전자상거래 추천자 시스템에 대한 분석

권치명*

Simulation Study on E-commerce Recommendation System

Chi-myung Kwon

Abstract

추천자 시스템은 E-commerce 사이트에서 소비자가 관심을 가지는 상품에 대한 정보를 수집하여 소비자가 구매할 것으로 예상되는 상품을 추천하는 목적으로 개발되었다. 추천자 시스템을 구축하여 성공적으로 활용하기 위해서 해결해야 할 과제로 취급 상품이 대량인 경우에 알고리즘의 효율성 문제라고 볼 수 있는데 본 연구는 문서 검색에서 사용되는 LSI(latent semantic indexing) 분석법을 이용하여 추천자 시스템을 개선하는 방안을 연구하고자 한다. LSI 분석법을 이용하여 고객-상품 구매행렬에서 고객이 상품을 구매하는 경향을 효과적으로 파악할 수 있다면 목표고객에 대한 인접고객군을 생성하는 계산 노력은 현저히 감소되어 추천자 알고리즘이 실시간으로 고객 데이터베이스로부터 많은 인접 고객을 효율적으로 검색할 수 있을 것으로 기대된다. 본 연구는 E-commerce 사이트로부터 얻는 실제적인 고객 자료와 유사한 자료를 시뮬레이션을 통하여 재생하고 이를 바탕으로 LSI에 의한 추천자 시스템의 효율성을 분석하고자 한다.

Key Words: 전자상거래, 추천자 시스템, SVD, LSI

* 동아대학교 경영정보과학부

1. 서론

E-commerce 사이트에서 고객이 원하는 상품을 쉽게 구매할 수 있도록 도움을 주기위해서 제안된 추천자 시스템(recommender sys-tem)은 소비자가 관심을 가지는 상품에 대한 정보를 수집하여 소비자의 욕구에 적합할 것으로 생각되는 상품을 추천하는 목적으로 개발되었다[4]. 공동 필터링 기법(collaborative filtering technology)은 추천자 시스템 가운데 성공적으로 사용되고 있는 대표적인 기법이다. 이 기법은 통계적인 이론을 적용하여 목표고객(target customer)과 유사한 구매 경험이나 의견을 가지는 인접고객군을 파악하고 인접고객이 자주 구매하는 상품을 추천하거나 또는 연관 규칙을 사용하여 추천하는 시스템이다[10]. 추천자 시스템을 구축하여 성공적으로 활용하기 위해서 해결해야 할 과제로 취급 상품이 대량인 경우에 알고리즘의 효율성 문제라고 볼 수 있다 [18].

추천자 시스템은 알고리즘이 실시간으로 적용되어야 하는 만큼 알고리즘의 효율성은 시스템 반응시간과 직결되는 매우 중요한 의미를 갖는다. 고객과 상품의 수가 많아지면 인접고객의 발견에 많은 시간이 소요되어 알고리즘의 효율성이 떨어질 수 있다. 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 있어 고객-상품 구매행렬의 대부분 원소는 0의 값을 가지는 행렬이다. 또한 판매되고 있는 제품의 이름은 달라도 기능 면에서는 실제로 아주 유사한 제품일 수 있다.

고객-상품 구매행렬은 행렬의 차원이 매우 크고 대부분의 원소가 0인 점에서 문서의 검색에서 사용되는 용어-문서 행렬(term docu-ment matrix)과 유사하다고 할 수 있다. 또한 용어-문서 행렬에서 다른 용어이지만 개념적으로 유사한 주제를 기술하는데 사용되는 동의어(synonymy)는 제품의 이름은 다르지만 기능 면에서는 유사한 제품일 수 있다는 개념으로, 비슷한 취향을 가진 고객을 유사한 주제의 문서로 각각 대응시켜보면 고객-상품 구매행렬의 형태는 용어-문서 행렬과 유사한 점이 많다고 볼 수 있다.

상품 추천 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 기대한다. 이를 위해 본 연구에서는 문서 검색에서 사용되는 LSI(latent semantic indexing) 분석법을 이용하여 추천자 시스템을 개선하는 방안을 연구하고자 한다. E-commerce 사이트로부터 얻는 실제적인 고객 자료와 유사한 자료를 시뮬레이션을 통하여 재생하고 이를 바탕으로 LSI에 의한 추천자 시스템의 효율성을 분석하고자 한다. 아울러 데이터 셀의 차원이나, 인접고객 군의 크기에 따라서 알고리즘의 정확도와 recall 기준에서 LSI의 효율성을 평가해보고 개선 방안을 제시하고자 한다.

2. LSI

LSI은 문서(논문 또는 책의 제목) 검색 요구(query)에서 요구 용어(단어)들을 문서와 대응시켜 검색하는 대신 요구 단어들이 가지는 개념을 통계적으로 추정되는

어의적인 지표(semantic indexing)로 변환하여 관련 문서를 검색한다[2]. 이 방법은 유사한 문서에 사용되는 단어들의 어의는 연관성이 있는 구조를 가진다고 가정하고 문서에서 사용된 단어들 사이에 연관 구조를 비정칙분해(singular value decompo-sition: SVD)를 통하여 추정하고자 한다. SVD는 크기가 $(m \times n)$이며 $rank(A)=r$인 행렬 A를 다음과 같이 분해한다.

$$A = U\Sigma V^T \qquad (1)$$

여기서 $\Sigma = diag(\sigma_1, ..., \sigma_n), \sigma_i > 0, U^TU = V^TV = I_n \ \ (1 \leq i \leq r), \ \sigma_i = 0 (i \geq r+1)$ 이다. 직교행렬 U와 V의 처음 r개의 열은 각각 AA^T와 A^TA의 r개의 비음 고유치에 대한 고유벡터며, U와 V의 열을 각각 좌 비정칙 벡터(left singular vector) 우 비정칙 벡터(right singlular vector)라 한다. 식 (1)의 σ_i는 A의 비정칙치(sinular value)이며 $\sigma_i^2 (1 \leq i \leq n)$는 AA^T의 고유치이다.

LSI의 문서 데이타베이스에서 단어-문서 행렬 $A(a_{ij})$는 단어를 행으로 문서를 열로 나열한 다음 문서 j에 단어 i가 포함되는 빈도 a_{ij}를 구하여 이를 행렬의 원소로 한다. 보통 개별 단어는 각 문서에서 몇 개만 나타나므로 용어-문서 행렬은 대부분이 0의 값을 가지는 행렬이다. 위 식 (1)에서 A에 대한 SVD는 A의 비정칙 벡터로 된 U와 V, 그리고 A의 비정칙치(singular value) 원소로 된 대각행렬 Σ, 이 3개의 행렬 곱으로 행렬 A를 나타내는 것으로 A에 내재된 원래의 용어-문서의 연관관계를 선형 독립적인 r개의 벡터(또는 인자)들로 분해하는 의미를 지닌다. 만일 식 (1)에서 $\sigma_1 \geq \sigma_2 ... \geq \sigma_r$이면 A에 내재된 연관구조와 가장 근사하는 용어-문서 행렬 중에서 차원이 $k(\leq r)$인 행렬 A_k는 대각행렬 Σ에서 비정칙치의 값이 큰 k개의 $\sigma_i (1 \leq i \leq k)$로 구성된 대각행렬 Σ_k과 k개의 σ_i $(1 \leq i \leq k)$에 대한 k개의 좌-우 비정칙 벡터로 된 U_k와 V_k의 곱으로 나타난다[2]. 즉 A_k의 수리적인 형태를 그림으로 나타내면 다음과 같다.

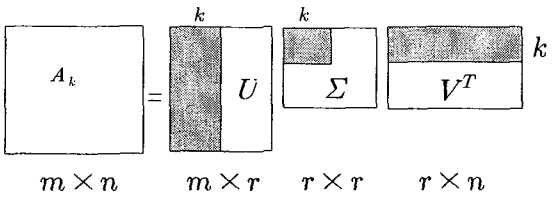

그림 1. 행렬 A_k의 수리적 형태

절단 SVD(truncated SVD)는 상호 관련성이 없는 인자를 k개 추출하여 좌-우 비정칙 벡터를 이용하여 용어-문서 행렬을 k-차원에서 벡터로 나타내는 것이라고 볼 수 있는데, 이는 행렬 A에 내재된 용어-문서의 연관성 구조에 대한 중요 정보를 k개의 인자를 통하여 추출하고 동시에 문서의 검색에서 요구 단어의 다양한 표현에서 나타날 수 있는 변이성(vrriability or noise)를 제거시키는 역할을 한다. 직관적으로 k는 문서에 사용된 단어의 수 m에 비해 아주 적은 값이며 용어 사용에서 사소한 차이는 무시된다고 볼 수 있다. 예를 들어 유사한 문서에서 사용된 용어는 같지 않더라도 k-차원의 인자 공간에서는 서로 인접해 있을 수 있다. SVD를 이용하여 용어 사이의 연관성(interrelationship)을 모델링하고 이

를 바탕으로 문서 검색 효율성을 제고하는 방법을 개발하는 것이 LSI의 주요 내용이다.

문서 검색에서 단어들로 구성된 사용자의 검색 요구 q는 하나의 문서로 생각할 수 있다. 검색 요구 q를 LSI는 k-차원 공간의 문서 벡터 $\hat{q}$로 변환하고, 즉

$$\hat{q} = q^T U_k \Sigma_k^{-1} \quad (2)$$

이를 변환된 데이타 파일 A_k에서 문서 벡터와 비교한다. 검색 요구 q와 유사성(similarity)을 가지는 문서를 검색결과로 사용자에게 제시한다. 가장 일반적인 유사성 측도로는 문서 벡터와 검색 요구 벡터 사이의 cosine을 사용하며 일정한 범위 내에 있는 인접한 문서를 검색결과로 출력하게 된다.

3. 추천자 시스템

추천자 시스템은 E-commerce 사이트에서 목표 고객에게 N 개의 구매 예상 선호 상품 리스트(list of top-N products)를 제공함으로써 고객이 구매하고자 하는 상품을 찾는데 도움을 주고자 한다. n명의 고객들이 m개의 상품에 대한 구매 내역은 크기가 $(n \times m)$인 고객-상품 구매행렬 $A(a_{ij})$로 표시할 수 있으며 만일 i번째 고객이 j번째 상품을 구매하는 경우, a_{ij}원소는 1이며 그렇지 않는 경우에는 0이다. 구매행렬의 원소는 대부분 0의 값을 가지는데, 만일 인접 고객이 구매한 상품의 수가 너무 적으면 인접 고객의 경향분석을 통한 상품 추천이 거의 불가능하게 되어 추천자 시스템의 적중률을 저하시켜 낮은 수준의 추천 정확도 결과를 가져올 수 있다. 또한 고객과 상품의 수가 많을 경우, 인접고객을 발견에 시간이 많이 소요되어 인접고객을 발견하는 알고리즘의 효율성이 저하된다. 특히 web 상에서 실시간으로 상품을 판매하는 경우 시스템 반응시간은 알고리즘의 질적인 문제와 함께 매우 중요한 의미를 가지므로 알고리즘의 계산노력를 현저히 감소시킬 수 있는 방안의 연구가 필요하다.

이러한 문제를 개선하기위해 문서검색에 사용되는 LSI 방법을 구매행렬로부터 인접고객군을 발견하는 문제에 적용하여 만일 LSI에 의한 차원의 축소가 원래 고객의 상품 구매 정보에 대한 특성을 유지하면서 목표고객에 인접한 고객 군을 효과적으로 발견할 수 있다면 추천자 알고리즘의 효율성을 개선하게 될 것으로 기대할 수 있다.

또한 LSI에 의해 생성된 인접고객 군은 상품추천에 있어서 또 다른 장점이 있다고 할 수 있다. 판매되고 있는 제품의 이름은 달라도 기능 면에서는 실제로 아주 유사한 제품일 수 있다. 구매행렬로부터 연관성 분석을 통하여 상품을 추천하는 시스템은 기능 면에서 유사한 제품을 다른 제품으로 취급함으로써 두 상품 사이에 내재한 관련성을 발견할 수 없는 단점이 있으나 LSI는 고객의 상품 구매에 대한 특성을 독립적인 인자로 기술함으로써 이러한 문제를 해결할 수 있다.

3.1 상품 구매행렬에 대한 절단 SVD

고객-상품 구매행렬 A에 대한 절단 SVD을 통하여 A와 유사한 구매 특성을 가지면서 차원$(k << m)$이 축소된 행렬 A_k로 변환시켜 준다고 볼 수 있다. 구매행렬 A에 대한 SVD는 식 (1)과 같이 분해되며 식 (1)에서 $\sigma_1 \geq \sigma_2 \ldots \geq \sigma_k$이면 A에 내재된 상품 구매

특성과 가장 근사하는 차원이 $k(\leq r)$인 구매 행렬 A_k는 다음과 같다.

$$A_k = U_k \Sigma_k V_k^T \quad (3)$$

(그림 1에서 등호 우측의 색깔이 있는 부분을 차례로 U_k, Σ_k, V_k^T로 표시). 위 식에서 A_k는 n명의 고객이 가지는 m개의 상품에 대한 의견 대신 k개의 변환상품(k-meta products)에 대한 구매 특성을 나타내는 변환 구매행렬으로 a) 행렬에서 열의 값이 모두 0이 아니며, b) 구매 상품의 수가 축소됨으로($k << m$) n명의 고객을 대상으로 인접고객을 발견하는 계산 노력을 감소킬 수 있고, 또한 c) 유사한 상품 사이에 있을 수 있는 내재적인 구매 특성을 고려함으로써 상품명이 달라도 기능이 유사한 상품은 같은 인자를 가지는 상품으로 취급함으로써 구매한 상품 사이의 연관성을 발견하는데 효율적 일 수 있다.

3.2 인접고객 군의 생성

목표고객의 인접고객 군을 발견하는 과정은 공동 필터링 추천자 시스템에서 가장 중요한 부분으로서 고객 사이의 유사성과 인접고객 군은 고객의 상품-구매 변환행렬을 바탕으로 계산된다. m개의 상품에 대한 목표고객의 구매 이력(크기 m인 벡터 q)을 k개의 변환상품으로 나타내는 상품구매 변환 벡터 $\hat{q}$ 를 계산하는 과정은 식 (2)와 같다. 두 고객 사이의 유사성은 보통 두 고객이 얼마나 인접하고 있음을 측정하는 측도로 k-차원 공간상에 두 고객벡터(상품구매 변환 벡터) a 와 b사이의 cosine 각도를 기준으로 인접성을 평가한다.

$$\cos(a,b) = \frac{a \cdot b}{\|a\|\|b\|} \quad (4)$$

목표고객과 다른 고객들 사이의 유사성을 계산하여 유사성이 높은 고객들을 차례로 목표고객의 인접고객 군으로 분류한다.

3.3 상품추천서 작성

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

4. 시뮬레이션 실험

4.1 데이타 셋

E-commerce 사이트로부터 얻는 실제적인 자료와 유사한 고객의 상품 구매 자료를 시뮬레이션을 통하여 재생하였다. 개별 고객의 상품 구매수는 평균이 5인 포아슨 분포를 따르는 것으로 가정하였으며 minimum support를 만족하는 large item set(LIS)의 크기는 평균이 2인 포아슨 분포로 가정하였다. 전체 고객의 LIS는 correltion level(EXPON(0.5))을 각 고객의 large item set에 할당하여 일부 상품은 직전 고객으로부터 선택하고 나머지 상품은 임의 선택하였다. 각 고객의 LIS에 가중치를 부여하여 i번째 고객의 LSI를 구하고 여기에 일정한 수의 item을 탈락시키기 위해 확률적인 corruption level을 부여하였다. 각 고객에 대하여 최종적으로 선택된 LIS에 상품을 확률적으로 추가 구매하여 전체적으로 구매

상품 수가 Poisson(5)를 따르도록 상품 구매 행렬을 재생하였다.

4.2 평가측도

상품-구매 전체 데이터 베이스를 training set과 test set 두 부분으로 나누고 training set을 대상으로 추천자 시스템을 적용하며 목표고객에 대하여 선호 상위 *N(top-N)*개의 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 상품중에서 목표고객이 실제로 구매한 상품의 집합을 적중집합(hit set)이라고 하면 recall과 precision은 다음과 같다.

$$recall = \frac{size\ of\ hit\ set}{size\ of\ test\ set} \qquad (5)$$

$$precision = \frac{size\ of\ hit\ set}{N} \qquad (6)$$

위 두 식에서 추천 상품의 수 N이 커지면 recall은 증가하나 precision은 감소함으로 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[4].

$$F1 = \frac{2*recall*precision}{recall + precision} \qquad (7)$$

4.3 실험 결과

전체 상품의 수를 500개, 고객의 수를 1000으로 지정하고 training set과 test set을 작성하였다. 시뮬레이션 결과, 인접고객 군의 크기와 추천 상품 수에 따른 3가지 평가 측도는 <표 1>과 같다.

<표 1> 인접고객군의 크기와 추천 상품 수에 따른 평가측도

인접고객 군의 크기	추천 상품 수				
	5	10	15	20	25
10	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
20	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
30	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
40	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
50	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

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

인접고객 군의 크기가 증가함에 따라 3가지 평가 측도가 증가하나 그 크기가 일정한 값 이상이 되면 이러한 측도는 별 변화가 없을 것으로 예상된다. 또한 추천 상품의 수도 일정 수 이상이 되면 recall이 증가하는 정도가 별로 크지 않음을 보이고 있으며 시스템의 정확도는 상당히 감소하고 있음을 보이고 있다. 전체적으로 F1 측도는 추천 상품 수가 5일때 최대로 나타나고 있다.

5. 결론

본 연구는 시뮬레이션을 통하여 얻은 고객 상품 구매에 대한 가상적인 자료를 바탕으로 문서 검색에서 사용되는 LSI 기법을 추천자

시스템에 활용하였다. LSI 기법을 적용하여 얻어진 고객-상품 구매행렬은 원래의 구매행렬과 비교하여 몇 가지 장점을 가지게 될 것으로 기대하는데 우선, 원래 구매행렬에 비해 낮은 차원으로 변환된 구매행렬로 인접고객군을 발견함으로써 인접고개군을 발견하는 계산 노력을 감소할 것으로 기대되며 아울러 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 잇을 것으로 기대한다.

시뮬레이션 결과, 상품과 고객의 수가 주어지면 추천자 시스템의 recall과 정확도는 인접고객군의 크기와 추천 상품 수가 증가함에 따라 증가 또는 감소하나 그 증감 비율은 고객군의 크기나 추천 상품 수가 일정한 수준에 이르면 별로 변화가 없는 것으로 나타나고 있다. 추천자 시스템을 구현하는 용도에 따라 이러한 정보는 유용하게 사용될 수 있다고 판단된다.

참고문헌

1. Agarwal, R., Imielinski, T. and Swami, A. fast Algorithm for Mining Association Rules. 1994.

2. Berry, M. Dumais, S. and O'Brian, G. Using Linear Algebra for Intelligent Information Retrieval. SIAM Review. 37(4). pp. 573-595. 1995.

3. Resnick, P. and Varian, H. Recommender Systems. Special Issue of Communications of the ACM. 40(3). 1997.

4. Sarwar, B., Karypis, G., Konran, J. and Riedl, J. Analysis of Recommendation Algorithm for E-Commerce. Proceedings of the ACM. 2000.