 대한 추천리스트

$MF(r_1)$을 시간 T에 C*에 있는 최빈 구매제품(most frequently purchased product) 이라 하자. 유사하게 $MF(r_2)$를 두 번째로 구매빈도가 높은 제품, $MF(r_N)$를 N번째 구매빈도가 높은 제품이라 하자. 그러면 목표고객에 대한 추천리스트는 $MF(r_1)$, $MF(r_2)$, $\cdots$ $MF(r_N)$로 구성되고, $MF(k)$는 다음과 같이 계산된다.

$$MF(k) = \sum_{j \in C^*} P_{ik}^{A_{k,T}} \times N_{ik}^{T}, \qquad (9)$$

여기서 $P_{ik}^{A_{j,T-k}} = \begin{cases} 1, \text{if } P_{ik} \in A_{j,T} \\ 0, \text{otherwise} \end{cases}$, N_{ik}^{T}는 제품 P_{ik}의 시간 T동안의 판매량이고, P_{ik}는 제품클래스 i의 k^{th} 말단 제품.

3. 성능 평가

3.1 평가 방법

우리는 제안한 방법론의 효과를 검증하기 위하여 실제 데이터를 사용하였다. 이 실험에 사용된 데이터는 H 백화점에서 판매된 여성용품에 대한 거래 레코드들이다. 2000년 5월에서 12월까지 8개월 동안의 거래 레코드들이다. 데이터 구조는 통상적인 거래데이터로 ID, 거래 일시, 거래시간, 제품명 등의 필드로 구성되어 있다. 고객들의 동적인 구매행위를 파악하기 위하여 사용된 데이터의 양은 18,843레코드이고 총 1,833명의 고객이 557종류의 상품을 거래하였다. 추천대상 고객은 고객의 동적인 구매행위를 파악하기 위하여 최근에 구매하고 자주 구매하는 단골 고객으로 한정하였다. 데이터의 분석단위는 월 단위로 설정하였는데, 이는 대상상품이 여성용품으로 매일 혹은 매주 구매하는 제품이 아니기 때문이다. 해당 백화점의 전문가의 인터뷰를 통하여 추천대상이 되는 단골 고객은 매달 적어도 한번은 구매를 하고 4개월 연속 구매를 하는 고객으로 정할 수 있었다. 310명의 고객이 여기에 해당되었으며, 이들이 구매한 제품의 종류는 총 고객이 구매한 제품과 같았다.

H 백화점에서 사용된 제품 계층도는 3단계로 구성되어 있다. 톱 레벨은 10개의 제품 클래스로 구성되어 있고, 그 다음 레벨은 25개의 클래스, 맨 마지막 레벨은 557개의 제품으로 구성되어 있다.

2000년 5월에서 8월까지의 기간을 모델 구축을 위한 모델구축기로, 9월에서 12월을 구축된 모델을 추천에 사용하는 시험기로 설정하였다. SOM 모델은 모델 구축기의 데이터에 적용되었고, 매월 자료를 4개의 fold로 구분하여 모델을 추정하는 4-fold 교차검증이 행해졌다. 또한 17개의 서로 다른 SOM 모델 클러스터에 대해서 클러스터의 숫자와 형태가 추천의 정확도에 미치는 영향을 시험하였다. 실제 적용에 있어서는 가장 정확도가 높은 SOM 모델을 정하고 이에 따라 목표고객에 대한 추천을 하는 것이 바람직하다. 그렇지만 본 연구에서는 SOM 모델에서 클러스터의 숫자가 추천의

정확도에 미치는 영향이 크다는 것을 보이기 위하여 여러 가지 클러스터에 대하여 실험을 진행하였다.

또한 최소 지지도를 2%로, 최소 신뢰도를 50%로 설정하였다. 이는 유사 연구인 Lawrence et al. (2001)에서의 제시한 것보다 더 높은 수치이다. 추천상품의 수는 10개로 한정하였다. 시험기간 동안 매달 구매한 목표고객의 숫자는 132명으로 이들을 목표고객의 집합으로 설정하여 추천하였다.

추천 집합의 품질을 평가하기 위하여 재현율(recall), 정확율(Precision)과 같은 상품 추천분야에서 자주 사용되는 지표를 사용하였다(Basu et al., 1998; Billsus and Pazzani, 1998; Lin et al., 2000, 2002; Sarwar et al., 2000). 재현율은 시간 T에 목표고객이 실제로 구매한 상품 중 추천한 상품의 비율로 정의된다. 이에 비해 정확율은 추천한 상품 중 목표고객이 구매한 상품의 비율로 정의된다. 재현율은 고객의 구매리스트 중 얼마나 많은 추천상품이 있는지를 측정하는데 비해, 정확율은 추천 리스트 중 얼마나 많은 상품을 구매했는지를 측정한다. 이들 평가지표는 계산이 간단하고 쉽게 이해되는 장점이 있는 반면, 추천 집합의 크기가 커지면 재현율은 올라가고 정확율을 떨어지는 경향을 나타내기 때문에 혼란스러울 때가 있다(Sarwar et al., 2000). 그래서 이들 지표를 동일한 가중치로 결합한 F1 척도(Billsus and Pazzani, 1998; Rijsbergen, 1979; Sarwar et al., 2000, 2001)도 자주 사용되는데 본 연구에서도 F1을 사용한다. 각각의 평가지표는 다음과 같이 정의된다.

$$\text{정확률} = \frac{\text{적중된 제품 수}}{\text{총 추천한 제품 수}} \tag{10}$$

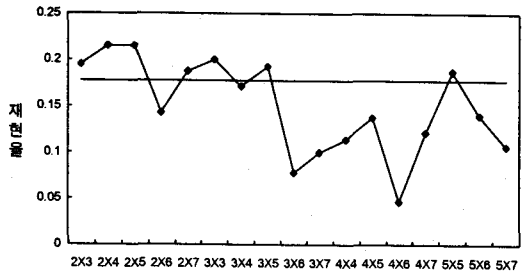
$$\text{정확률} = \frac{\text{적중된 제품 수}}{\text{구매한 제품 수}} \tag{11}$$

$$F1 = \frac{\text{Recall} \times \text{Precision}}{(\text{Recall} + \text{Precision})/2} \tag{12}$$

3.2 분석 결과

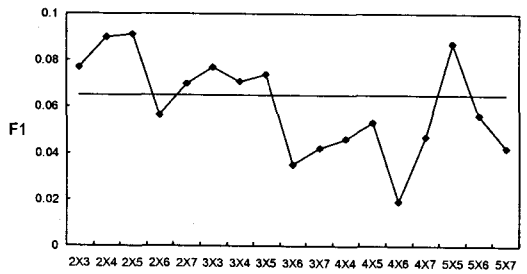
우리는 클러스터의 수가 변화함에 따른 추천의 효과성을 이전에 제시한 평가지표를 이용하여 평가하였다. 이에 덧붙여 본 연구에서 제시한 방법론의 정확도를 성능이 우수하다고 알려진 대표적인 CF 기법(Sarwar et al., 2000)과 비교하였다. [그림 3]은 실험결과를 나타낸다. 3개의 평가지표가 기존 CF 기법보다 좋은 성과를 보이는 경우는 6(2×3), 8(2×4), 9(3×3), 10(2×5), 14(2×7), 15(3×5), and 25(5×5)이었다. 이에 비하여 12(2×6, 3×4), 16(4×4), 18(3×6), 20(4×5), 21(3×7), 24(4×6), 28(4×7), 30(5×6)의 경우는 기존 기법보다 더 열악한 성과를 보였다. 12(3×4) 클러스터의 경우는 재현율은 벤치마크 CF 기법보다 낮았으나, 나머지 두 개 지표는 더 우수했다.

우리는 클러스터의 숫자가 증가하면 정확도가 높아지거나, 그 반대의 경우 등의 클러스터 숫자의 변화에 따른 일반적인 경향을 발견할 수 없었다. 이러한 결과를 보건대 클러스터의 숫자는 Top N 추천 방법론의 정확도에 영향을 미친다는 것을 알 수 있었다. 직관적으로 볼 때 이러한 결과는 상당히 이성적인 것으로 판단된다. 왜냐하면 모든 군집화 방법론이 클러스터의 숫자에 따라 정확도가 달라짐을 발견하고 있고, 최적 클러스터 숫자를 발견해내는 것이 미해결 영역이기 때문이다(Nour and Madey, 1996). 이러한 결과는 SOM 모델이 목표고객의 구매행위 궤적을 얼마나 설명하느냐에 반영하는 것으로 보인다. 군집화 방법론의 과도적합문제를 방지하기 위해서는 고객의 구매행위 궤적을 정확하게 설명할 수 있는 SOM 클러스터의 수를 결정하는 것이 필요하다.



(a) 재현율(recall)

(b) 정확율(precision)



(c) F1

[그림 3] 제안된 방법론과 기존 CF방법론의 비교 (직선은 최상 성능의 CF방법론의 결과)

<표 1>은 벤치마크 CF 기법에 대한 SOM 클러스터 각각에 대한 양측 t-검정의 결과를 보여준다. 8(2×4), 10(2×5)와 25(5×5) 클러스터의 경우 F1과 정확율은 벤치마크 CF보다 통계적으로 유의한 것으로 나타났다. 제안된 방법론은 기존 CF 방법론 보다 클러스터의 수가 적절히 선택된다면 더 우수한 성과를 보이는 것으로 나타났다. 제안된 방법론 중 가장 좋은 성과를 보인 10(2×5) 클러스터의 경우 벤치마크 CF 기법에 비하여 F1과 정확율이 평균적으로 각각 40%, 52% 향상된 것으로 나타났다.

<표 1> 클러스터 숫자에 따른 t-양측검정

	F1	재현율	정확율
SOM2X3	+	+	+
SOM2X4	+(*)	+	+(*)
SOM2X5	+(*)	+	+(**)
SOM2X6			
SOM2X7	+	+	+
SOM3X3	+	+	+
SOM3X4	+		+
SOM3X5	+	+	+
SOM3X6			
SOM3X7			
SOM4X4			
SOM4X5			
SOM4X6			
SOM4X7			
SOM5X5	+(*)	+	+(*)
SOM5X6			
SOM5X7			

*: $p < 0.05$, **: $p < 0.01$.

4. 결론

고객들의 선호도는 시간에 따라 변화한다. 이 연구에서 우리는 상품추천의 정확도를 향상시키기 위하여 시간에 따라 변화하는 고객의 선호도를 반영하는 방법론을 제안하였다. 또한 시간에 따라 변화하는 고객의 선호도를 측정하는 방법과 여러 가지 문제에 따른 답을 도출해내었다.

몇몇 가능한 확장 연구가 있을 수 있다. 이 연구의 결과로부터 우리는 어떤 상품을 추천하면 고객이 구매할 가능성이 큰지를 알 수 있었다. 그렇지만 언제 추천하면 구매가능성이 높은지는 알 수 없다. 이 부분에 대한 추가연구가 있을 수 있을 것이다. 고객의 과거 구매행태에 대한 좀더 상세한 연구를 한다면 언제 상품 추천에 적합한지를 알 수 있게 된 것이다. 덧붙여 모든 모델기반의 연구는 시간이 지남에 따라 구축된 모델의 정확도가 떨어지게 된다. 이는 과거의 행태를 분석하여 미래를 예측한다는 방법론의 한계이기도 하다. 변화하는 고객의 선호도에 따라 모델을 언제 재구축하여야 하는지, 어떻게 모델을 동적으로 관리할 것인지에 대한 추가연구도 필요할 것이다.

참고문헌

[1] 김재경, 이희애, 안도현, 조윤호, "설명기능을 추가한 협업 필터링 기반 개인별 상품 추천 시스템-WebCF-Exp", *경영학연구*, 35권 2호 (2006), 493~519.

[2] 김재경, 안도현, 조윤호, "개인별 상품추천 시스템 WebCF-PT : 웹마이닝과 상품계층도를 이용한 협업필터링", *경영정보학연구*, 15권 1호(2005), 63~79.

[3] 신택수, 장근녕, 박유진, "선호도 추정모형과 협업필터링 기법을 이용한 고객추천 시스템. 선호도 추정모형과 협업필터링 기법을 이용한 고객 추천시스템", *한국지능정보시스템학회논문지*, 12권 4호(2006), 1~14.

[4] 김룡, 부종수, 홍종규, 박원익, 김영국, "시간 스키마 기법 2단계 클러스링 적용 추천시스템의 성능향상", *한국컴퓨터종합학술대회*, 32권 1(B)호(2005), 205~207.

[5] Agrawal, R., T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases", Proceedings of the ACM SIGMOD Conference on Management of Data, (1993), 207~216.

[6] Balabanović, M. and Y. Shoham, "Content-based, collaborative recommendation", *Communacations of the ACM*, Vol.40, No.3 (1997), 66~72.

[7] Basu, C., H. Hirsh and W. Cohen, "Recommendation as classification : using social and content-based information in recommendation", In *Proceedings of the 1998 Workshop on Recommender Systems*, 11-15, AAAI Press, 1998.

[8] Billsus, D. and M. J. Pazzani "Learning collaborative information filters", Proceedings of the 15th International Conference on Machine Learning, (1998), 46~54.

[9] Cho, Y. H., J. K. Kim and S. H. Kim, "A Personalized Recommender System based on Web Usage Mining and Decision Tree Induction", *Expert Systems With Applications*, Vol.23, No.3(2002), 329~342.

[10] Cho, Y. H. and J. K. Kim, "Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce", *Expert Systems With Applications*, Vol.26, No.3(2004), 234~246.

[11] Chow, G. C., "Tests of equality between sets of coefficients in two regressions", *Econometrica*, Vol.28, No.3(1960), 591~605.

[12] Hill, W., L. Stead, M. Rosenstein and G. Furnas, "Recommending nad Evaluating Choices in a Communication of Use", In proceedings of CHI 95, 1995.

[13] Kohonen, T., "The Self-Organizing Map", *Proceedings of the IEEE*, Vol.78, No. 9(1990), 1464~1480.

[14] Lawrence, R. D., G. S. Almasi, V. Kotlyar, M. S. Viveros and S. S. Duri, "Personalization of Supermarket Product Recommendation", *Data mining and Knowledge Discovery*, Vol.5, No.1-2(2001), 11~32.

[15] Lin, W., S. A. Alvarez and C. Ruiz, "Efficient Adaptive-Support Association Rule Mining for Recommender Systems", *Data Mining and Knowledge Discovery*, Vol.6, No.1(2002), 83~105.

[16] Mobasher, B., R. Cooly and J. Srivastava, "Automtic Personalization based on web mining", *Communications of ACM*, Vol.43, No.8(2000), 142~151.

[17] Nour, M. A. and G. R. Madey, "Heuristic and optimization approaches to extending the Kohonen self organizing algorithm", *European Journal of Operational Research*, Vol.93, No.2(1996), 428~448.

[18] Resnick, P., N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl, "Grouplens : An open architecture for collaborative filtering of netnews", In Proceedings of the ACM 1994 Conference on Computer Supported Cooperative Work, (1994), 175~186.

[19] Rijsbergen, C. J., *Information retrieval*, 2nd ed., London : Butterworths, 1979.

[20] Sarwar, B., G. Karypis, J. Konstan and J. Riedl, "Item-based collaborative filtering recommendation algorithm", In Proceedings of The Tenth International World Wide Web Conference, (2001), 285~295.

[21] Sarwar, B., G. Karypis, J. Konstan and J. Riedl, "Analysis of Recommendation Algorithms for E-commerce", In Proceedings of ACM E-commerce 2000 conference, (2000), 15~167.

[22] Sarwar, B., G. Karypis, J. Konstan and J. Riedl, "Application of Dimensionality Reduction in Recommender System - A case study", In proceedings of the ACm web KDD-2000 Workshop, 2000.

[23] Schmittlein, D. C. and R. A. Peterson, "Customer base analysis : An industrial purchase process application", *Marketing Science*, Vol.13, No.1(1994), 41~67.

[24] Shardanand, U. and P. Maes, "Social Information Filtering : Algorithms for Automating 'Word of Mouth'", In proceedings of CHI 95, 1995.

[25] Wang, D., S. Lee, C. Lin, "Goal-oriented sequential pattern for network banking churn analysis", Expert Systems with applications forthcoming, 2003.

Abstract

Considering Customer Buying Sequences to Enhance the Quality of Collaborative Filtering

Yeong-Bin Cho* · Yoon-Ho Cho**

The preferences of customers change over time. However, existing collaborative filtering (CF) systems are static, since they only incorporate information regarding whether a customer buys a product during a certain period and do not make use of the purchase sequences of customers. Therefore, the quality of the recommendations of the typical CF could be improved through the use of information on such sequences.

In this study, we propose a new methodology for enhancing the quality of CF recommendation that uses customer purchase sequences. The proposed methodology is applied to a large department store in Korea and compared to existing CF techniques. Various experiments using real-world data demonstrate that the proposed methodology provides higher quality recommendations than do typical CF techniques with better performance.

Key words : Recommender systems, Purchase sequence, Collaborative Filtering, SOM, Association Rules

* Department of Business Administration, Konkuk University
** School of Business Administration, Kookmin University