한국시뮬레이션학회 논문지
Vol. 15, No. 3, pp. 23-30 (2006. 9)

LSI 기법을 이용한 전자상거래 추천자 시스템의 시뮬레이션 분석

권치명[1†]

Simulation Study on E-commerce Recommender System by Use of LSI Method

Chi-myung Kwon

ABSTRACT

A recommender system for E-commerce site receives information from customers about which products they are interested in, and recommends products that are likely to fit their needs. In this paper, we investigate several methods for large-scale product purchase data for the purpose of producing useful recommendations to customers. We apply the traditional data mining techniques of cluster analysis and collaborative filtering(CF), and CF with reduction of product-dimensionality by use of latent semantic indexing(LSI). If reduced product-dimensionality obtained from LSI shows a similar latent trend of customers for buying products to that based on original customer-product purchase data, we expect less computational effort for obtaining the nearest-neighbor for target customer may improve the efficiency of recommendation performance. From simulation experiments on synthetic customer-product purchase data, CF-based method with reduction of product-dimensionality presents a better performance than the traditional CF methods with respect to the recall, precision and F1 measure. In general, the recommendation quality increases as the size of the neighborhood increases. However, our simulation results shows that, after a certain point, the improvement gain diminish. Also we find, as a number of products of recommendation increases, the precision becomes worse, but the improvement gain of recall is relatively small after a certain point. We consider these informations may be useful in applying recommender system.

Key Words : E-commerce, Recommender System, Collaborative Filtering, SVD, LSI

요 약

추천자 시스템은 전자상거래 사이트에서 고객의 상품 구매 정보를 수집하여 고객에 대한 예상 구매 상품을 추천하는 목적으로 개발되었다. 본 연구는 대형 전자상거래 사이트에서 고객의 상품 구매 이력이 활용 가능한 경우에 전통적인 통계기법인 군집분석 및 고객 간의 상품 구매 상관성을 이용하는 기존 추천자 시스템(협력적 필터링 기법)과 문서 검색에서 사용되는 LSI분석에 기반한 협업 필터링 기법을 상품 추천에 적용하여 각 기법의 상품 추천 효율성을 비교 분석하였다. 문서-용어 행렬과 유사한 구조를 가지는 고객-상품 구매 행렬에 문서 검색에 사용되는 LSI 분석법은 고객의 상품 구매 경향을 원 상품 수보다 축소된 차원의 변환 상품을 통하여 파악함으로써 목표고객에 대한 인접고객군의 생성 노력을 현저히 감소시킬 수 있어 결과적으로 실시간으로 적용되는 추천자 알고리즘의 효율성을 개선할 수 있을 것으로 기대할 수 있다. 가상적인 고객-상품 구매 리스트를 대상으로 실행한 시뮬레이션 실험 결과에서도 알고리즘의 효율성 평가측도인 recall과 정확도 및 F1에서 LSI 기반 협력적 필터링 기법이 기존의 방법보다 우수한 결과를 나타내었다. 시뮬레이션 결과, 인접고객 군의 크기가 일정한 수준에 이르면 그 크기를 증가시키더라도 알고리즘의 효율성은 별로 개선되지 않으며 또한 추천 상품 수가 일정 수준에 도달하면 추천 정확도가 낮아지는 정도에 비해 recall의 개선도는 별 변화가 없는 것으로 나타나고 있다. 추천자 시스템을 구현하는 용도에 따라 이러한 정보는 유용하게 사용될 수 있다고 판단된다.

주요어 : 전자상거래, 추천자 시스템, 협력적 필터링, 비정칙분해, LSI

* 이 논문은 2004년 한국학술진흥재단의 학술연구비(선도연구자과제)에 의하여 지원되었음.
2005년 12월 7일 접수, 2006년 6월 9일 채택
[1] 동아대학교 경영정보과학부
주 저 자 : 권치명
교신저자 : 권치명
E-mail; cmkwon@dau.ac.kr

1. 서 론

E-commerce 사이트에서 추천자 시스템(recommender system)은 고객이 관심을 가지는 상품에 대한 정보를 수집하여 고객이 구매할 것으로 예상되는 상품을 추천하기 위한 목적으로 개발되었다. 연관성(association) 분석기법과 협력적 필터링(collaborative filtering: CF)기법은 추천자 시스템 가운데 성공적으로 사용되고 있는 대표적인 기법이다.[11]

연관성 분석 기법은 구매 빈도(support)가 높은 상품군을 대상으로 고객의 습관이나 기호를 탐색하여 특정 상품 군을 구입하는 고객이 구입할 것으로 예상되는 상품 추천에 두 상품군 간의 연관성을 이용하는 기법이다. 이 기법은 두 상품군 집합간의 연관성 규칙을 발견하는 것으로 그 응용 범주는 자료의 특성 및 차원, 적용 수준, 제약조건 등에 따라 달라진다. CF기법은 고객이 구매한 상품에 대한 평가측도(rating)를 데이터베이스로 구축한 다음 목표고객(target customer)과 유사한 상품 구매 기호나 취향을 가지는 고객을 인접고객군(neighborhood)으로 분류하고 이들이 구매한 상품을 대상으로 목표고객의 선호도가 높은 상품을 추천하는 시스템으로 영화나 음반, 도서와 같은 분야의 상품 추천에서 널리 활용되고 있다.[3,4]

대형 전자상거래 사이트에서 실제로 고객들이 상품을 구매한 거래내역서(고객-상품 구매행렬)를 살펴보면 상품 구매를 활발히 하는 고객의 경우에도 판매되고 있는 제품의 1%에도 미치지 못하는 상품을 구매하고 있다. 개별 상품의 구매 빈도가 낮을 경우, 상품군을 중심으로 추천규칙을 발견하는 연관성분석 기법은 강력한 연관규칙을 발견하도라도 해당 상품군의 support가 낮아 상품 추천 규칙은 별 의미가 없다(전체 상품 중에서 평균적으로 1%를 구매하는 가상적인 시뮬레이션 데이터에 대하여 SAS의 E-miner version 9.1을 적용한 결과, 최대 support는 0.4% 이었음). 영화 등과 같이 특정한 분야에서는 상품에 대한 고객의 평가 자료로부터 목표고객의 상품 선호도를 추정하는 데 큰 어려움이 없지만 다양한 종류의 상품을 취급하는 전자상거래 사이트에서는 많은 상품에 대한 고객의 평가 자료가 불충분한 경우가 많다. 이러한 경우 CF 알고리즘은 평가자료 대신 상품 구매내역서를 바탕으로 인접고객군을 구하여 이로부터 목표고객의 구매 예상 상품을 추천한다.[7]

추천자 시스템을 성공적으로 활용하기 위해서 해결해야 할 과제로 취급 상품이 대량인 경우에 알고리즘의 효율성 문제라고 볼 수 있다.[12] 추천자 시스템은 알고리즘이 실시간으로 적용되어야 하는 만큼 알고리즘의 효율성은 시스템 반응시간과 직결되는 매우 중요한 의미를 갖는다. 따라서 고객과 상품의 수가 많을 경우, 많은 시간이 소요되는 인접고객 발견 과정을 개선하여 알고리즘의 효율성을 재고할 필요가 있다.

고객-상품 구매행렬의 특징을 살펴보면 행렬의 차원이 매우 크고, 전체 상품에 비해 특정 고객이 구매한 상품의 수가 적어 행렬의 대부분 원소는 0의 값을 가지며, 또한 일부 상품은 이름은 달라도 기능 면에서는 아주 유사할 수 있다. 문서 검색에서 사용되는 용어-문서 (term- document) 행렬도 대부분의 원소가 0인 점과 비슷한 취향을 가진 고객을 유사한 주제의 문서로 생각할 수 있으며, 그리고 용어-문서 행렬에서 다른 용어이지만 개념적으로 유사한 주제를 기술하는데 사용되는 동의어는 상품의 이름은 다르지만 기능 면에서는 유사한 상품일 수 있다는 개념으로 대응시키면 고객-상품 구매행렬의 형태는 용어-문서 행렬과 유사한 점이 많다고 볼 수 있다.

본 연구에서는 대형 전자상거래 사이트에서 고객의 상품 구매 이력이 활용 가능한 경우에 문서 검색에서 사용되는 LSI(latent semantic indexing) 분석법을 이용하여 추천자 시스템을 개선하는 방안을 연구하고자 한다. 문서에 포함된 용어들의 어의 특성을 축소된 차원으로 표현하는 LSI 분석법이 고객의 상품 구매 경향을 효과적으로 파악하는데 활용할 수 있다면 이는 결국 목표고객에 대한 인접고객 군 생성의 시간과 질적인 면을 개선할 수 있을 것으로 기대할 수 있을 것 이다. 또한 목표고객과 인접고객이 공동 구매한 상품의 수가 적을 경우에도 인접고객의 경향분석을 통한 상품 추천이 가능 할 것으로 기대한다.

추천자 알고리즘의 반응 시간과 질적인 면(추천된 상품에 고객이 원하는 제품이 포함되는 비율)은 서로 상충되는 된다고 볼 수 있는데 이러한 문제를 동시에 해결할 수 있는 방안은 실용적인 측면에서도 매우 유용할 것으로 생각한다. 이를 위해 본 연구에서는 전자상거래 사이트로부터 얻는 실제적인 자료와 유사한 자료를 시뮬레이션을 통하여 재생하고 이를 바탕으로 LSI에 의한 추천자 시스템의 효율성을 분석하여 추천자 시스템을 개선하는 방안을 연구하고자 한다.

2. 관련 연구

2.1 협력적 필터링 기법

인접고객군에 기반을 두는 CF 기법은 크게 3단계를 거쳐 목표고객에게 상품을 추천한다. 개별 상품 i에 대한 인

접고객 u의 평가지수 $r_{u,i}$를 바탕으로 먼저 목표고객 a와 상품 구매를 유사하게 하는 인접고객군을 발견하고, 인접고객들의 개별 상품에 대한 가중 평가를 사용하여 목표고객의 상품 i에 대한 구매 선호지수 $p_{a,i}$를 구한 다음, 선호지수가 높은 상품들을 차례로 목표고객 a에게 추천한다.[8] m 개의 상품에 대한 인접고객 u와 목표고객 a의 평가지수의 평균을 각각 $\bar{r_u}$와 $\bar{r_a}$라 할 때 두 고객의 상품 평가지수 사이의 피어슨 상관계수는 다음과 같다.

$$w_{a,u} = \sum_{i=1}^{m}(r_{a,i} - \bar{r_a})(r_{u,i} - \bar{r_u})/\sigma_a\sigma_u \quad (1)$$

GroupLens[6,10]는 목표고객 a가 상품 구매에 있어서 크기가 n인 인접고객군과 얼마나 유사한 성향을 나타내는 지수로 상관계수의 가중평균을 이용하여 목표고객 a의 상품 i에 대한 선호지수

$$p_{a,i} = \bar{r_a} + \sum_{u=1}^{n}(r_{u,i} - \bar{r_u})w_{a,u} / \sum_{u=1}^{n} w_{a,u} \quad (2)$$

를 예측하고 선호지수가 높은 상품을 추천하는 시스템을 처음으로 제안하였다.

두 고객의 유사성을 평가하는 여러 형태의 지수가 활용되고 있으며 이러한 추천자 시스템으로 Ringo music recommender[13]와 Bellcore Video Recomender[8]를 대표적으로 뽑을 수 있다. (인접고객군 CF 기법에 대한 자세한 내용은 참고문헌[7]을 참조).

2.2 LSI

LSI는 검색 요구 단어들을 문서의 제목과 대응시켜 검색하는 대신 검색 단어들이 가지는 개념을 통계적으로 추정되는 어의적 지표(semantic indexing)로 변환하여 관련 문서를 검색하는 방법으로 제안되었다[2]. 이 방법은 유사한 문서에 사용되는 단어들의 어의는 연관성이 있는 구조를 가진다고 가정하고 문서에서 사용된 단어 사이에 연관구조를 단어-문서 행렬 A의 비정칙분해(singular value decomposition: SVD)를 통하여 추정한다. 사용 단어의 수가 m이고 전체 문서의 수가 n인 크기 $(m \times n)$인 단어-문서 행렬 $A\,(a_{ij})$에서 a_{ij}는 단어 i가 문서 j에 포함되는 빈도이다. 만일 $rank\,(A)$=r이면 행렬 A의 SVD는 다음과 같다.

$$A = U\Sigma V^T \quad (3)$$

여기서 $U^TU = V^TV = I_n$ 이며 $\Sigma = diag(\sigma_1, , \ldots, \sigma_n), \sigma_i > 0 (1 \leq i \leq r), \sigma_i = 0 (i \geq r + 1)$이다. 식 (3)의 σ_i는 A의 비정칙치이며 σ_i^2 $(1 \leq i \leq n)$는 AA^T의 고유치이다. A에 대한 SVD는 A에 내재된 원래의 단어-문서의 연관관계를 선형 독립적인 r개의 벡터로 분해하는 것으로 만일 식 (3)에서 $\sigma_1 \geq \sigma_2 \ldots \geq \sigma_r$이면 A에 내재된 연관구조와 가장 근사하는 용어-문서 행렬 A_k $(k \leq r)$의 수리적인 형태는 다음 그림과 같다[5].

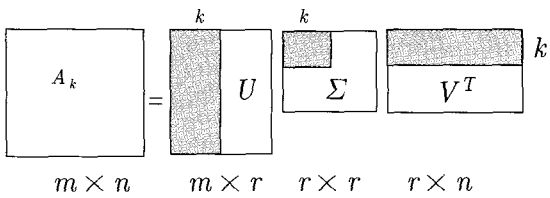

그림 1. 행렬 A_k의 수리적 형태

행렬 A_k는 대각행렬 Σ에서 비정칙치의 값이 큰 k개의 $\sigma_i(1 \leq i \leq k)$로 구성된 대각행렬 Σ_k과 k개의 $\sigma_i(1 \leq i \leq k)$에 대한 k개의 좌-우 비정칙 벡터로 구성된 행렬 U_k와 V_k의 곱으로 표현할 수 있으며 이는 행렬 A에 내재된 용어-문서의 연관성 구조에 대한 중요 정보를 k개의 인자를 통하여 추출하고 문서의 검색에서 요구 단어의 다양한 표현에서 나타날 수 있는 변이성을 제거시키는 역할을 한다. 직관적으로 k는 문서행렬에서 사용된 단어수 m에 비해 상당히 적은 값이며 유사한 문서에서 사용된 용어는 같지 않더라도 k-차원의 인자공간에서는 서로 인접해 있을 수 있다. LSI는 m-차원의 검색 요구 문서 q를 k-차원 공간의 문서 벡터

$$\hat{q} = q^TU_k\Sigma_k^{-1} \quad (4)$$

로 변환하고, 이를 A_k에서의 문서 벡터와 비교하여 검색 문서 벡터 q와 유사한 문서를 검색결과로 제시한다. LSI

는 유사성 측도로 문서 벡터와 검색 요구 벡터가 이루는 각도의 cosine을 일반적으로 사용하며 이 값이 일정한 범위 내에 있는 인접한 문서를 검색결과로 출력한다.

3. LSI 기반 상품 추천자 시스템

Web 상에서 실시간으로 상품을 판매하는 전자상거래 시스템에서 시스템의 반응시간은 알고리즘의 질적인 문제와 함께 매우 중요한 의미를 가진다. 문서검색에 사용되는 LSI 방법은 문서행렬에 내재된 용어-문서의 연관성 구조를 발견하고 이를 이용하여 원 문서의 어의적 특성을 지니는 낮은 차원의 문서행렬로 원 문서행렬을 변환한다. 차원이 축소된 변환 문서행렬을 이용함으로써 검색문서와 인접한 문서를 발견하는 알고리즘의 계산 노력을 감소시킬 수 있으며 아울러 변환 문서행렬에 내재된 어의적 특성은 검색문서와 유사한 문서를 인접문서로 분류하는데 질적인 측면에서 기여할 수 있다.

문서검색에 사용되는 LSI 방법을 추천자 시스템에 적용하여 LSI에 의한 차원의 축소가 원래 고객의 상품 구매 정보에 대한 특성을 유지하면서 목표고객에 인접한 고객군을 효과적으로 발견할 수 있다면 추천자 알고리즘의 효율성을 개선하게 될 것으로 기대할 수 있다.

본 절에서는 추천자 시스템의 상품 추천과정을 3 단계로 나누어 살펴보고자 한다. 먼저 고객-상품 구매행렬에 대한 SVD를 통하여 구매행렬의 차원을 축소하는 방법을 분석하고 다음으로 목표고객에 대한 인접고객군을 생성하는 문제를 다루고 마지막으로 인접고객군의 구매 상품군으로부터 N개의 구매 선호 상품 리스트를 목표고객에게 제공하는 과정을 제시하고자 한다.

3.1 고객-상품 구매행렬의 변환과 특성

n명의 고객이 m개의 상품에 대한 구매 내역은 크기가 (nxm)인 고객-상품 구매행렬 $A(a_{ij})$로 표시할 수 있다. i번째 고객이 j번째 상품을 구매하는 경우, 구매행렬의 a_{ij} 원소는 1이며 그렇지 않는 경우에는 0이다. 고객-상품 구매행렬 A에 대한 SVD는 식 (3)과 같이 표현할 수 있으며 만일 여기에서 대각행렬 $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_n)$의 대각선 원소의 값이 $\sigma_1 \geq \sigma_2 \ldots \geq \sigma_k$인 관계를 나타내면 A에 내재된 고객의 상품 구매 특성과 가장 근사하며 차원이 $k(\leq r)$인 변환-구매행렬 A_k는 다음과 같이 표현할 수 있다.

$$A_k = U_k \Sigma_k V_k^T \tag{5}$$

(그림 1에서 등호 우측의 색깔이 있는 부분을 차례로 U_k, Σ_k, V_k^T로 표시).

이와 같이 절단된(truncated) SVD의 A_k는 n명의 고객의 m개의 상품에 대한 구매 경향 대신 k개의 변환상품(k-meta products)에 대한 구매 특성을 나타내는 변환상품 구매행렬로 생각할 수 있다. 즉 LSI는 고객의 상품 구매 특성을 k개의 독립적인 변환상품 구매인자로 기술한다고 볼 수 있다. 변환상품의 수 k는 원 구매 상품의 수 m보다 축소됨으로 n명의 고객을 대상으로 인접고객을 발견하는 계산 노력을 감소 시킬 수 있다. 또한 유사한 상품 사이에 있을 수 있는 내재적인 구매 특성을 고려함으로써 상품명이 달라도 기능이 유사한 상품을 같은 인자를 가지는 상품으로 취급함으로써 결과적으로 구매한 상품 사이의 연관성을 발견하는데 효율적 일 수 있다.

3.2 인접고객군 생성

목표고객의 인접고객 군을 발견하는 과정은 추천자 시스템에서 가장 중요한 부분으로서 목표고객의 상품 구매 성향과 유사한 성향을 가지는 고객들이 인접고객 군으로 분류된다. 상품-구매 변환행렬 A_k를 이용하면 개별 고객의 상품 구매 성향은 k-차원의 벡터로 표현할 수 있으며 m개의 상품에 대한 목표고객의 구매 이력(크기 m인 벡터 q)도 식 (4)에 의하여 k-차원의 상품-구매 벡터로 변환될 수 있다.

두 고객 사이의 유사성(similarity)은 두 고객의 상품 구매성향이 얼마나 인접하고 있음을 측정하는 측도로 추천자 시스템에서는 보통 k-차원의 공간상에서 두 고객 벡터(상품구매 변환 벡터) a와 b가 이루는 $cosine$ 각도를 기준으로 인접성을 평가한다[10].

$$cos(a, b) = a \cdot b / (\mid a \mid \mid b \mid) \tag{6}$$

전체 고객에 대해 인접고객 군을 찾는 두 가지 방법 중에서 자주 사용으로는 우선 목표고객 중심법을 들 수 있다[7]. 이 방법은 목표고객을 중심으로 인접성이 높은 고객을 차례로 p명 선택하여 크기가 p인 p-인접고객군을 생성한다. 다음으로 인접고객 중심법은 목표고객과 그와 가장 인접한 고객을 크기가 2인 첫 번째 인접 고객군의 구

성원소로 선택하고 이들의 중심을 인접고객군의 중심으로 정한다. 전 단계에서 구한 인접고객군의 중심과 가장 인접한 고객을 새로운 인접고객군의 요소로 편입하여 다시 인접고객군의 중심을 계산하는 방법으로 크기가 p인 인접고객군을 생성할 때까지 반복과정을 통하여 p-인접고객군을 구하는 방법이다. 이 방법은 인접고객군을 생성하는데 가장 인접한 고객이 영향을 미치게 된다.

본 연구에서는 목표고객과 다른 고객들 사이의 유사성을 계산하여 유사성이 높은 고객들을 차례로 목표고객의 인접고객 군으로 분류하는 목표고객 중심법을 이용하여 인접고객 군을 생성하고자 한다. 만일 고객의 구매 이력이 데이터 파일에 없는 새로운 고객이라면 상품 사이트에 대한 고객의 browsing으로부터 얻는 정보를 구매이력으로 대신하는 것이 대안으로 제시되고 있다[14].

3.3 상품추천서 작성

인접고객 군으로부터 상위 N개의 상품을 목표고객에게 추천하는 방법은 최빈 상품 추천(most frequent items recommendation)방법을 사용하고자 한다. 최빈 상품 추천은 인접고객군의 구매 행렬을 조사하여 구매 상품의 도수분포를 구한 다음 이를 이용하여 목표 고객이 구매한 경험이 없으며 구매 빈도가 높은 상위 N개의 상품을 추천하는 방안이다.

4. 시뮬레이션 실험

본 연구에서는 E-commerce 사이트로부터 얻는 실제적인 고객 자료와 유사한 자료를 시뮬레이션을 통하여 재생하고 이를 바탕으로 LSI에 의한 추천자 시스템의 효율성을 분석하였다. 아울러 인접고객군의 크기와 추천 상품의 수에 따라 알고리즘의 정확도, recall 및 F1 평가측도는 어떻게 변화하는지를 조사하여 알고리즘의 적용성에 대한 방안을 분석하고자 한다.

4.1 고객-상품 구매 리스트 작성

시뮬레이션 실험의 편의상 전체 상품의 수를 500개, 고객의 수를 1000개로 지정하고 같은 크기의 training set과 test set을 작성하였다. 개별 고객의 구매 상품수는 평균적으로 전체 상품의 1% 정도를 구매하는 것으로 가정하고 구매 상품 수는 평균이 5인 포아슨 분포를 이용하여 재생하였다. 구매 상품수가 0인 경우는 단 1개의 상품만을 구매하는 것으로 처리하였다. 전체 자료에서 minimum support를 만족하는 large item set(LIS)의 크기는 평균이 2인 포아슨 분포를 따르는 것으로 가정하였다. 전체 고객에 대한 LIS를 작성하기위해서 우선 첫 번째 LIS는 임의로 선택하고 그 다음 LIS는 직전 LIS로부터 상관비율(correlation level)만큼 구매상품을 선택하고 나머지 상품은 임의 선택하여 전체적으로 구매상품의 수가 평균이 2인 포아슨분포를 따르도록 하였다. 상관비율은 평균이 0.5인 지수분포로부터 확률적으로 재생하였다.

이와 같이 얻은 전체 LSI에 가중치 w_i를 부여하여 i번째 LSI가 개별 고객의 고객-구매 상품이력에 포함될 확률을 정하였다. 평균이 1인 지수분포를 이용하여 얻은 i번째 확률변수 E_i로부터 가중치는 $w_i = E_i / \sum E_i$ 이다. LIS에 포함된 모든 물품을 항상 구매하지 않는다는 상황을 유사하게 모델화 하기위해서는 개별 고객의 LSI에 확률적인 corruption level을 부여하여 일부 품목을 탈락시켰다. Corruption level c는 N(0.5, 0.1)을 따르는 확률변수로 만일 c가 일량분포 U(0,1)를 따르는 확률변수의 값보다 크면 LIS의 모든 품목을 구입하고 아니면 LSI의 첫 번째 품목을 구매 리스트에서 삭제하고 다시 corruption level를 부여하는 반복과정을 통하여 LIS의 상품리스트를 결정하였다. 최종적으로 결정된 LIS에 확률적으로 추가 상품을 구매하여 전체적으로 구매 상품 수가 Poisson (5)을 따르도록 각 고객의 상품 구매 행렬을 재생하였다. (자세한 고객-상품구매 리스트의 생성과정은 참고문헌[1] 참조).

4.2 평가측도

고객의 상품 구매 자료를 training set과 test set, 두 부분으로 나누고 training set을 대상으로 추천자 시스템을 적용하며 목표고객에게 선호 예상 상위 N개의 상품을 추천하였다. 추천된 상품을 실제로 목표고객이 구매한 비율과 목표고객이 실제로 구매한 상품 중에서 추천 상품의 비율은 추천자 시스템의 효율성을 평가하는 측도가 될 수 있다. 추천된 상품 중에서 목표고객이 실제로 구매한 상품의 집합을 적중집합(hit set)이라고 하면 recall과 precision은 다음과 같이 정의된다[9].

$$recall = \frac{size\ of\ hit\ set}{size\ of\ test\ set} \tag{7}$$

$$precision = \frac{size\ of\ hit\ set}{N} \tag{8}$$

위 두 식에서 추천 상품의 수 N이 커지면 recall은 증가하나 precision은 감소하므로 이 두 측도는 서로 상충된다고 볼 수 있다. 이러한 점을 고려하여 정보검색의 효율성을 평가하는 측도로 개별 목표고객에 대한 F1 측도를 계산하고 이들의 전체 평균을 추천자 시스템 알고리즘의 평가 측도로 사용하였다[14].

$$F1 = \frac{2^* recall^* precision}{recall + precision} \quad (9)$$

4.3 실험 결과

LSI에 의한 상품 추천자 시스템의 효율성을 비교하기 위하여 고객-상품 구매행렬을 대상으로 CF 기법과 군집분석(cluster analysis: CA)기법 그리고 LSI 기반 추천자 시스템을 각각 적용하여 상품 추천 효율성 평가측도를 계산하였다. 군집분석에서는 유사성 측도로 두 벡터 사이의 각도(cosine)를 사용하여 인접고객 군을 발견하고 이들이 구매한 최빈 N개의 상품을 목표고객에게 추천하는 방법을 사용하였다. 인접고객군의 크기를 10에서 50까지 10단위씩 증가시키고 추천 상품 수는 5에서 25까지 5만큼씩 증가시키면서 세 가지 방법에 의한 시뮬레이션 결과는 <표 1>과 같다.

<표 1>에서 LSI 기반 추천자 시스템이 recall, precision 및 F1 평가측도에서 CF 기법과 CA 기법보다 상품 추천 효율성이 우수하게 나타나고 있으며 CF 기법은 CA 기법과 추천 효율성이 비슷한 경향을 보이나 추천 효율성은 다소 좋은 것으로 나타나고 있다.

인접고객군의 크기가 증가함에 따라 LSI의 경우 recall 값은 증가하나 대략 인접고객군의 크기가 일정한 값(30) 이상이면 거의 변화가 없는 것으로 나타나고 있다. 반면 CA와 CF 추천자 알고리즘의 경우, recall 값은 인접고객군의 크기가 20 이상이 되면 다소 감소하는 경향을 보이고 있다(그림 2 참조). LSI, CA와 CF 추천 시스템에 의한 precision과 F1 평가 측도는 recall과 비슷한 경향을 보이고 있으며 인접고객군의 크기가 30 이상이 되면 큰 변화가 없는 것으로 나타나고 있다(그림 3과 그림 4 참조).

표 1. 인접고객군의 크기(c)와 추천 상품수(N)에 따른 평가측도

c	기법	추천 상품 수(N)				
		5	10	15	20	25
10	LSI	3.178	3.447	3.679	3.896	4.092
		0.636	0.345	0.245	0.195	0.164
		1.056	0.627	0.460	0.371	0.315
	CA	1.691	2.472	2.888	3.095	3.184
		0.338	0.247	0.193	0.155	0.127
		0.564	0.449	0.361	0.295	0.245
	CF	1.777	2.540	2.919	3.064	3.117
		0.355	0.254	0.195	0.153	0.125
		0.592	0.462	0.365	0.292	0.240
20	LSI	3.929	4.621	4.707	4.736	4.758
		0.786	0.462	0.314	0.237	0.190
		1.310	0.840	0.588	0.451	0.366
	CA	1.523	2.371	2.932	3.284	3.454
		0.305	0.237	0.195	0.164	0.138
		0.508	0.431	0.366	0.313	0.266
	CF	1.658	2.489	3.010	3.324	3.447
		0.332	0.249	0.201	0.166	0.138
		0.553	0.453	0.376	0.317	0.265
30	LSI	4.074	4.836	4.960	4.989	4.989
		0.815	0.484	0.331	0.249	0.200
		1.358	0.879	0.620	0.475	0.384
	CA	1.444	2.231	2.808	3.238	3.519
		0.289	0.223	0.187	0.162	0.141
		0.481	0.406	0.351	0.308	0.384
	CF	1.697	2.382	2.912	3.316	3.572
		0.339	0.238	0.194	0.166	0.143
		0.566	0.433	0.364	0.316	0.275
40	LSI	4.081	4.909	4.975	5.019	5.029
		0.816	0.491	0.372	0.251	0.201
		1.360	0.893	0.622	0.478	0.387
	CA	1.413	2.156	2.664	3.119	3.434
		0.283	0.216	0.178	0.156	0.137
		0.471	0.392	0.333	0.297	0.264
	CF	1.793	2.343	2.819	3.242	3.551
		0.359	0.234	0.188	0.162	0.142
		0.598	0.426	0.352	0.309	0.273
50	LSI	4.071	4.918	4.970	5.001	5.024
		0.814	0.492	0.331	0.250	0.201
		1.357	0.892	0.621	0.476	0.386
	CA	1.257	2.146	2.583	2.977	3.311
		0.251	0.215	0.172	0.149	0.132
		0.419	0.390	0.323	0.234	0.255
	CF	1.888	2.305	2.795	3.174	3.478
		0.378	0.238	0.186	0.159	0.139
		0.629	0.434	0.349	0.302	0.268

(각 cell의 값은 위에서 차례로 recall, precision, F1임)

LSI 기반 추천자 시스템에 의한 상품 추천에서 인접고객군의 크기와 추천 상품 수에 따른 recall과 precision 및 F1 평가측도의 변화는 <그림 5>-<그림 7>와 같이 나타나고 있다. 추천 상품 수가 증가함에 따라 recall이 증가하고 precision과 F1은 감소하고 있음을 알 수 있으며 추

천 상품의 수가 일정할 경우에 인접고객군의 크기(c)가 30이상이면 3가지 평가측도는 별 변화가 없는 것으로 보인다. 추천 상품의 수가 15정도 이상이면 recall의 증가 정도가 별로 크지 않으며 전체적으로 시스템의 정확도와 F1 측도는 추천 상품 수가 5일 때 최대로 나타나고 있다.

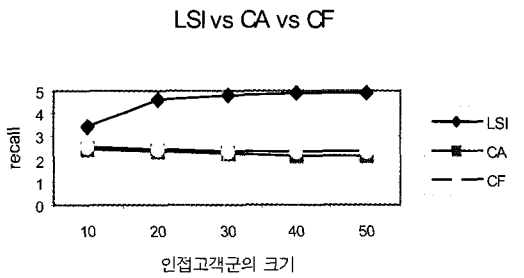

그림 2. N=10에서 인접고객군의 크기에 따른 recall

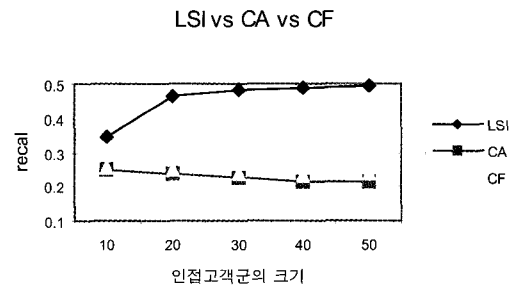

그림 3. N=10에서 인접고객군의 크기에 따른 precision

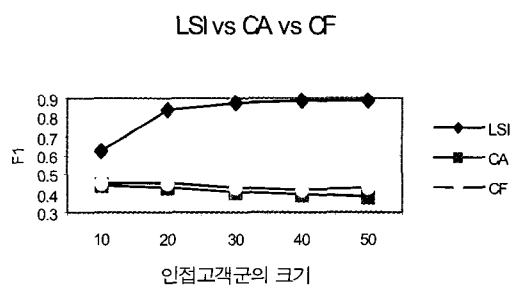

그림 4. N=10에서 인접고객군의 크기에 따른 F1

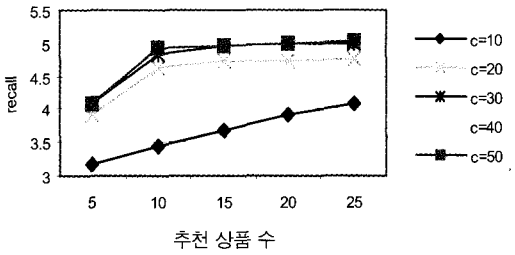

그림 5. 추천 상품 수에 따른 recall

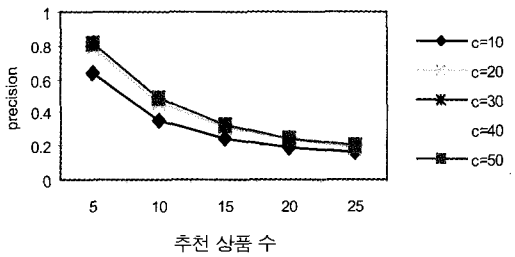

그림 6. 추천 상품 수에 따른 precision

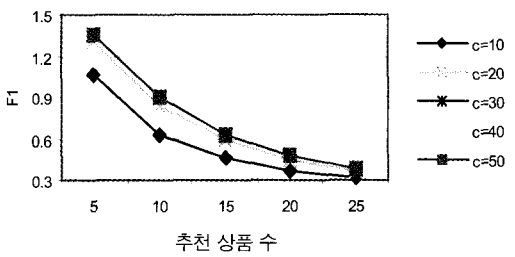

그림 6. 추천 상품 수에 따른 F1

5. 결 론

본 연구는 시뮬레이션을 통하여 얻은 고객의 상품 구매에 대한 가상적인 자료를 바탕으로 문서 검색에서 사용되는 LSI 기법을 추천자 시스템에 활용하였다. LSI 기법을 적용하여 얻어진 고객-상품 변환 구매행렬은 원 구매행렬과 비교하여 몇 가지 장점을 가지게 될 것으로 기대할 수 있다. 우선 원 상품 구매행렬에 비해 차원이 낮은

변환 상품 구매행렬은 인접고객군을 발견하는 계산 노력을 감소시킬 수 있으며 아울러 원 상품 구매행렬이 갖는 sparsity 문제를 완화시킬 수 있을 것으로 판단된다. 또한 상품명은 다르나 기능이 유사한 상품을 구매하는 경우에도 두 상품 사이에 내재한 연관성을 고려할 수 있음으로 추천 시스템의 정확성을 개선시킬 수 있을 것으로 기대한다.

고객들의 가상적인 상품 구매 리스트를 대상으로 실행한 시뮬레이션 실험 결과에서도 목표고객의 인접고객군 발견에 원 고객-상품 구매행렬보다는 SVD에 의한 변환 상품 구매행렬을 이용하는 것이 추천자 시스템의 효율성을 평가하는 recall, precision 및 F1 측도에서 CA와 CF 기법보다 우수하게 나타나고 있어 이러한 주장을 지지하고 있다.

일반적으로 추천 상품 수가 증가함에 따라 추천자 시스템의 recall은 증가하고 정확도는 감소하는데 그 증감 정도는 목표고객에 대한 인접고객군의 크기나 추천 상품 수가 일정한 수준에 이르면 큰 변화가 없는 것으로 나타나고 있다. 추천자 시스템을 구현하는 용도에 따라 이러한 정보는 유용하게 사용될 수 있다고 판단된다.

참 고 문 헌

1. Agawal, R and R. Srikant (1994), "Fast Algorithms for Association Rules", Proceedings of the 20th VLDB Conference, pp. 487-499.
2. Berry, M., Dumais, S. and G. O'Brian (1995), "Using Linear Algebra for Intelligent Information Retrieval," SIAM Review, Vol. 37, pp. 573-595.
3. Billsus, D. and M. Pazzani (1998), "Learning Collaborative Information Filters," Proceedings of ICML, pp. 46-53.
4. Breese, J., Heckerman, D. and C. Kadie (1998), "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, pp. 43-52.
5. Deerwester, S., Dumais, S. and Furnas, G. and T. Landauer (1990), "Indexing by Latent Semantic Analysis," J. of the American Society for Information Science. Vol.41, No. 6, pp. 391-407.
6. Good, N., Schafer, B., Konstan, L., Borchers, A., Sarwar, B., Herlocker J. and J. Riedl (1999), "Combining Collaborative Filtering with Personal Agents for Better Recommendations," Proceedings of the AAAI Conference, pp. 439-446.
7. Herlocker, J.H., Joseph, A., Borchers, Al and J. Riedl (1999), "An Algorithm Framework for Performing Collaborative Filtering," Proceedings of the Conference on Research and Development in Information Retrieval.
8. Hill, H., Stead, L., Rosenstein, M and G. Furnas (1999), "Recommending and Evaluating Choices in a Virtual Community of Use," Proceedings of ACM, pp. 194-201.
9. Kautz, H., Selman, B. and M. Shah (1997), "Combining Social Networks and Collaborative Filtering," Communications of the ACM, Vol. 40, No. 3, pp. 63-65.
10. Konstan, J., Miller, B., Martz, D., Herlocker, J., Gordon, L. and J. Riedl (1997), "GroupLens: Applying Collaborative Filtering to Usenet News," Communication of the ACM, Vol.40, No.3, pp. 77-87.
11. Resnick, P. and H. Varian (1997), "Recommender Systems," Special Issue of Communications of the ACM, Vol. 40, No. 3.
12. Sarwar, B., Karypis, G., Konran, J. and J. Riedl (2000), "Analysis of Recommendation Algorithm for E-Commerce," Proceedings of the 2nd ACM Conference on Electronic Commerce.
13. Shadanand, U. and P. Maes (1995), "Social Information Filtering: Algorithm for Automating Word of Mouth," Proceedings of ACM, pp. 210-217.
14. Shafer, J., Konstan, J. and J. Riedl (1999), "Recommender Systems in E-Commerce," Proceedings of ACM E-Commerce Conference.

권 치 명 (cmkwon@dau.ac.kr)

1978 서울대학교 공과대학 산업공학과 학사
1983 서울대학교 대학원 산업공학과 석사
1991 VPI& SU 산업시스템공학과 박사
1984～현재 동아대학교 경영대학 경영정보과학부 교수

관심분야 : Simulation Modeling & Output Analysis, Simulation Optimization, FMS