웹마이닝과 상품계층도를 이용한 협업필터링 기반 개인별 상품추천시스템

안도현[a], 김재경[b], 조윤호[c]

[a,b] 경희대학교 경영대학
130-701, 서울특별시 동대문구 회기동 1번지
Tel: +82-2-961-9355, Fax: +82-2-967-0788, E-mail: {adh, jaek}@khu.ac.kr

[c] 국민대학교 e-비즈니스학부
136-702, 서울특별시 성북구 정릉동 861-1
Tel: +82-2-910-4950, Fax: +82-2-910-4519, E-mail: www4u@kookmin.ac.kr

Abstract

Recommender systems are a personalized information filtering technology to help customers find the products they would like to purchase. Collaborative filtering is known to be the most successful recommendation technology, but its widespread use has exposed some problems such as sparsity and scalability in the e-business environment. In this paper, we propose a recommendation methodology based on Web usage mining and product taxonomy to enhance the recommendation quality and the system performance of original CF-based recommender systems. Web usage mining populates the rating database by tracking customers' shopping behaviors on the Web, so leading to better quality recommendations. The product taxonomy is used to improve the performance of searching for nearest neighbors through dimensionality reduction of the rating database. Several experiments on real e-commerce data show that the proposed methodology provides higher quality recommendations and better performance than original collaborative filtering methodology.

1. 서 론

인터넷의 급속한 성장과 발달로 이를 기반으로한 전자상거래가 빠르게 증가하고 있다. 인터넷 쇼핑몰이나, 컨텐츠 제공업체(Content Provider)와 같은 기업들은 전자상거래의 급성장으로 인하여 생존우위와 정보과부하 현상을 해결하기 위한 새로운 마케팅 전략이 필요하게 되었다. 또한 고객들은 온라인기업이 증가함에 따라 다양하고, 많은 상품을 선택할 수 있는 기회가 주어지지만 본인의 요구에 가장 적합한 상품을 찾기 위해서 더 많은 정보를 처리해야 하는 부담이 생기게 되었다. 이와같은 이유로 기업은 고객별로 차별화된 원투원 마케팅(One-To-One Marketing)과 고객의 입장에서 고객을 이해하고 고객과의 관계를 강화시켜 나가는 CRM(Customer Relationship Management, 고객관계관리) 전략 등을 사용함으로써 기업의 경쟁력 강화에 주력하고 있다. 특히 CRM의 여러분야 중 전자상거래시스템에서 구매촉진(Campaign Management)을 위해 사용되고 있는 핵심기술이 추천시스템이 있다. 추천시스템은 통계적 기법과 지식탐사기술를 이용하여 고객의 요구에 가장 부합되는 상품을 추천해주는 시스템으로서, 고객들의 편의를 도모하고 교차판매 및 매출 증대에 초점을 맞춘 시스템이다[김재경 외, 2003; Sarwar, et al., 2000].

현재까지 추천시스템에 대한 연구가 활발하게 진행되고 있으며, 아마존(Amazon), 시디나우(CD Now) 등 해외의 유수한 사이트뿐만 아니라 삼성몰, 한솔CS클럽 등 국내의 일부 쇼핑몰에서도 널리 적용되고 있다. 이러한 추천시스템의 핵심은 추천알고리즘에 있으며, 현재까지 가장 선호되고 있는 추천알고리즘은 협업필터링(Collaborative Filtering: CF)으로 영화, 웹사이트, 상품, 뉴스 등 여러분야에서 사용되고 있다[Billsus & Pazzani, 1998; Cho et al., 2002; Hill et al., 1995; Lawrence et al., 2001; Resnick et al., 1994]. 협업필터링 기법은 목표고객과 유사한 선호도를 보이는 이웃고객들이 구매한 상품들 중 구매할 가능성이 가장 높은 상품을 추천하는 기법이다. 그러나 이러한 협업필터링 기법은 다른 알고리즘보다 우수함에도 불구하고 다음과 같은 문제점들이 존재한다[Cho et al., 2002; Sarwar et al., 2000]. 첫째, 입력 데이터의 희박성(sparsity) 문제이다. 웹사이트에서 판매되는 상품의 수가 기하급수적으로 증가함에 따라 고객의 선호도가 입력되지 않은 상품의 개수가 상대적으로 많아짐으로 인해 이웃 고객군을 형성하는 과정에서 아주 적은 수의 평가 데이터만을 사용함으로써 유사도 측정에 신뢰성이 떨어지고, 이는 결국 상품 추천의 질을 떨어뜨리는 요인으로 작용한다. 둘째, 시스템의 확장성(scalability) 문제이다. 고객과 상품의 수가 증가함에 따라 이웃 고객군을 찾기 위한 연산량은 기하급수적으로 늘어날 수 밖에 없기 때문에 실시간으로 추천을 목적으로 하는 상품추천시스템에서는 심각한 시스템 확장성 문제에 직면하게 된다.

따라서 본 연구에서는 협업필터링의 문제점들을 해결하기 위해 웹마이닝과 상품계층도를 이용하여 협업필터링 기반 상품추천시스템의 효과 및 성능을 개선하는 방법을 제안하고자 한다.

2. 관련 연구

2.1 상품추천시스템

협업필터링(Collaborative Filtering: CF)은 웹을 기반으로 하는 전자쇼핑몰에서 이용되고 있는 성공적인 상품추천기법중의 하나로써, 목표고객과 유사한 구매이력을 보이는 이웃 고객들의 상품에 대한 선호를 바탕으로 목표고객에게 유용한 상품을 추천하는

방법이다[김재경 외, 2003; Resnick et al., 1994; Sarwar et al., 2000]. 일반적으로 이러한 협업필터링 기반 상품추천과정은 크게 입력 데이터 구성, 이웃 집단 탐색, 추천 상품 결정 단계로 나눠볼 수 있으며, 이러한 과정을 자세히 살펴보면 다음과 같다.

1) 입력 데이터 구성(Representation): 협업필터링 기반 상품추천시스템에서의 입력데이터는 보통 n개의 상품에 대하여 m명 고객의 구매 트랜잭션의 집합으로 구성되며, 보통 $n \times m$의 고객-상품 행렬 R로 표현될 수 있다.
2) 이웃 집단 탐색(Neighborhood Formation): 고객간의 유사도를 계산하여 이웃 집단을 탐색하는 과정이다. 두 고객 a와 b의 유사도를 측정하는 방법으로써 피어슨 상관계수(Pearson Correlation), 코사인(Cosine) 등을 사용한다.
3) 추천 상품 결정(Generation of Recommendation): 상품추천을 위한 마지막 단계로서 설정된 이웃 집단으로부터 상위 N개의 추천 상품 목록을 이끌어내는 단계이다.

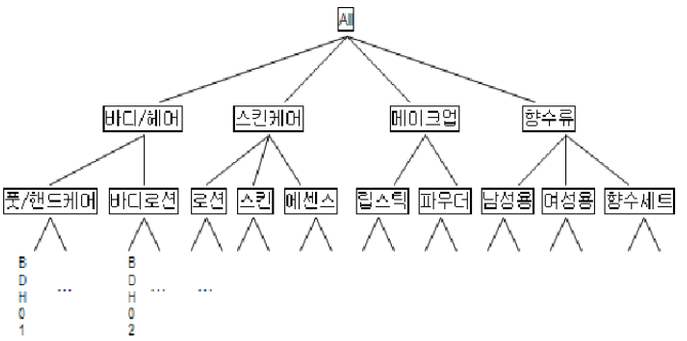
그러나 이러한 협업필터링 기반 추천시스템은 인터넷 쇼핑몰의 상품과 고객수의 급속한 증가로 인해 입력 데이터의 희박성과 시스템 확장성 문제를 노출시키고 있다[Billsus & Pazzani, 1998; Sarwar et al., 2000a].

2.1 웹마이닝

인터넷 쇼핑몰에 있어서 획득 가능한 고객정보에는 오프라인 기업에서와 같은 구매이력 외에도 고객들이 쇼핑몰 사이트를 방문할 때 발생하는 웹로그 정보가 있다. 이 정보는 인터넷 기업에 있어서 고객분석을 할 때 간과해서는 안될 매우 중요한 정보이며, 특히 상품추천의 경우, 고객의 성향이나 상품선호도 분석에 있어 핵심적인 기초정보로 활용할 수 있다. 따라서 본 연구에서는 이러한 웹로그 정보를 적극적으로 상품추천에 활용하기 위해 웹마이닝 기법을 이용한다. 웹마이닝(Web Mining)은 웹 사이트에서 고객이동경로 즉 클릭스트림 정보를 가지고 있는 웹로그로부터 고객들의 접속관계· 패턴· 규칙 등을 찾아내고 모형화해 유용한 마케팅정보로 변환시키는 일련의 과정을 말한다[Mobasher et al., 2000]. 웹마이닝의 전체 프로세스는 데이터 전처리와 패턴발견 두 가지로 분류된다. 데이터 전처리는 획득한 웹로그 데이터를 분석에 적합한 형태로 정제하여 불필요한 데이터를 제거함으로써 원천데이터의 용량을 감소시킬 수 있다[Cooley et al., 1999]. 두번째, 패턴분석과정은 연관성규칙, 연속적인 패턴발견등을 이용하여 고객의 쇼핑행위에서 구매 패턴을 발견할 수 있다[Mobasher et al.,2000]. 이러한 웹마이닝을 통하여 기존에 발견할 수 없었던 고객선호도를 보다 많이 확보할 수 있으므로, 협업필터링 기반 상품추천시스템의 데이터 희박성 문제를 해결 할 수 있을 것으로 기대된다.

2.3 상품계층도

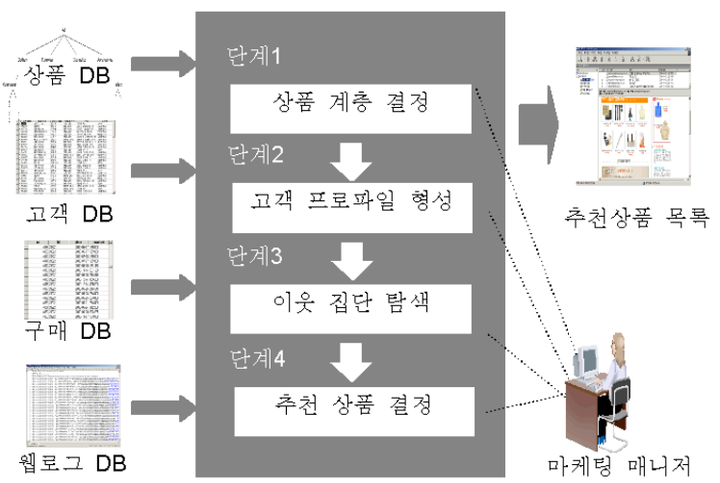
상품계층도는 개별상품을 추상 개념이 낮은 상품 클래스로 분류하고 이들 상품 클래스를 다시 추상 개념이 좀더 높은 상품 클래스로 분류한 계층 구조를 말하며 일반적으로 개별 상품, 상품 카테고리(Category), 상품군 등의 순서로 형성된다. [그림 1]은 상품계층도의 예를 보여주고 있다. 최근 많은 데이터 마이닝 관련 연구들이 데이터 분석에서의 상품 계층도의 필요성 및 중요성을 지적하고 있다[Han & Fu, 1999; Lawrence, et al., 2001]. 상품계층도상에서 유사한 선호도 패턴을 갖는 개별상품들을 특정 상품군으로 군집화하여 입력 데이터의 차원을 축소하면 이웃 집단 탐사 과정에서 계산속도의 향상을 도모하여 시스템 확장성 문제를 해결 할 수 있을 것이다.



[그림 1] 상품계층도

3. 제안하는 상품추천방법

본 연구에서는 기존 협업필터링의 문제점들을 해결하기 위해 웹마이닝과 상품계층도를 이용한 협업필터링 기반 상품추천방법(Web-based Collaborative Filtering using Product Taxonomy: WebCF-PT)을 소개한다. [그림 2]는 WebCF-PT추천방법의 개략적 절차를 나타내며, 첫 단계는 '상품 계층 결정' 단계로 데이터베이스 내의 모든 상품을 마케팅 전문가가 특정그룹으로 분류· 재구성하여 입력 데이터를 감소시키는 단계이다. 두번째 단계는 '고객 프로파일 형성' 단계로 인터넷 쇼핑몰에서 개별고객의 쇼핑행위를 추적하여 얻은 데이터를 통해 목표고객의 상품 선호도 정보를 발견하고, 분석한 정보를 이용하여 고객 프로파일을 형성하는 단계이다. 세번째 단계는 '이웃 집단 탐색' 단계로 형성된 고객 프로파일을 이용하여 고객간의 유사도를 계산하고 이를 기반으로 목표고객과 유사한 성향을 가진 고객들을 선택하여 이웃을 형성하는 단계이다. 마지막 단계는 '추천 상품 결정' 단계로 형성된 이웃들의 쇼핑행위를 기반으로 선호도가 높은 상위 N 개의 상품을 선택하여 추천 상품의 목록을 결정하는 단계이다.



[그림 2] WebCF-PT추천시템의 개략적 절차

4. 실험 및 평가

4.1 데이터

실험 및 평가를 위해서 다양한 여성 상품을 판매하는 C인터넷 쇼핑몰의 웹로그 데이터와 상품 데이터를 사용했다.

웹로그 데이터: 2001년 5월 1일부터 30일까지 IIS 4 대의 웹서버로부터 124 개 웹로그 파일을 수집하였다. 웹로그 파일의 전체 크기는 약 64,730MB이고 전체 HTTP 요청수는 약 420,000,000,000건 이다. 웹로그 파일들은 데이터 전처리 작업을 통해 시간, 고객 ID, 상품 ID, 그리고 쇼핑 단계 정보가 있는 거래 데이터베이스를 구축하였다. 이 데이터베이스는 66,329 명 거래 고객 데이터를 가지고 있다. 전체 2,249,540개 레코드 중에 7208개의 구매 레코드, 60,892개의 장바구니 레코드, 그리고 2,181,440 개의 클릭 레코드가 있다. 2001년 5월 1일부터 24일까지는 학습 기간으로 사용하고, 25일부터 30일까지는 테스트 기간으로 사용했다. 위의 두 기간 동안 최소 1번 이상 구매한 116 명 고객을 목표고객으로 정했다. 최종 목표고객으로 생성된 학습 셋에는 8960개 거래 레코드와 156개 구매 레코드가 있다.

상품 데이터: C 인터넷 쇼핑몰은 3216 가지 상품을 판매하고 있다. 상품계층도는 3개의 계층으로 되어 있다. 상위계층은 10가지 상품클래스, 중간계층은 72가지 상품클래스, 하위계층은 3216가지 상품이 있다.

4.2 평가 기준

정확도 평가 기준: *Recall*은 고객이 실제 구매한 상품중에 추천한 상품이 얼마나 되는 것을 의미하고, *Precision*은 추천한 상품중에 고객이 실제로 얼마나 구매 했는가를 의미한다. 이러한 기준은 계산도 간단하고 효과적이지만 추천 셋의 크기가 증대함에 따라 *Recall*은 증가되지만 *Precision*은 오히려 줄어드는 문제를 안고 있다 [Sarwar et al., 2000]. 따라서 이 연구는 Billsus & Pazzani[1998] 등이 이용했던 *Recall*과 *Precision*의 조화평균인 *F1* 을 정확도 평가 기준으로 채택했다.

$$F1 = \frac{2 \times recall \times precision}{recall + precision} \quad (1)$$

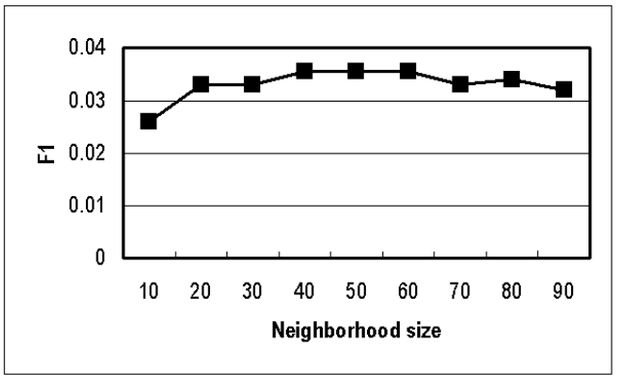
성능 평가 기준: 시스템 확장성(scalability) 문제를 평가하기 위해 시스템 성능 평가 기준으로 응답 시간(Response time)을 사용하였다. 응답 시간은 학습 셋에서 상품추천까지의 전과정을 계산하는데 걸리는 시간을 의미한다.

4.3 실험 결과

실험을 통해 최적의 파라미터를 선정한 후 전통적 협업필터링 기반 방법과 본 연구에서 제안하는 WebCF-PT 방법과의 정확도 및 성능을 비교하였다.

4.3.1 이웃 크기에 따른 정확도

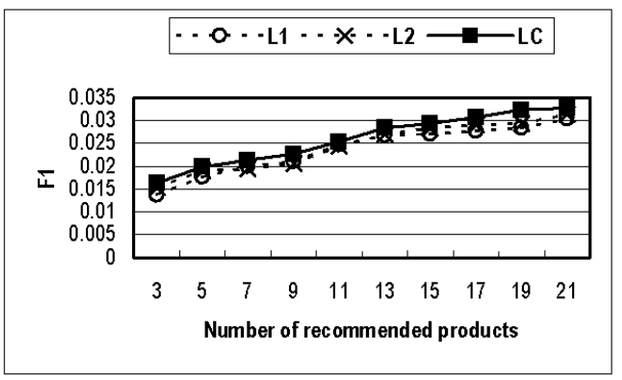
이웃 크기는 추천 정확도에 중요한 영향을 준다는 연구[Herloker et al., 1999] 결과에 따라 [그림 3]와 같이 이웃 크기에 따른 추천 결과의 정확도를 실험하였다. 실험 결과를 보면, 이웃 크기에 따라 추천 정확도가 크지는 않지만 변화하는 것을 알 수 있다. 일반적으로 추천 정확도는 이웃 크기가 클수록 정확도는 높아지지만, 어느정도 고점에 도달하면 점점 낮아진다. 이 실험에서는 이웃 크기 50일 때 정확도가 가장 높으므로 다음 단계 실험에서는 이웃 크기 파라미터를 50으로 고정하여 실험을 진행하였다.



[그림 3] 이웃 크기에 따른 정확도

4.3.2 상품 계층에 따른 정확도

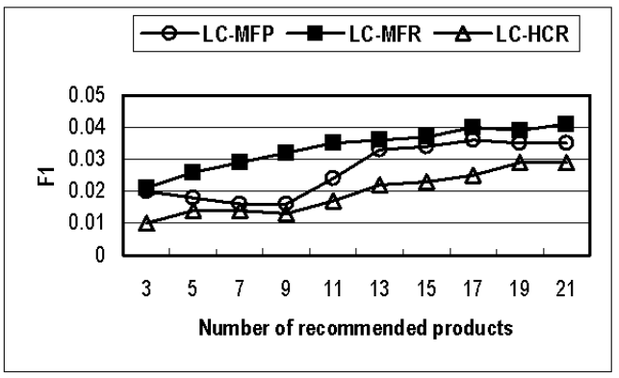
상품 계층에 따른 정확도를 평가하기 위해서 세가지 유형의 상품 계층에 대해 실험하였다. 높은 수준의 그레인(L1), 낮은 수준의 그레인(L2), 혼합 수준의 그레인(LC)으로 나누어 정확도를 평가 하였으며, [그림 4]는 이러한 세가지 유형에 따라 실험한 결과를 보여주고 있다. 결과를 분석해 보면, 세가지 유형 중에서 혼합 수준의 그레인(LC)의 정확도가 조금 높기 때문에 다음 단계의 실험에서는 혼합 수준의 그레인(LC)를 선택하여 실험을 진행하였다.



[그림 4] 상품 계층에 따른 정확도

4.3.3 추천 상품 결정 방법에 따른 정확도

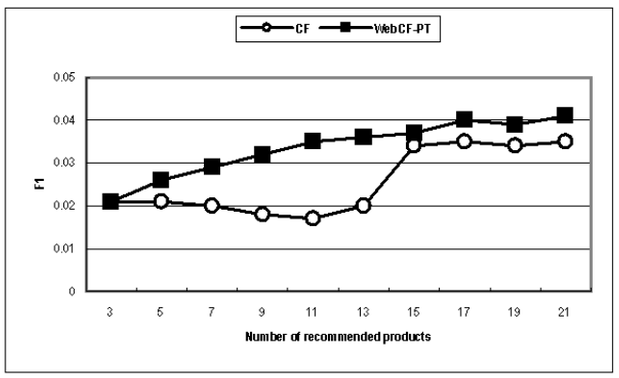
상품 추천 방법론의 마지막 단계인 추천 상품 결정 방법을 결정하기 위해 MFP, MFR, HCR 세가지 방법에 대해 상대적 정확도를 실험하였다. [그림 5]는 이러한 실험 결과를 보여주고 있으며, MFR 방법의 정확도가 상대적으로 가장 높은 것으로 나타났다.



[그림 5] 추천 상품 결정 방법에 따른 정확도

4.3.4 전통적 CF 방법과 WebCF-PT 방법의 정확도 비교

전통적 CF 방법과 WebCF-PT 방법의 정확도를 비교하기 위해 앞의 실험을 통해 결정한 이웃 크기와 상품 계층 그리고 추천 상품 결정 방법 중 MFR 을 이용하여 실험하였다. [그림 6]은 전통적 CF 방법과 WebCF-PT 방법의 정확도를 실험을 통해 비교한 결과를 보여준다. 그래프를 분석해 보면, 제안한 WebCF-PT 방법이 전통적 CF 방법 보다 정확도가 평균 32% 우수하다는 것을 알 수 있다.



[그림 6] 전통적 CF 방법과 WebCF-PT 방법의 정확도 비교

4.3.5 전통적 CF 방법과 WebCF-PT 방법의 성능 비교

전통적 CF 방법과 WebCF-PT 방법의 성능을 비교하기 위하여 5.2.2장에서 설명한 응답 시간 (Response time)을 평가 기준으로 이용하였다. 실험 결과는 [표 1]에 나타나듯이, 제안한 WebCF-PT 방법이 전통적 CF 방법 보다 약 18배 정도 빠른 응답을 보였다. 이러한 결과는 전자상거래의 빠른 성장으로 인한 고객과 상품 수의 급속하게 증가되고 있는 상황에서 중요한 의미를 가지고 있다.

[표 1] 전통적 CF 방법과 WebCF-PT 방법의 성능 비교

	CF	WebCF-PT
응답 시간(초)	91.53	4.87

5. 결 론

전자상거래의 급성장으로 기업들의 생존을 위한 경쟁은 더욱 심화되어 다른 경쟁업체보다 경쟁우위를 가질 수 있는 마케팅 전략이 필요하게 되었고, 고객은 상품 정보의 과다로 인하여 효과적으로 상품을 선택할 수 없게 되는 상품 과부하 현상을 야기시켰다. 이러한 문제를 해결하기 위한 정보 기술 중의 하나가 고객의 취향에 부합하는 상품을 찾도록 도와 주는 상품추천시스템이다. 현재까지 상품추천시스템을 구현하기 위한 다양한 기술들이 개발되어 왔는데, 이중에서 협업필터링이 가장 성공적인 상품추천 기술로 알려지고 있다. 협업필터링 기법은 목표고객과 유사한 선호도를 보이는 이웃고객들이 구매한 상품들 중 구매할 가능성이 가장 높은 상품을 추천하는 기법이다. 그러나 이러한 협업필터링 기법은 다른 알고리즘보다 우수함에도 불구하고 다음과 같은 문제점들이 존재한다. 첫번째는 입력 데이터의 희박성 문제이고, 두번째는 시스템 확장성 문제이다. 본 연구에서는 이러한 협업필터링의 문제점들을 해결하기 위해 웹마이닝과 상품계층도를 활용한 협업 추천방법론을 제시하고 구현하였으며, 실제 전자상거래 데이터를 이용하여 기존 협업필터링 방법론과 실험적으로 비교하였다. 실험 결과 제시한 방법론이 기존 협업필터링 방법론 보다 우수한 성능 결과를 나타내었다. 하지만 실험 데이터 셋의 한계가 있을 수 있으므로 추후에는 다양한 데이터 셋과 영역에 제시한 방법론의 적용과 다른 방법론과의 비교 연구가 필요하며, 실제 마케팅 캠페인 프로젝트를 수행하며 제시한 방법론을 적용하고자 한다.

참고문헌

1. 김재경, 안도현, 조윤호, “Development of a Personalized Recommendation Procedure Based on Data Mining Techniques for Internet Shopping Malls”, 한국지능정보시스템학회, 제9권, 제3호, 2003, pp.177-191.
2. Billsus, D., and Pazzani, M. J., “Learning collaborative information filters”, In Proc. 15th International Conference on Machine Learning, 1998, pp.46-45.
3. Cho, Y. H., Kim, J. K., and Kim, S. H., “A personalized recommender system based on Web usage mining and decision tree induction”, Expert Systems with Applications, 2002, Vol.23, No.3, pp.329-342.
4. Cooley, R., Mobasher, B., and Srivastava, J., “Data Preparation for Mining World Wide Web Browsing Patterns”, Knowledge and Information Systems, 1999, Vol.1, No.1.
5. Han, J. and Fu, Y., “Mining Multiple-Level Association Rules in Large Databases”, IEEE Transaction on Knowledge and Data Engineering, 1999, Vol.11, No.5, pp. 798-804.
6. Herlocker, J. L., Konstan, J. A., Borchers, A., Riedl, J., “An Algorithmic Framework for Performing Collaborative Filtering”, In Proc. Conference on Research and Development in Information Retrieval, 1999, pp. 230-237.
7. Hill, W., Stead, L., Rosenstein, M., Furnas, G. W., “Recommending and Evaluating Choices in a Virtual Community of Use”, In Proc. Human Factors in Computing Systems, 1995, pp. 194-201.
8. Lawrence, R. D., Almasi, G. S., Kotlyar, V., Viveros, M. S., and Duri, S. S. “Personalization of supermarket product recommendations”, Data Mining and Knowledge Discovery, 2001, Vol.5, No.1, pp.

11-32.
9. Mobasher, B., Cooley, R., and Srivastava, J., "Automatic Personalization based on Web usage mining", Communications of the ACM, 2000, Vol.43, No.8, pp. 142-151.
10. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J., "GroupLens: An Open Architecture for Collaborative Filtering of Netnews", In Proc. Computer Supported Cooperative Work, Chapel Hill, NC, 1994, pp.175-186.
11. Sarwar, B., Karypis, G., Konstan, J. A., & Riedl, J., "Analysis of recommendation algorithms for e-commerce", In Proc. ACM E-Commerce, 2000, 158-167.