조작된 선호도에 강건한 협업적 여과 방법

A Robust Collaborative Filtering against Manipulated Ratings

김 흥 남* 하 인 애** 조 근 식***
Heung-Nam Kim Inay Ha Geun-Sik Jo

요 약

협업적 여과는 추천 시스템을 구축하는데 가장 널리 보급된 정보 여과 기법으로 사용자 각 개인의 관심에 적합한 정보 및 아이템을 추천함으로써 사용자들의 의사 결정에 도움을 준다. 그러나 협업적 여과 기법은 우수한 추천 성능에도 불구하고, 최근에는 실링 공격이라 일컫는 악의적인 목적을 가진 사용자들의 추천 결과 조작에 쉽게 노출될 수 있는 문제가 새로운 이슈로 대두되고 있다. 본 논문에서는 협업적 여과의 실링 공격 문제들을 보완하기 위해, 추천 시스템에서 발생할 수 있는 실링 공격의 유형을 분석하고 악의적인 사용자의 조작된 선호도가 시스템에 미치는 영향을 최소화하기 위한 강건한 신뢰 모델 구축 방법을 제시한다. 그리고 그 모델을 적용하여 신뢰할 수 있는 아이템 추천 및 선호도 예측 방법을 제안한다.

ABSTRACT

Collaborative filtering, one of the most successful technologies among recommender systems, is a system assisting users in easily finding the useful information and supporting the decision making. However, despite of its success and popularity, one notable issue is incredibility of recommendations by unreliable users called shilling attacks. To deal with this problem, in this paper, we analyze the type of shilling attacks and propose a unique method of building a model for protecting the recommender system against manipulated ratings. In addition, we present a method of applying the model to collaborative filtering which is highly robust and stable to shilling attacks.

☞ KeyWords : 협업적 여과(collaborative filtering), 추천 시스템(recommender system), 실링 공격(shilling attack), 선호도 조작(manipulated rating)

1. 서 론

협업적 여과(collaborative filtering)는 추천 시스템을 구축하는데 가장 널리 보급된 정보 여과 기법으로 학문적으로뿐만 아니라 상업적으로도 가장 많이 이용되고 있다. 이는 특정 사용자와 유사한 성향을 가진 다른 사용자들의 의견을 기반으로 하는 사회적 정보 여과 방법으로써 자동화된 "Word of Mouth" 현상을 이용한 추천이라 할 수 있다[13]. 최초로 유즈넷 기사를 자동으로 추천한 GroupLens 시스템[5][11]의 개발 이후 사용자의 과거 구매 정보를 바탕으로 책을 추천해주는 Amazon.com[1], 사용자의 선호도(rating) 점수를 활용하여 DVD를 추천해주는 온라인 비디오 대여 서비스 Netflix.com[2], 사용자가 선택하여 들은 음악을 기반으로 추천해주는 last.fm[3], 자동화된 온라인 뉴스 서비스 Google News[4] 등 협업적 여과는 다양한 도메인의 추천 시스템에 성공적으로 적용되어 사용되고 있다.

* 정 회 원 : 인하대학 BK21 박사후 연구원
nami@eslab.inha.ac.kr(교신저자)
** 정 회 원 : 인하대학교 대학원 컴퓨터정보공학과
박사과정 inay@eslab.inha.ac.kr
*** 정 회 원 : 인하대학교 컴퓨터정보공학과 교수
gsjo@inha.ac.kr
[2009/04/06 투고 - 2009/04/15 심사 - 2009/05/28 심사완료]

1) http://www.amazon.com
2) http://www.netflix.com
3) http://www.last.fm
4) http://news.google.com

협업적 여과는 쉽게 분석될 수 없는 아이템들을 사용자 커뮤니티의 정보, 즉 사용자들의 선호도를 통해 사용자가 원하는 아이템을 선별하고 추천한다는 점에서 널리 사용됨에도 불구하고, 실제 응용에서 온라인 추천 시스템으로부터 얻어지는 추천을 신뢰할 수 있는가에 대한 의문이 제기될 수 있다. 그리고 최근에는 실링 공격(shilling attack)[6][9]이라 일컫는 악의적인 목적을 가진 사용자들의 추천 결과 조작에 쉽게 노출될 수 있는 문제가 새로운 이슈로 대두되고 있다.

실링 공격이란 자신의 이익과 관련된 아이템의 추천을 높이거나 경쟁 상대의 아이템의 추천을 저해시키기 위해, 또는 전체 시스템의 추천의 성능을 저해시키기 위한 목적으로 악의적인 사용자들이 본인들의 선호도를 조작하는 것을 말한다[6]. 아이템의 생산자들은 온라인 시장에서 그들의 상품이 경쟁사보다 많이 팔리기를 원한다. 따라서 일부 생산자들은 상품의 품질과 상관없이 자신들의 아이템이 많이 추천될 수 있게 부도덕한 방법으로 시스템에 영향을 미치게 하려고 노력한다. 그 예로, 생산자와 관련된 악의적인 사용자들이 자신들의 선호도를 거짓으로 시스템에 제공하여 자신의 상품 추천을 높이거나 경쟁사의 상품 추천을 낮추도록 유도할 수 있다. 이는 일반 사용자들의 입장에서는 잘못된 아이템 추천으로 인한 시간과 돈의 낭비를 초래할 수 있으며, 시스템 운영자의 입장에서는 일반 사용자의 시스템에 대한 신뢰가 떨어지게 된다.

이러한 협업적 여과의 실링 공격 문제들을 보완하기 위해, 본 논문은 추천 시스템에서 발생할 수 있는 실링 공격의 유형을 분석하고 악의적인 사용자의 조작된 선호도가 시스템에 미치는 영향을 최소화하기 위한 강건한 신뢰 모델 구축 방법을 제시한다. 그리고 그 모델을 적용하여 신뢰할 수 있는 아이템 추천 및 선호도 예측 방법을 제안한다.

본 논문의 전체 구성은 다음과 같다. 2장에서는 이론적 고찰로 협업적 여과 방법에서 발생할 수 있는 악의적인 사용자들의 조작된 선호도 삽입에 대한 문제를 논한다. 3장에서는 신뢰 모델을 구축하는 방법에 대하여 기술하고, 신뢰 모델을 적용함으로써 강건한 협업적 여과 및 추천 방법을 기술한다. 4장에서는 실험을 통해 제안된 신뢰 모델을 적용한 협업적 여과 방법들의 예측 성능, 악의적인 공격에 대한 강건성을 기존의 협업적 여과 방법들의 성능과 비교 분석하여 그 성능을 입증한다. 마지막으로 5장에 결론을 맺고 향후 연구에 대해 논한다.

2. 배경 지식 및 문제점 기술

2.1 협업적 여과 기반의 추천 시스템

비록 협업적 여과 기법은 다양한 도메인의 정보 여과 문제에 사용되며 연구되고 있지만, 일반적으로 사용자들의 아이템에 대한 각 개인의 과거 선호도 평점(rating) 정보를 사용한다. m명의 사용자 집합 $U=\{u_1, u_2, \cdots, u_m\}$의 각 사용자가 n개의 아이템 집합 $I=\{i_1, i_2, \cdots, i_n\}$의 각 아이템에 대한 선호도는 $m \times n$ 사용자-아이템 선호도 행렬(user-item rating matrix) $\mathbf{R}$로 표현할 수 있다[4]. 여기서 행은 사용자를, 열은 아이템을 의미하며, $R_{u,j}$는 사용자 u의 아이템 j에 대한 수치적 선호도 값을 나타낸다. 행렬의 요소 $R_{u,j}$는 양수 R_{min}에서 양수 R_{max}사이의 정수 값 또는 아직 평가하지 않은 경우 중 하나에 속한다.

$$R_{u,j} \in \{ R_{min}, \ldots, R_{max}, \varnothing \}$$

사용자 u가 아이템 j에 대하여 R_{min}에 가까운 선호도 점수를 부여하였다면 그 아이템은 사용자 u의 취향에 맞지 않음을 의미한다. 반대로 R_{max}에 가까운 선호도 점수로 평가하였다면 자신의 취향에 부합되는 아이템임을 의미한다. 만약 $R_{u,j}$의 값이 $\varnothing$라면 사용자 u는 아직 아이템 j에 대해 평가하지 않았음을 의미하며, 궁극적으로 사용자 u에 대하여 $R_{u,*}=\varnothing$인 아이템들이 추천 대상이 되는 아이템들이라 말할 수 있다.

2.1.1 사용자 기반의 협업적 여과

Resnick et al.[11]에 의해 제안된 전통적인 사용자 기반의 협업적 여과는 사용자 기반의 관점에서 어떤 아이템에 대한 목적사용자(target user)의 선호도 점수를 예측하는 추천 방법이다.

우선, *K*명의 근접 이웃(K Nearest Neighbors)으로 지칭되는 유사한 사용자 집단 KNN을 구성한다. 이를 위해 각 사용자 간의 유사성 정도를 측정하는데, 식 1과 같이 두 사용자가 서로 얼마나 관계가 있는가를 판단하는 피어슨 상관 계수(Pearson correlation coefficient)[1] 기반의 유사도 측정 방법이 가장 대표적으로 사용된다.

$$sim(u,w)=\frac{\sum_{i\in I}(R_{u,i}-\overline{R_u})(R_{w,i}-\overline{R_w})}{\sqrt{\sum_{i\in I}(R_{u,i}-\overline{R_u})^2}\sqrt{\sum_{i\in I}(R_{w,i}-\overline{R_w})^2}} \quad (1)$$

사용자 *u*와 *w*의 피어슨 상관 계수는 두 사용자가 공통으로 선호 평가한 아이템 집합 I에 포함된 아이템들에 대한 각 사용자의 선호도 점수를 활용한다. 즉, $R_{u,i}$와 $R_{w,i}$는 각 사용자 *u*, *w*의 아이템 *i*에 대한 선호도를 수치로 나타낸 것이다. 그리고 $\overline{R_u}$, $\overline{R_w}$는 각 사용자의 선호 평균점수 나타낸다. 계산 결과는 -1에서 1 사이의 실수로, *u*의 점수가 증가하면 *w*의 점수도 증가하는 양의 상관관계를 갖는 경우 1, 음의 상관관계인 경우 -1에 가까우며, 전혀 관계가 없을 경우 0을 반환한다.

위와 같은 방법으로 목적사용자와 유사한 사용자들을 이웃 집단으로 구성하며, 이들의 유사도 값과 선호도 점수는 아직 선호 평가하지 않은 특정 아이템에 대한 목적사용자의 선호도가 어느 정도인지를 예측하는 데 사용된다. 선호도 예측에 사용되는 대표적인 예측 방법은 다음과 같다.

$$\overset{\vee}{R}_{u,j}=\overline{R_u}+\frac{\sum_{w\in KNN}(R_{w,j}-\overline{R_w})\cdot sim(u,w)}{\sum_{w\in KNN}|sim(u,w)|} \quad (2)$$

2.1.2 아이템 기반의 협업적 여과

아이템 기반의 협업적 여과의 주요 동기는 사용자가 과거에 이미 선호했던 아이템들과 유사한 아이템들을 선호할 가능성이 높다는 사실에 있다. 따라서 사용자 간의 유사도가 아닌 아이템 간의 유사도를 계산하여 얻어진 *K*개의 유사한 아이템 집합(K Most Similar Items)을 바탕으로 미리 특정 아이템에 대한 유사도 모델을 구축한다는 점에서 사용자 기반의 협업적 여과와 구별된다.

두 아이템 *i*와 *j*의 유사도 측정식은 사용자 전체 집합 U에 포함된 모든 사용자가 두 아이템 *i*와 *j*에 매긴 선호도 점수에 따라 코사인 기반을 이용하여 다음과 같이 계산될 수 있다[12].

$$sim(i,j)=\cos(\vec{i},\vec{j})=\frac{\sum_{u\in U}R_{u,i}\cdot R_{u,j}}{\sqrt{\sum_{u\in U}R_{u,i}^{\,2}}\sqrt{\sum_{u\in U}R_{u,j}^{\,2}}} \quad (3)$$

Sawar et al.[12]에 의해 제안된 목적사용자 *u*의 특정 아이템 *j*에 대한 예측 점수 $\overset{\vee}{R}_{u,j}$는 식 4와 같이 유사 아이템 집합 MSI에 포함된 모든 아이템 *i*에 대한 *u*의 점수에 아이템 *i*와 *j*의 유사도를 가중치로 하여 계산한다.

$$\overset{\vee}{R}_{u,j}=\frac{\sum_{i\in MSI}sim(i,j)\cdot R_{u,i}}{\sum_{i\in MSI}|sim(i,j)|} \quad (4)$$

아이템 기반의 협업적 여과는 미리 유사도 모델을 구축하여 사용함으로써 사용자 기반의 방법에 비해 매우 빠른 추천이 가능하다. 또한 아이템 집합의 변화는 사용자 집합의 변화에 비해 매우 적으므로, 이미 구축되어 있는 아이템 유사도 모델을 재사용하더라도 초기의 추천 성능에 비해 정확도가 떨어지지 않는다. 따라서 모델의 구축은 추천과 함께 실시간으로 이루어질 필요가 없다[3]. 그러나 모델을 구축하는 데 많은 시간이 소요되며, 사용자의 새로운 선호도를 실시간으로 반영하지 못한다는 단점을 가지고 있다.

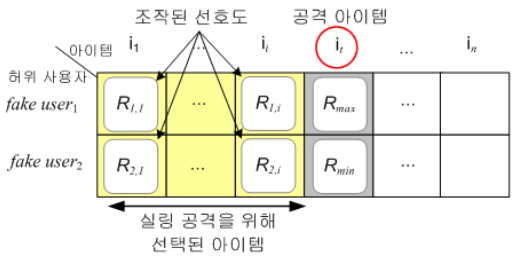
2.2 협업적 여과에서 추천 신뢰성

협업적 여과는 다수의 사용자가 제공하는 방대한 양의 정보에서 비교적 손쉽게 의미 있는 정보

를 얻을 수 있도록 하고, 이를 통해 사용자 간의 의견 교환을 가능하게 한다는데 있어서 큰 이점을 갖는다. 그러나 협업적 여과 추천 시스템에 정보를 제공하는 익명의 사용자는 동일한 의견을 갖는 다수의 사용자 프로파일을 생성하거나, 무의미한 정보를 갖는 프로파일을 생성함으로써 추천에 영향을 줄 수 있는 신뢰성의 문제가 제기된다. 또한 상업적인 목적을 갖는 추천의 제공자가 자신에게 유리하거나, 혹은 경쟁자에게 불리한 의견을 갖는 프로파일을 삽입하여 이익 창출에 도움을 얻고자 시도할 가능성이 있다. 실제로 최근의 많은 연구에서, 널리 사용되고 있는 추천 알고리즘이 평범한 수준의 공격에도 추천 결과의 조작이 가능하다는 사실이 증명되었다[7][8].

본 논문에서 다루는 사용자 공격은 전통적인 정보 보안 분야에서 언급되는 서비스 거부(denial of service), 시스템 해킹, 패스워드 크래킹(cracking) 등을 지칭하고 있지 않다. 일반적으로 추천 시스템에서 지칭되는 실링 공격(shilling attack)[6][9] 및 프로파일 삽입 공격(profile injection attack)[7][8][14]을 일컫는다. 즉, 추천 시스템에서 시스템 운영자가 아닌 사용자들이 특정 상품의 추천을 조작하기 위한 목적으로 새로운 프로파일을 삽입하는 것을 말한다.

본 논문에서 사용하는 허위사용자(fake user)란 시스템에서 특정 아이템의 추천을 인위적으로 높이거나 낮추기 위한 목적으로 일반 아이템들에 대하여 조작된 선호도를 평가하는 사용자를 의미한다. 또한 시스템의 전반적인 추천 품질을 낮추기 위한 의도로 자신의 선호도를 조작하는 사용자도 포함한다. 그리고 조작된 선호도(manipulated rating)란 허위 사용자들에 의해 생성된 선호도를 일컫는다. 마지막으로 공격 아이템(attack item)은 허위 사용자들에 의해 추천이 조작되는 아이템을 일컫는다[2].



(그림 1) 허위 사용자의 공격 아이템에 대한 조작된 선호도

2.2.1 공격 유형과 모델

실링 공격의 유형에는 크게 3가지로 언급된다. 첫째, 공격 아이템의 추천 및 선호도 예측을 높이기 위한 push 공격. 둘째, 공격 아이템의 추천 및 선호도 예측을 낮추기 위한 nuke 공격. 그리고 마지막으로 전체 시스템의 추천 및 예측의 성능을 저해시키기 위한 시스템 품질 공격(quality of recommendations)이다[6].

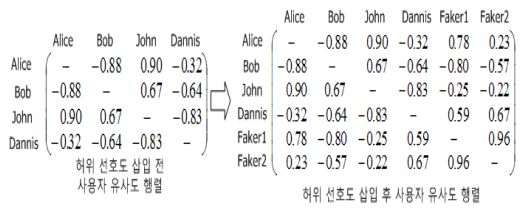
전형적인 실링 공격의 유형인 push 공격은 허위 사용자들이 공격 아이템의 추천을 높이거나 예측 선호도 값을 높이기 위한 목적으로 프로파일을 조작하여 삽입하는 것을 일컫는다[8]. 특정 아이템의 생산자 및 그와 이익이 관련된 사람들은 그들의 아이템이 일반 사용자들에게 많이 노출되거나 추천되어 판매 이익을 높이기를 원한다. 이를 위해, 허위 사용자들은 특정 아이템, 즉 공격 아이템의 선호도를 최고 평점으로 부여하고 전체 아이템들 중의 일부에 대하여 조작된 선호도를 생성한다. 이렇게 함으로써, 허위 사용자들과 유사한 사용자들의 공격 아이템에 대한 선호도 예측시 자신들의 최고 점수가 반영되어 높은 예측 점수를 유도한다.

표 1은 사용자-아이템 선호도에서 허위 사용자들과 관련된 영화 “타짜”가 일반 사용자에게 많이 관람되기를 바라며 Faker1과 Faker2의 프로파일이 추가된 예이다.

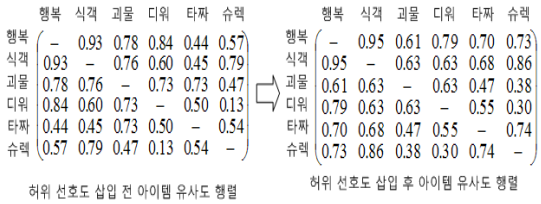
(표 1) push 공격을 위한 허위 사용자 프로파일 예 (Rmin=1, Rmax=5)

	행복	식객	괴물	디워	타짜	슈렉
Alice	3	5	-	1	?	5
Bob	4	4	5	5	3	-
John	2	5	4	-	1	4
Dannis	-	1	2	-	5	4
Faker1	3	4	-	3	5	4
Faker2	3	4	-	-	5	4

그리고 그림 2와 그림 3은 각각 허위 사용자들이 삽입되기 전과 삽입된 후의 사용자-사용자 유사도 행렬과 아이템-아이템 유사도 행렬을 보여주고 있다.



(그림 2) 허위사용자들에 의해 변경되는 사용자 간의 피어슨 상관 계수



(그림 3) 허위사용자들에 의해 변경되는 아이템 간의 코사인 유사도

그림에서 보이는 것과 같이 허위 사용자들이 삽입됨으로써 사용자 간의 유사도와 아이템 간의 유사도는 변경된다. 즉, 허위 사용자들의 선호도 조작으로 인해 특정 사용자들의 유사 사용자 이웃 집단과 특정 아이템에 대한 유사 아이템 이웃 집단이 변경되게 된다.

만약 현재 시스템이 2.1.1절에 기술한 사용자 기반의 협업적 여과 방법을 이용한다고 가정해 보자. 유사 사용자 집단의 크기 KNN을 2라고 가정하고 실링 공격이 이루어지기 전에 목적사용자 Alice의 "타짜" 선호도를 식 2를 이용하여 예측한다면, 예측된 선호도는 다음과 같다.

$$\hat{R}_{Alice,\text{타짜}} = 3.5 + \frac{(1\times 0.90)+(5\times -0.32)}{|0.90|+|-0.32|} = 2.93$$

이번에는 허위 사용자들 Faker1과 Faker2가 삽입된 후를 고려해보자. 이번에는 Alice의 유사 이웃들이 John과 Faker1이 되기에 "타짜"에 대한 Alice의 선호도 예측은 다음과 같다.

$$\hat{R}_{Alice,\text{타짜}} = 3.5 + \frac{(1\times 0.90)+(5\times 0.78)}{|0.90|+|0.78|} = 6.35$$

즉, "타짜"의 선호도를 최고로 높게 평가한 Alice와 유사한 이웃 Faker1의 영향으로 예측 선호도 값이 크게 상승하게 된다.

만약 현재 시스템이 2.1.2절에 기술한 아이템 기반의 협업적 여과 방법을 이용한다고 가정해 보자. 유사 아이템 집단의 크기를 3이라고 가정하고 실링 공격이 이루어지기 전에 목적사용자 Alice의 "타짜" 선호도와 그 후의 선호도를 식 (4)를 이용하여 예측하면 각각 다음과 같이 계산된다.

공격전: $$\frac{(0\times 0.73)+(1\times 0.50)+(5\times 0.54)}{0.73+0.50+0.54} = 1.80$$

공격전: $$\frac{(3\times 0.70)+(5\times 0.68)+(5\times 0.74)}{0.70+0.68+0.74} = 4.81$$

즉, 공격 아이템에 대한 선호도 예측 값이 1.8에서 4.81로 높아지게 되었다. 이는 허위 사용자들이 Alice가 높게 선호 평가한 아이템들과 공격 아이템 "타짜"를 유사한 아이템이 되도록 유도한

결과라 할 수 있다.

push 공격의 반대 개념인 nuke 공격은 허위 사용자들이 공격 아이템의 추천을 낮추거나 예측 선호도 값을 낮추기 위한 목적으로 프로파일을 조작하여 삽입하는 것을 일컫는다[8]. 특정 아이템의 생산자 및 그와 이익이 관련된 사람들은 자신의 아이템과 경쟁이 되는 아이템이 상대적으로 적게 노출되거나 추천되기를 바란다. 즉, 시스템에서 경쟁사의 아이템의 추천을 낮추어 상대적으로 자신의 아이템에 대한 이익을 유도하는 것이다. 이를 위해, 허위 사용자들은 경쟁되는 아이템, 즉 공격 아이템의 선호도를 최저 평점 R_{min}으로 부여하고 push 공격 때와 유사하게 전체 아이템들 중의 일부에 대하여 조작된 선호도를 생성한다. 이렇게 함으로써, 허위 사용자들과 유사한 사용자들의 공격 아이템에 대한 선호도 예측시 자신들의 최저 평가 점수가 반영되어 낮은 예측 점수를 유도한다. 예를 들어 표 1에서 Faker1과 Faker2의 "타짜"에 대한 선호도 값을 5에서 1로 변경하면 Alice와 John에 대한 nuke 공격이 이루어지게 된다.

시스템 추천 품질 공격은 push 공격과 nuke 공격과는 약간 다른 유형이라 할 수 있다. 이는 특정 아이템에 대한 공격이라기보다는 전체적인 시스템 자체의 추천 또는 선호도 예측에 대한 성능을 방해하려는 목적으로 프로파일을 조작 삽입하는 것을 말한다. 예를 들면, 영화에 대한 아이템을 판매하고 대여해 주는 경쟁 관계의 온라인 추천 시스템 S_1과 S_2가 있다고 하자. 시스템 S_1의 입장에서는 시스템 S_2보다 자신의 시스템으로 사용자들이 많이 방문하기를 바란다. 따라서 허위 사용자들은 경쟁 시스템 S_2에 전반적인 조작된 선호도를 생성하여 일반 사용자들의 추천 신뢰가 떨어지도록 유도한다.

2.2.2 조작된 선호도 모델

허위 사용자들이 자신의 조작된 선호도를 어떻게 생성하느냐에 따라 시스템에 미치는 영향이 다르다. 허위 사용자들의 의도가 최대한 시스템에 영향을 미치게 하려면 조작된 선호도를 추천 시스템에 적절하게 조작하여 삽입하여야 한다.

기본적으로 push 경우에는 그림 1의 *faker user*$_1$과 같이 공격 아이템을 R_{max}의 선호 점수로 조작하고, nuke 경우에는 *faker user*$_2$와 같이 R_{min}의 선호 점수로 조작한다. 그 이외의 아이템에 대해서는 선택되는 아이템들과 그 아이템들의 조작된 선호도 평점에 따라 허위 사용자 모델이 구분된다.

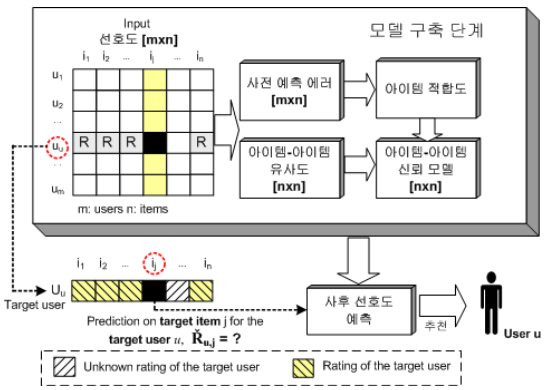
가장 단순한 조작된 선호도 모델은 임의의 공격(random attack)이다. 이는 공격 아이템을 제외한 전체 아이템에 대한 전체 사용자의 선호도 평균 값을 기준으로 선호도 분산 범위 안의 값을 임의로 선택된 아이템들의 조작된 선호도로 이용하는 것이다. 하지만 임의의 선호도 모델은 push 및 nuke 공격에 크게 영향을 주지 못한다[9].

임의의 공격보다 강력한 조작된 선호도 공격 모델은 임의로 선택된 아이템들의 선호도 값을 그 아이템의 평균 선호도와 근사한 점수로 평가하는 것이다. 이를 평균 공격(average attack)이라 일컫는다[7]. 예를 들어, 표 1에서 Faker1과 Faker2는 평균 선호도 모델이라 할 수 있다. 즉, Faker1이 선택한 아이템 "행복"에 대해서는 그 아이템의 평균 선호도인 3점을, 또 다른 아이템 "식객"에 대해서는 "식객"의 평균 선호도 근사 값인 4점을 조작된 선호도로 이용한다. 이러한 선호도 모델은 사용자 기반의 협업적 여과 기법에 대한 공격에 매우 효과적이지만 아이템 기반의 협업적 여과 기법에는 그다지 효과적이지 못하다[7].

임의의 선호도 모델을 변경한 bandwagon 공격은 임의의 선호도 모델에서 선택된 아이템들 중 일부를 가장 인기있는 아이템들로 변경한다. 그리고 그 변경된 아이템들에 대한 선호도를 R_{max}로 부여한다. 일반적으로 인기 있는 아이템들은 많은 사용자들에 의해 높은 선호도를 보이기에 이러한 아이템들을 선택하여 선호도 조작을 하면 많은 일반 사용자들의 유사 사용자가 될 확률이 높다

[8]. bandwagon 선호도 모델과 정 반대로 가장 인기가 낮은 아이템들을 선택하여 그 아이템들의 선호도를 R_{min}으로 평가할 수 있다. 일반적으로 인기가 낮은 아이템은 상대적으로 적은 사용자들에 의해 낮은 선호도를 보이기에 이런 아이템을 선택한 사용자들의 유사 사용자가 될 확률이 높다. Mobasher et al.은 bandwagon 선호도 모델을 이용한 사용자 기반의 협업적 여과에 대한 공격이 평균 선호도 모델처럼 효과적이라는 것을 증명하였다[8].

3. 강건한 협업적 여과 알고리즘



(그림 4) 신뢰 모델을 적용한 협업적 추천 시스템

본 논문에서 제안된 협업적 추천 방법은 2단계로 나눠 질 수 있다. 첫 번째 단계는 오프라인에서 사전에 미리 모델을 구축하는 과정이고 두 번째 단계는 온라인에서 사용자에게 새로운 아이템 추천 및 선호도 예측하는 과정이다. 본 장에서는 협업적 여과 기반의 추천 시스템에서 허위 사용자들(fake users)의 조작된 선호도(manipulated ratings)에 강건한 신뢰 모델 구축 방법에 대하여 기술한다. 그리고 구축된 모델들을 이용하여 온라인 추천 및 예측하는 방법을 기술한다. 그림 4는 제안된 협업적 여과 방법의 전체 흐름을 보여주고 있다.

3.1 사전 선호도 예측 에러

본 논문에서는 사용자들이 아직 선호도를 평가하지 않은 아이템들을 예측하기에 앞서 사용자들이 이미 평가한 아이템에 대한 사전 선호도를 예측한다. 실제로 사용자들이 평가한 선호도 점수에 대하여 제안된 알고리즘에 얼마나 정확한 예측을 했는가에 대한 평가를 통해 새로운 아이템에 대한 예측에 활용한다. 따라서 본 논문은 사용자에 대한 아이템의 선호도 예측을 다음과 같이 두 개로 정의할 수 있다.

- **사전 선호도 예측(A Priori Prediction).** 목적 사용자 u의 $R_{u,j} \neq \varnothing$인 아이템 j에 대한 선호도 예측을 말하며, $P_{u,j}$로 표기한다. 즉, 사용자 u가 이전에 선호도를 이미 평가한 아이템에 대한 예측을 일컫는다.
- **사후 선호도 예측(A Posteriori Prediction).** 목적사용자 u의 $R_{u,j}=\varnothing$인 아이템 j에 대한 선호도 예측을 말하며, $\breve{R}_{u,j}$로 표기한다. 즉, 사용자 u가 아직 선호도를 평가하지 않은 아이템에 대한 예측을 일컫는다.

3.1.1 사전 선호도 예측

사용자들이 이미 선호 평가한 아이템들에 대하여 “All but 1” 프로토콜을 통하여 사전 선호도 예측을 수행한다. “All but 1”은 목적사용자가 이미 선호 평가한 아이템 중 임의로 하나의 아이템을 선택하여 테스트로 사용하고, 그것을 제외한 나머지 데이터를 이용하여 테스트 아이템에 대한 예측을 수행하는 방법이다[1]. 사전 선호도 예측 값은 2.1절에 설명한 것과 같이 사용자 기반의 협업 예측 방법 또는 아이템 기반의 협업 예측 방법을 이용하여 계산할 수 있다. 하지만 본 논문에서는 보다 정확한 사전 예측을 위해 사용자 기반의 접근과 아이템 기반의 접근을 혼합한 방법으로 예측을 수행한다. 즉, 특정 아이템에 대해 나와 비슷한 선호를 가진 사용자들의 선호 경향과 그 아이

템과 유사한 아이템들에 대한 나의 이전 선호 경향을 기반으로 예측하고자 하는 아이템의 예측 선호도 값을 측정한다. 이를 수식으로 정의하면 식 5와 같이 표현할 수 있다.

$$P_{u,j} = R^{j}_{knn(u)} + \frac{\sum_{i \in MSI_u(j)} (R_{u,i} - R^{i}_{knn(u)}) \times sim(i,j)}{\sum_{i \in MSI_u(j)} sim(i,j)} \qquad (5)$$

$P_{u,j}$는 사용자 u의 아이템 j에 대한 사전 선호도 예측 값이고, $MSI_u(j)$는 아이템 j의 유사 아이템들 중에 사용자 u가 선호 평가한 아이템 집합을 의미한다. $sim(i, j)$는 두 아이템 간의 코사인 유사도 값이다. $R^{i}_{knn(u)}$는 사용자 u와 유사한 사용자들의 아이템 i에 대한 평균 선호도 점수이고, $R^{j}_{knn(u)}$는 사용자 u와 유사한 사용자들의 아이템 j에 대한 평균 선호도 점수를 의미한다.

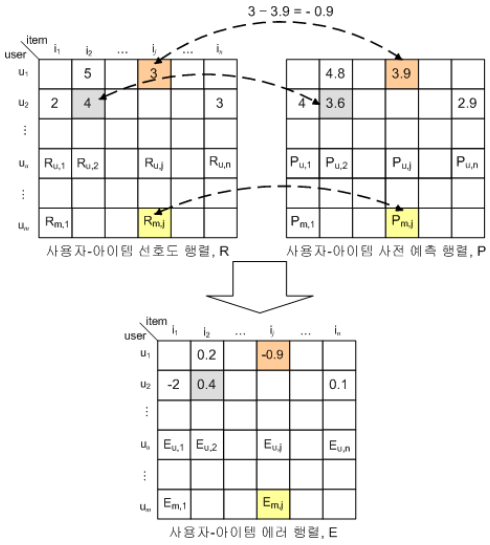
m명의 사용자 집합 $U=\{u_1, u_2, \cdots, u_m\}$의 각 사용자가 n개의 아이템 집합 $I=\{i_1, i_2, \cdots, i_n\}$의 각 아이템에 대한 사전 선호도 예측 값은 $m \times n$ 사용자-아이템 사전 예측 행렬(user-item a priori prediction matrix) **P**로 표현할 수 있다. 여기서 행은 사용자를, 열은 아이템을 의미하며, $P_{u,j}$는 사용자 u의 아이템 j에 대한 사전 예측한 선호도 값을 나타낸다. $P_{u,j}$는 R_{min}에서 R_{max}사이의 실수 값 또는 아직 평가하지 않은 경우 중 하나에 속한다.

$$P_{u,j} \in \{ R_{min}, \ldots, R_{max}, \varnothing \}$$

사용자 u의 아이템 j에 대한 사전 예측한 값이 R_{min}일 경우 그 아이템은 사용자 u의 취향에 어울리지 않음을 예측한 경우이며, R_{max}는 그 반대의 경우라 말할 수 있다. 만약 예측 불가능한 경우가 발생했을 경우, 본 논문에서는 그 사용자의 평균 선호도 값을 $P_{u,j}$로 이용한다.

3.1.2 사전 선호도 예측 에러

사용자들에 대한 사전 선호도 예측이 완료되면 사전 선호도 예측 값이 실제 사용자의 선호도와 얼마나 유사하게 예측되었는가를 실제 선호도 값과 사전 예측된 값의 오차로 계산할 수 있다.



(그림 5) 사전 선호도 예측 에러 계산 과정

실제 선호도 값에 대응하는 사전 예측 선호도 값의 쌍 $<R_{u,j}, P_{u,j}>$에 대하여 각각의 예측 에러는 그림 5와 같이 실제 선호도 값에서 예측된 선호도의 차로 계산한다. 즉, 이를 수식화 하면 다음 식 6과 같다.

$$E_{u,j} = R_{u,j} - P_{u,j} \qquad (6)$$

$R_{u,j}$는 사용자 u의 아이템 j에 대한 실제 선호도 값을 의미하며, $P_{u,j}$는 아이템 j에 대한 사용자 u의 사전 선호도 예측 값을 의미한다. 형식적으로, 사용자-아이템 선호도 행렬과 사용자-아이템 사전 예측 행렬로부터 사용자들에 대한 아이템의 예측 에러는 사용자-아이템 에러 행렬(user-item error matrix) E로 표현될 수 있다. 행렬의 요소 $E_{u,j}$가 가질 수 있는 값은 $(R_{min} - R_{max})$에서 $(R_{max} - R_{min})$사이의 실수 값 또는 아직 평가하지 않은 경우 중에 하나에 속한다.

$$E_{u,j} \in \{ (R_{min} - R_{max}), \ldots, 0, \ldots, (R_{max} - R_{min}), \varnothing \}$$

사전 예측 값이 과대평가(overestimatation) 되었다면, 예측 에러 값 $E_{u,j}$는 음의 값을 가지며, 과소

평가(underestimatation) 되었다면 $E_{u,j}$는 양의 값을 가지게 된다. $E_{u,j}$=0일 경우는 시스템이 실제 선호도 값을 정확하게 예측했음을 의미하며, 0에 가까울수록 사전 선호도 예측의 정확도가 높다고 말할 수 있다.

3.2 아이템 사전 예측 적합도

사전 선호도 예측 에러란 특정 아이템에 대해 나와 비슷한 선호를 가진 사용자들의 선호 경향과 특정 아이템과 유사한 아이템들에 대한 나의 선호 경향을 기반으로 예측 했을 때 발생한 에러이다. 따라서 실제 선호도와의 예측 에러가 작다는 것은 그 아이템에 대하여 적절한 유사 이웃과 적절한 유사 아이템을 기반으로 예측했음을 의미한다[10]. 아이템 적합도(Item Suitability: IS)란 특정 아이템에 대한 시스템이 측정한 사전 선호 예측들이 얼마나 적절히 계산하였는지의 비율을 의미하는 것으로 다음 식 7에 의해 정의된다.

$$IS_{\theta}(j) = \frac{|X_j \cap X_j^{\theta}|}{|X_j|} \quad (7)$$

$IS_{\theta}(j)$는 에러 임계값 Θ에 대한 아이템 j의 적합도를 의미한다. X_j는 에러 행렬 E에서 아이템 j에 대하여 사전 선호도 예측이 수행된 사용자들의 집합이고 X_j^{θ}는 X_j에 속하는 사용자들 중에 사전 예측된 에러의 절대값이 에러 임계값 Θ보다 작은 사용자들의 집합을 의미한다.

$$X_j = \{u \mid u \in U \wedge E_{u,j} \neq \varnothing\}$$

$$X_j^{\theta} = \{u \mid u \in X_j \wedge |E_{u,j}| < \Theta\}$$

에러 임계값 Θ는 0보다 큰 실수이다. 사전 예측 에러 값에 절대값을 고려하는 이유는 앞에서 언급했듯이 예측 에러는 과대 평가시 음의 값을 가지며, 과소 평가시 양의 값을 가지기 때문이다. 따라서 어느 정도의 에러 임계값을 설정해야 하는 것이 무엇보다 중요한 문제이다.

아이템 적합도는 0에서 1 사이의 값을 가진다. $IS_{\theta}(j)$가 0의 값을 가질 때에는 아이템 j에 대한 모든 사전 선호도 예측 에러가 크다는 것을 의미하며, $IS_{\theta}(j)$가 1의 값을 가질 때에는 아이템 j에 대한 모든 사전 선호도가 만족할 만한 에러 수준으로 예측되었음을 의미한다.

3.3 아이템 기반 신뢰 모델 구축

본 절에서는 아이템 사전 예측 적합도를 이용하여 아이템 간의 신뢰도를 측정하고 아이템 기반의 신뢰 모델 구축 절차를 기술한다.

3.3.1 아이템 신뢰도

아이템 기반의 협업 여과가 일반적인 조작된 선호도에 강건하다고 하지만 이 역시 특성화된 공격 모델에 취약할 수 있다. 만약 일반 사용자들이 대체로 높게 선호 평가한 아이템의 집단과 대체로 낮게 선호 평가한 아이템 집단의 정보를 얻어낸다면 아이템 기반의 협업 여과도 쉽게 공격받을 수 있다. 허위 사용자들은 높게 선호 평가된 아이템들을 가능한 한 공격 아이템과 유사하게 만들어야 하고, 낮게 선호 평가된 아이템들은 가능한 한 공격 아이템과 유사하지 않게 만들면 된다. 더군다나 만약 공격 아이템이 시스템에 새로이 삽입된 아이템이라면 일반 사용자들이 선호 평가한 정보가 아직 충분히 쌓이지 않았기에 시스템은 더 큰 영향을 받게 된다.

이를 보완하기 위해 아이템 간의 유사도와 특정 아이템의 사전 예측 적합도 간의 가중치된 조합 평균(weighted harmonic mean)으로 두 아이템 간의 신뢰도를 측정한다. 즉, van Rijsbergen의 F-measure[15]를 활용한 두 아이템 간의 신뢰도 측정 방법을 수식화 하면 식 8과 같다.

$$T_{j,i}^{(\beta)} = \frac{(\beta^2 + 1) \times sim(j,i) \times IS_{\theta}(i)}{\beta^2 \times sim(j,i) + IS_{\theta}(i)} \quad (8)$$

$T_{j,i}^{(\beta)}$는 아이템 j에 대한 아이템 i의 신뢰도이다. $sim(j,i)$는 아이템 j와 i의 유사도이며 $IS_{\theta}(i)$는 아이템 i의 적합도이다. 여기서 β는 아이템 유사도와 아이템 적합도의 상대적 비중을 나타내는 변수이다. 만약 β의 값이 0이라면 $T_{j,i}^{(\beta)}$는 아이템 간의 유사도 $sim(j,i)$와 같으며, β의 값이 $+\infty$라면 $T_{j,i}^{(\beta)}$는 $IS_{\theta}(i)$와 같게 된다. 아이템 유사도와 아이템 적합도의 조합 평균(harmonic mean)으로 아이템 신뢰도 값을 측정하려면 β의 값을 1로 설정하면 된다. 아이템간의 신뢰도 역시 0에서 1 범위의 값으로 계산되며, $T_{j,i}^{(\beta)}$와 $T_{i,j}^{(\beta)}$은 같지 않다.

아이템 j에 대한 아이템 i의 신뢰도가 높다는 것은 두 아이템 j와 i의 유사도뿐만 아니라 아이템 i의 적합도 역시 높다는 것이다. 이는 허위 사용자들이 자신들의 목적을 달성하기 위해 일부 아이템들을 공격 아이템과 유사하게 만들었다 할지라도 실제 일반 사용자의 공격 아이템에 대한 사전 선호도 예측에는 영향을 주기 힘들다. 즉, 허위 사용자들의 목적은 공격 아이템에 대한 일반 사용자들의 사후 예측에 영향을 미치게 하려는 것이지 이미 공격 아이템을 선호 평가한 사용자들의 사전 선호도 예측에 영향을 주려는 것은 아니다.

알고리즘 1은 특정 아이템과 다른 아이템과의 신뢰도를 측정하는 방법이다.

(알고리즘 1) 아이템 적합도 및 아이템 간의 신뢰도 측정

```
    Input: parameter β; error threshold θ; item-item similarity
    matrix D; user-item error matrix E; total item list I
    Output: item-item trust matrix T^(β)

    computeItemtoItemTrust(β, θ, E, D, I)
    IS_θ[] ← itemSuitability(I, E, θ)
01  for each j ∈ I
02      t1← 0; t2 ← 0;
03      for i ← 1 to n
04          t1 ← (β^2+1) × D_j,i × IS_θ[i]
05          t2 ← (β^2 × D_j,i) + IS_θ[i]
06          T^(β)_j,i ← t1 / t2
07  return T^(β)
08
    itemSuitability (I, E, θ)
    for each i ∈ I
01      irrelevant ← 0
02      relevant ← 0
03      for u ← 1 to m
04          if E_u,i =0 then continue;
05          else if |E_u,i| ≤ θ then relevant++
06          else irrelevant++
07      IS_θ[i] = relevant / (relevant + irrelevant)
08  return IS_θ[]
09
```

선호도 행렬 **R**과 예측 에러 행렬 **E**로부터 계산된 아이템 간의 신뢰도는 n×n 아이템-아이템 비대칭 신뢰 행렬(item-item trust asymmetric matrix) $\mathbf{T}^{(\beta)}$로 표현할 수 있다. 여기서 행과 열 모두 아이템을 의미하며, 행렬 요소 $T_{j,i}^{(\beta)}$는 아이템 j에 대한 아이템 i의 신뢰도 값을 나타낸다.

3.3.2 아이템 기반 신뢰 모델

특정 아이템과 모든 다른 아이템과의 신뢰도가 측정되면 그 다음으로 이웃 집단의 규모를 결정한다. 여기서 이웃 집단의 의미는 특정 아이템과 가장 신뢰도가 높은 아이템들이다. 이를 본 논문에서는 신뢰 아이템 집단(Most Trusted Items: MTI)이라 지칭한다. 신뢰도가 구해진 모든 아이템들, 즉 신뢰 행렬 $\mathbf{T}^{(\beta)}$ 자체를 사후 예측에 이용할 수 있으나, 이는 정확도나 실링 공격에 대한 강건성(robustness) 면에서도 권장할 방법이 아니다. 최대한 허위 사용자들의 조작된 선호도의 영향을 둔화시키기 위한 적절한 신뢰 아이템 집단의 크기를 선택해야 한다.

알고리즘 2는 아이템 기반 신뢰 모델 구축 과정을 나타낸 것이다. 알고리즘 입력으로 앞 절에서 구축한 신뢰 행렬 $\mathbf{T}^{(\beta)}$와 모델 크기를 결정하기 위한 변수 K이다. 그리고 출력 결과로 n×n 아이템-아이템 행렬로 표현될 수 있는 신뢰 모델 $\hat{\mathbf{T}}^{(K)}$이다. 신뢰 모델 $\hat{\mathbf{T}}^{(K)}$의 각 요소는 0이거나 두 아이템 간의 신뢰도 값이 되고 j번째 행은 아이템 j와 가장 신뢰도가 높은 K개의 아이템을 저장한다. 즉, 각 1×n 행 벡터는 최대 K개의 0이 아닌 값을 가지며, 0이 아닌 값을 가지는 열은 그 행의 아이

템의 신뢰 아이템 집단에 속함을 의미한다.

(알고리즘 2) 아이템 기반 신뢰 모델 구축

```
    Input: model size K; item-item trust matrix T^(β); total
    item list I
    Output: item-based trust model T̆^(k)

    buildTrustModel (K, T^(β), I)
01  for each j ∈ I
02      for i ← 1 to n
03          T̆^(K)_j,i ← T^(β)_j,i
04      for i ← 1 to n
05          if i = j
06              then T̆^(K)_j,i ← 0
07          else if T̆^(K)_j,i ≒ among the K highest values in
            T̆^(k)_j,*
08              then T̆^(k)_j,i ← 0
09  return T̆^(k)
```

3.4 신뢰 모델을 이용한 협업적 추천 시스템

본 절에서는 구축된 신뢰 모델을 이용하여 사용자의 선호도를 예측하고 아이템을 추천 하는 방법에 대해서 기술한다.

3.4.1 신뢰 모델 이용한 사후 선호도 예측

협업적 여과에서 마지막 단계는 사용자의 아이템 선호도를 예측하는 과정이다. 예측 대상이 되는 아이템은 특정 사용자가 아직 선호 평가 하지 않은 아이템을 말한다. 시스템에 있는 사용자들의 과거 정보를 바탕으로 특정 사용자가 특정 아이템을 얼마나 선호하는지를 정확하게 예측하는 것은 협업적 여과에서 무엇보다 중요한 문제이다. 이는 사용자들이 아이템을 구매, 선택, 또는 관람하는 의사 결정(decision making)에 직접적으로 영향을 주기 때문이다.

많은 선행 연구에서 언급했듯이, 아이템 기반의 접근 방법은 기본적으로 나의 선호도 경향을 기반으로 하기에 상대적으로 사용자 기반의 접근 방법에 비하여 실링 공격에 강건하다[7][8]. 즉, 일반 사용자들의 선호도는 허위 사용자들이 조작할 수 있는 것이 아니기에 허위 사용자들이 그들의 목적을 달성하기 위해서는 많은 지식과 많은 노력이 필요하다. 따라서 본 논문에서는 목적사용자가 특정 아이템을 얼마나 선호할 것인가를 추측하기 위해 예측 하고자 하는 아이템과 신뢰도 값이 높은 아이템들에 대한 목적사용자의 과거의 선호 경향을 활용한다.

$$\breve{R}_{u,j} = \frac{\sum_{i \in MTI_u(j)} R_{u,i} \times \breve{T}_{j,i}}{\sum_{i \in MTI_u(j)} \breve{T}_{j,i}} \quad (9)$$

식 9는 특정 아이템 j와 가장 신뢰도가 높은 아이템들의 각 신뢰도를 가중치로 목적사용자 u의 과거 선호도들의 합으로 예측하는 방법이다. 이는 허위 사용자의 조작된 선호도가 직접적으론 예측에 반영되지 못해 실링 공격에 다소 강건한 예측 방법이라 할 수 있다.

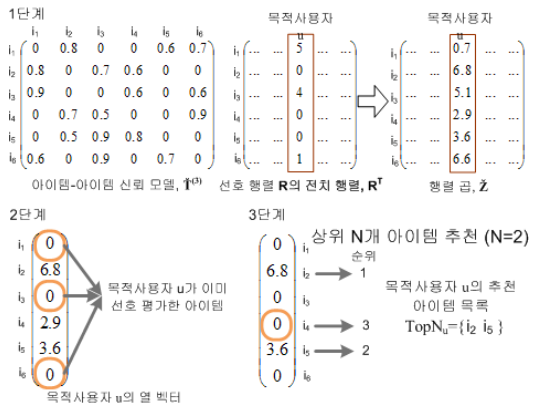
3.4.2 신뢰 모델 이용한 아이템 추천

협업적 여과 추천 시스템은 일종의 개인화된 정보 여과 기술로써 최종적으로 특정 아이템에 대한 선호도를 예측 하거나 목적사용자의 N개의 추천 아이템 리스트를 제공한다. 앞 절에서는 특정 아이템을 목적사용자가 얼마나 선호하는지를 예측하는 과정을 기술하였다. 본 절은 신뢰 모델을 이용하여 개인마다 특성화 된 N개의 아이템 리스트를 식별하고, 그것들을 개개인의 사용자에게 추천하는 과정을 기술한다. 이는 일반적으로 상위 N개 아이템 추천(Top-N Recommendation)이라 일컫는다[3]. 알고리즘 3은 신뢰 모델을 적용하여 상위 N개의 아이템을 추천하는 과정이다.

(알고리즘 3) 상위 N개 아이템 추천

Input: target user *u*, user-item rating matrix **R**; item-based trust model $\check{\mathbf{T}}^{(K)}$; a number of recommended items *N*;
Output: a set of items for user *u* $\mathbf{TopN_u}$

generatingTopNItem(*u*, *N*, $\check{T}^{(k)}$)
01 let $\mathbf{R}^T$ be a transpose of a matrix **R**
02 **for** *j* ← 1 to n
03 $\check{Z}_{j,u} \leftarrow 0$
04 **for** y ← 1 to n
05 $\check{Z}_{j,u} \leftarrow \check{Z}_{j,u} + (\check{T}^{(K)}_{j,y} \times R^T_{y,u})$
06 **for** *j* ← 1 to n
07 **if** $R^T_{j,u} \neq \varnothing$
08 **then** $\check{Z}_{j,u} \leftarrow 0$
09 **else if** $\check{Z}_{j,u} \neq$ among the *N* highest values in $\check{Z}_{*,u}$
10 **then** $\check{Z}_{j,u} \leftarrow 0$
11 **for** *j* ← 1 to n
12 **if** $\check{Z}_{j,u} \neq 0$ then
13 $\mathrm{TopN_u} \leftarrow \mathrm{TopN_u} \cup \{\mathrm{item}j\}$
14 **return** $\mathrm{TopN_u}$



(그림 6) 상위 N개 아이템 추천 과정 (N=2일 경우)

알고리즘은 크게 그림 6과 같이 세 단계로 이루어진다. 첫 번째 단계에서는 우선 m×n 사용자-아이템 선호도 행렬 **R**에서 $R_{u,j}=\varnothing$인 값들을 모두 0으로 설정한다. 그 다음에 n×n 아이템-아이템 신뢰 모델 $\check{\mathbf{T}}$와 **R**의 전치 행렬인 n×m 아이템-사용자 행렬 $\mathbf{R}^T$와 행렬 곱 연산을 한다.

$$\check{\mathbf{Z}} = \check{\mathbf{T}}\ \mathbf{R}^T$$

행렬 곱으로 만들어진 n×m 행렬 $\check{\mathbf{Z}}$에서 행은 아이템을 열은 사용자를 나타낸다. 따라서 *j*번째 행의 *u*번째 열의 값은 *j*번째 아이템과 가장 신뢰도가 높은 아이템들의 각 신뢰도를 가중치로 한 *u*번째 사용자의 과거 선호도들의 합이 된다. 두 번째 단계에서는 행렬 $\mathbf{R}^T$에서 0이 아닌 요소에 해당하는 행렬 $\check{\mathbf{Z}}$의 값을 0으로 변경한다. 즉, 사용자가 이미 선호 평가한 아이템은 추천 대상이 아니므로 그 아이템들에 대해 0으로 변경하는 단계이다. 마지막 단계에는 행렬 $\check{\mathbf{Z}}$의 각 열(사용자)을 기준으로 가장 높은 수치의 값을 가지는 *N*개의 요소를 제외한 나머지를 0으로 변경한다. 즉, 신뢰도를 가중치로 한 사용자의 선호도 합이 가장 높은 *N*개의 아이템을 식별하는 단계이다. 모든 단계가 수행되면 행렬 $\check{\mathbf{Z}}$의 각 n×1 열 벡터에서 0이 아닌 아이템들이 최종 그 열에 해당하는 사용자의 *N*개의 추천 아이템 리스트이다. 여기서 각 사용자에게 추천되는 아이템의 개수가 실제로 *N*개 보다 적을 수 있다. 이런 경우는 사용자가 아직 선호 평가하지 않은 아이템의 수가 *N*개 보다 적은 경우이거나, 신뢰도를 가중치로 한 사용자의 선호도 합이 0의 값보다 큰 경우가 *N*개 보다 적을 경우이다.

4. 실험 분석 및 결과

본 장에서는 제안된 협업적 아이템 신뢰 모델을 적용한 추천 시스템의 성능에 대한 실험을 하고 그 결과에 대한 분석을 기술한다. 제안된 모델들의 성능 평가는 크게 두 범위로 나뉘어 진행되었다. 우선, 제안된 모델 기반의 협업적 여과 방법이 얼마나 정확한 예측 및 추천을 제공하는가에 대한 품질을 평가하였다. 그 다음, 허위 사용자들에 대한 공격에 얼마나 영향을 받는가에 대한 시스템 강건성에 대한 평가를 하였다.

4.1 실험 평가 방법

4.1.1 실험 데이터 집합

제안된 모델들의 성능 평가 실험은 실제 데이터인 MovieLens* 데이터 집합(dataset)을 사용하였다. MovieLens는 GroupLens[11]에서 연구를 목적으로 운영하는 웹 기반의 추천 시스템으로, 사용자들이 자신이 본 영화에 대하여 1(R_{min})부터 5(R_{max})사이의 선호도를 평가할 수 있다. MovieLens 데이터는 각 사용자마다 최소 20개 이상의 아이템에 대한 선호도 정보를 가지고 있으며, 많은 협업적 여과 관련 연구에서 실험 데이터로 활용되고 있다. 데이터에는 943명의 사용자들이 1,682개의 아이템에 대한 100,000개의 선호도 점수를 포함하고 있다.

실험은 각 데이터 집합을 80% 학습 데이터 집합(training set)과 20% 테스트 데이터 집합(test set)으로 나누어, 학습 데이터만을 이용하여 테스트 집합에 있는 아이템에 대한 사용자의 선호도를 예측하였다.

4.1.2 성능 측정식

제안된 모델을 적용한 추천 시스템의 성능을 평가하기 위해 다음과 같은 측정식들을 이용하였다.

가. 예측 정확성 평가 방법

협업적 여과 추천 시스템에서 가장 보편화된 예측의 정확성 평가 방법은 MAE(Mean Absolute Error)이다[1][4]. MAE는 테스트 집합에 있는 실제 선호도와 시스템에 의해 예측된 선호도의 평균 절대 편차를 평가하는 통계적인 측정 방법이다. MAUE(u)는 테스트 집합에 있는 사용자 u에 대한 평균 절대 사용자 오차(Mean Absolute User Error)로 식 10에 의해 정의된다.

$$MAUE(u) = \frac{\sum_{j \in IT_u} | R_{u,j} - \check{R}_{u,j} |}{| IT_u |} \quad (10)$$

IT_u는 테스트 데이터에서 사용자 u가 선호 평가한 아이템 집합이며 <$R_{u,j}$, $\check{R}_{u,j}$>는 사용자 u의 아이템 j에 대한 실제 선호도 값과 예측된 선호도 값의 쌍이다. 따라서 전체 테스트 집합에 있는 사용자들의 MAE는 식 11을 이용하여 계산할 수 있다.

$$MAE = \frac{\sum_{u \in UT} MAUE(u)}{| UT |} \quad (11)$$

UT는 테스트 데이터에 있는 사용자 집합을 의미한다. MAE가 최소화될수록 시스템의 예측 정확도가 높다고 말할 수 있다.

나. 시스템 강건성 평가 방법

실링 공격에 대한 강건성에 대한 성능 평가로는 예측 변화(prediction shift)의 측정식을 이용한다[9]. 테스트 데이터의 예측 변동은 조작된 선호도가 포함되기 전의 선호도 예측 값과 조작된 선호도가 포함되었을 때 예측된 값의 차로 다음과 같이 계산될 수 있다.

$$\Delta_{u,j} = \check{R}_{u,j} - \check{R}'_{u,j}$$

$\check{R}_{u,j}$는 원래의 학습 데이터 집합을 이용하였을 때 사용자 u의 아이템 j에 대한 선호도 예측 값이고 $\check{R}'_{u,j}$는 허위 사용자들의 프로파일이 삽입되었을 때의 예측 값이다. $\Delta_{u,j}$가 양의 값이면 nuke 공격이, 음의 값이면 push 공격이 이루어 졌음을 의미한다[8]. 본 실험 평가에서는 nuke 또는 push 공격에 상관없이 허위 사용자들의 조작된 선호도에 시스템이 얼마나 영향을 받는가에 대한 시스템 품질 공격에 대하여 평가한다. 이는 테스트 데이터의 아이템 j에 대한 전체 사용자에 대한 절대 예측 변동(absolute prediction shift)으로 측정할 수 있으며 식 12와 같이 정의된다.

* http://www.movielens.org

$$APS(j) = \frac{\sum_{u \in UT_j} |\Delta_{u,j}|}{|UT_j|} \quad (12)$$

UT_j는 테스트 데이터에서 아이템 j를 선호 평가한 사용자들의 집합을 의미한다. 최종적으로 전체 테스트 데이터의 실링 공격에 대한 영향은 식 13을 이용하여 측정하였다.

$$MAPS = \frac{\sum_{j \in IT} APS(j)}{|IT|} \quad (13)$$

IT는 테스트 데이터에 있는 아이템들의 집합이다. MAPS(Mean Absolute Prediction Shift) 값이 0에 가까울수록 실링 공격의 영향을 적게 받는다고 할 수 있다.

4.1.3 비교 평가 시스템

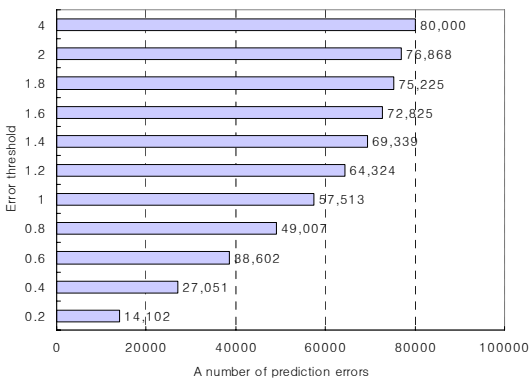
제안된 모델들의 예측 정확성과 실링 공격의 강건성을 평가하기 위해, 2.1절에서 기술한 협업적 여과 추천 시스템에서 가장 널리 알려진 두 가지 방법, 피어슨 상관 계수 유사도를 이용한 사용자 기반의 협업적 여과 방법[1][4]과 코사인 유사도를 이용한 아이템 기반의 협업적 여과 방법[3][12]의 성능과 비교하였다. 실험 결과에서 기존의 알고리즘인 사용자 기반의 협업적 여과는 UserCF, 아이템 기반의 협업적 여과는 ItemCF, 제안된 협업적 여과는 ITrustCF로 각각 표기하였다.

4.2 파라미터 값의 변화에 따른 예측 성능

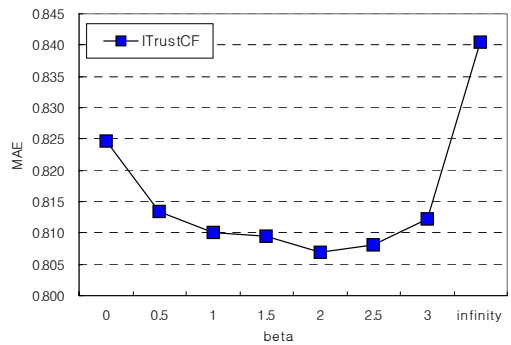
이번 절에서는 제안된 신뢰 모델 구축시 영향을 미치는 파라미터들의 값의 변화에 따른 성능 평가 실험 및 분석에 대하여 기술한다.

4.2.1 β 값의 변화에 따른 성능 평가

신뢰도 모델 구축시 β 값에 따라 두 아이템 간의 신뢰도 값이 다르게 된다. 즉, 아이템 적합도에 더 높은 비중을 둘 것인가 아니면 아이템 유사도에 더 비중을 둘 것인가에 따라 아이템 선호도 예측 값이 달라질 수 있다. 따라서 본 절의 실험에서는 β 값을 0.5에서 3.0으로 변화시키면서 예측 성능의 변화를 분석해 보았다. 또한, 아이템 유사도만을 고려했을 때(β=0)와 아이템 적합도만을 고려했을 때(β=+∞)의 예측 성능 차이를 비교하였다. 비록 아이템 사전 예측 적합도가 에러 임계값 Θ에 따라 다소 차이가 발생하지만, 그림 7과 같이 에러 임계값에 따른 사전 예측 에러의 수를 바탕으로 Θ를 1.4로 설정하였다. 그리고 신뢰 모델의 크기는 60으로 설정하였다.



(그림 7) Θ값에 따른 사전 예측 에러 수



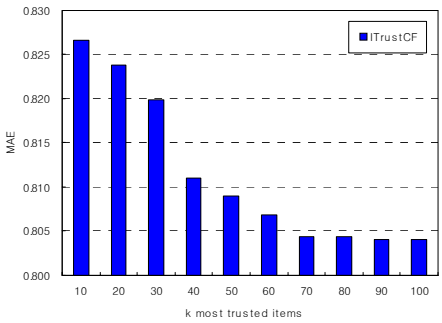
(그림 8) β 값의 변화에 따른 예측 MAE

그림 8은 β값의 변화에 따른 ITrustCF의 MAE 변화를 보여준다. 실험결과 β 값이 증가 할수록

예측 성능이 향상되었으며, β값이 2를 기점으로 서서히 예측 성능이 저하되었다. β값이 1을 기점으로 변화를 분석해 보면, 아이템 적합도와 아이템 유사도를 동일한 비중으로 고려했을 때(β=1)보다 아이템 적합도에 조금 더 가중을 두었을 때가 더 높은 예측 성능을 보였다. 실험 결과 흥미로운 것은 아이템 유사도만을 고려했을 때(β=0)와 아이템 적합도만을 고려했을 때(β=+∞)의 예측 성능 결과이다. 아이템 적합도가 아이템 유사도와 같이 고려되었을 때는 더 좋은 성능을 보였지만, 아이템 적합도만을 고려하였을 때에는 예측 성능이 현저하게 떨어졌다. 이번 실험 결과로부터 사전에 적절하게 예측된 아이템 정보가 그 아이템에 대한 사후 선호도 예측의 정확성 향상에 긍정적인 영향을 미침을 알 수 있다.

4.2.2 신뢰 모델 크기에 따른 성능 평가

본 절의 실험은 신뢰 모델의 크기가 예측 정확성에 미치는 영향을 분석한다. 앞 절의 실험 결과를 토대로 β 값을 2로 사용한 신뢰 행렬 $T^{(2.0)}$로부터 신뢰 아이템 집단의 크기를 10에서 100까지 10개씩 증가시키면서 MAE를 측정하였다. 그림 9는 실험 결과를 나타낸다.



(그림 9) 모델 크기 변화에 따른 MAE 변화

결과 그래프에서 보이듯이, 신뢰 아이템 집단의 크기가 너무 작으면 많은 아이템 정보가 모델에서 제거되어 예측 성능이 좋지 않았으며, 신뢰 아이템 집단의 크기가 70인 $\check{T}^{(70)}$ 이후부터 예측 성능 변화가 거의 없었다.

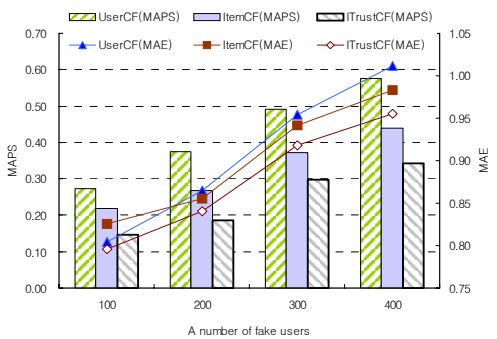
4.3 조작된 선호도에 대한 강건성 비교 평가

이번 절에서는 허위 사용자의 악의적인 공격에 대한 제안된 신뢰 모델의 예측 정확성 및 강건성 성능 평가에 대하여 기술한다. 실험에 사용된 조작된 선호도 모델은 임의의 공격(random attack)을 이용하였다. 따라서, 전체 아이템 중에 임의의 50개의 아이템을 선택하여, 전체 사용자의 선호도 평균 값을 기준으로 선호도 분산 범위 안의 값을 각 허위 사용자의 선택된 50개의 아이템에 대한 선호도 값으로 이용하였다. 그리고 조작된 선호도를 가진 사용자들을 MovieLens 원본 학습 데이터 집합에 100, 200, 300, 400명을 각각 삽입함으로써 시스템 품질 공격에 미치는 영향을 분석하였다.

표 2는 원본 학습 데이터 집합에 조작된 허위 사용자들을 삽입시키기 전의 UserCF, ItemCF, 그리고 ITrustCF에 대한 가장 좋은 예측 MAE 결과를 나타낸다. 그리고 그림 10은 허위 사용자 수의 변화에 따른 UserCF, ItemCF, ITrustCF의 MAPS와 MAE를 보여주고 있다.

(표 2) 조작된 선호도를 삽입하기 전 각 방법의 선호도 예측 MAE

	UserCF	ItemCF	ITrustCF
MAE	0.7534	0.8230	0.8044



(그림 10) 허위 사용자 공격 수에 따른 MAPS 와 MAE

허위 사용자들의 수가 증가함에 따라 그림 10의 MAPS(막대 그래프)의 변화를 살펴보면, 세 가지 방법 모두 조작된 선호도가 영향을 미쳤음을 알 수 있다. 즉, UserCF의 경우는 허위 사용자들이 일반 사용자의 유사한 이웃으로 선택되는 것이 성공했고, ItemCF는 조작된 선호도에 의해 아이템 간의 유사도가 변경되었음을 의미한다. 그리고 ITrustCF 역시, 아이템 간의 유사도 또는 아이템 적합도가 변경되었음을 의미한다.

실링 공격에 강건한 여과 방법은 정확한 예측을 제공하면서(MAE) 허위 사용자들이 미치는 영향을 최소화해야 한다(MAPS). 이는 반대로, 허위 사용자들로 하여금 그들의 목적 달성을 위해 엄청난 비용을 요구하게 만든다면 실링 공격에 강건하다 말할 수 있다. 이런 측면을 고려했을 때, 200명의 허위 사용자를 삽입되었을 경우 UserCF의 예측 변화 정도가 ITrustCF 경우는 400명의 허위 사용자를 삽입해야지 비슷한 효과를 볼 수 있었다. ItemCF와 ITrustCF의 예측 변화를 비교하였을 때에도, ITrustCF의 예측 변화가 더 낮았으며, 허위 사용자들의 수가 증가할수록 더 강건함을 알 수 있었다. 결론적으로, ITrustCF 방법이 UserCF 방법에 비해 평균 19%, ItemCF 방법에 비해 평균 8% 낮은 예측 변화를 보였다.

5. 결론 및 향후 연구

본 논문에서는 협업적 여과에서 발생할 수 있는 대표적인 문제점인 실링공격을 보완하기 위해 특정 아이템에 대한 사전 선호 예측들이 얼마나 적절히 계산되었는지의 비율을 의미하는 아이템 적합도와 두 아이템 간의 유사도를 이용한 신뢰 모델을 구축하는 방법들을 제시하였으며, 또한 그 모델들을 적용하여 개인화된 아이템을 추천하는 협업적 여과 방법을 제안하였다.

실링 공격에 대한 시스템의 강건성을 평가하기 위해, MovieLens 학습 데이터에 인위로 조작된 허위 사용자 프로파일을 삽입하면서 기존의 방법들과 제안된 방법의 예측 변화를 살펴보았다. 비교 결과, 제안된 방법이 실링 공격에 대해 가장 강건함을 보였다. 따라서 본 연구 결과를 통해 사전에 예측된 아이템들의 에러 정보를 활용한 신뢰 모델은 허위로 조작된 선호도가 시스템에 미치는 영향을 최소화 시킬 수 있고, 이는 실링 공격에 강건하고 신뢰할 수 있는 추천 시스템 구축에 활용될 수 있음을 알 수 있다.

하지만 다량의 조작된 선호도가 삽입된다면 제안된 방법 역시 예측 및 추천의 결과가 변경될 수 밖에 없다. 따라서 최근에 활발하게 진행되고 있는 허위 사용자의 발견(detection) 기법에 관한 향후 연구가 필요하다[2][14]. 이는 제안된 모델을 구축시 허위 사용자들의 모델, 조작된 선호도의 패턴 등의 분석을 통하여 허위 사용자로 의심하는 정보를 제거하거나 낮은 가중치를 줌으로써 시스템의 안정성 및 강건성을 향상 시킬 수 있을 것으로 기대된다.

참 고 문 헌

[1] Breese, J. S., Heckeman, D., and Kadie, C., "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence, pp.43-52, 1998.

[2] Burke, R., Mobasher, B., Williams, C., and Bhaumik, R., "Classification Features for Attack Detection in Collaborative Recommender Systems," Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and DataMining, pp.542-547, 2006.

[3] Deshpande, M. and Karypis, G., "Item-based Top-N Recommendation Algorithms," ACM Transactions on Information Systems, Vol. 22, pp.143-177, 2004.

[4] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T., "Evaluating Collaborative Filtering

Recommender Systems," ACM Transactions on Information Systems, Vol.22, ACM Press, pp.5-53, 2004.

[5] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., and Riedl, J., "GroupLens: Applying Collaborative Filtering to Usenet News," Communications of the ACM, Vol.40, pp.77-87, 1997.

[6] Lam, S. and Ridl, J., "Shilling Recommender Systems for Fun and Profit," Proceedings of the 13th International World Wide Web Conference, ACM press, pp.393-402, 2004.

[7] Mobasher, B., Burke, R., Bhaumik, R., and Samdvig, J. J., "Attacks and Remedies in Collaborative Recommendation," IEEE Intelligent Systems, Vol.22, pp.56-63, 2007.

[8] Mobasher, B., Burke, R., Bhaumik, R., and Williams, C., "Toward Trustworthy Recommender Systems: An Analysis of Attack Models and Algorithm Robustness," ACM Transactions on Internet Technology, Vol.7, 2007.

[9] O'Mahony, M., Hurley, N., Kushmerick, N., and Silverstre, G., "Collaborative Recommendation: A Robustness Analysis," ACM Transactions on Internet Technology, Vol.4, pp.344-377, 2004.

[10] O'Donvan, J. and Smyth, B., "Mining Trust Values from Recommendation Errors," International Journal on Artificial Intelligence Tools, pp.945-962, 2006.

[11] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J., "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," Proceedings of the ACM Conference on Computer Supported Cooperative Work, ACM press, pp.175-186, 1994.

[12] Sarwar, B., Karypis, G., Konstan, J., and Reidl, J., "Item-based Collaborative Filtering Recommendation Algorithms," Proceedings of the 10th International World Wide Web Conference, ACM press, pp.285-295, 2001.

[13] Shardanand, U. and Maes, P., "Social Information Filtering: Algorithms for Automating Word of Mouth," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp.210-217, 1995.

[14] Williams, C., Mobasher, B., Burke, R., "Defending Recommender Systems: Detection of Profile Injection Attacks," Journal of Service Oriented Computing and Applications, Vol.1, Springer, pp.157-170, 2007.

[15] van Rijsbergen, C. J., "Information Retrieval," 2nd Ed. Butterworth, 1979.

◐ 저 자 소 개 ◑

김 흥 남

2002년 인하대학교 컴퓨터공학과 졸업(학사)
2004년 인하대학교 대학원 전자계산공학과 졸업(석사)
2009년 인하대학교 대학원 정보학과 졸업(박사)
2009～현재 인하대학 BK21 박사후 연구원
관심분야 : 추천시스템, 데이터 마이닝, 시맨틱 웹, etc.
E-mail : nami@eslab.inha.ac.kr

하 인 애

2005년 수원대학교 컴퓨터과학과 졸업(학사)
2007년 인하대학교 대학원 컴퓨터정보공학과 졸업(석사)
2007～현재 인하대학교 대학원 컴퓨터정보공학과 박사과정
관심분야 : 추천시스템, 시맨틱 웹, 개인화 시스템, 이러닝 etc.
E-mail : inay@eslab.inha.ac.kr

조 근 식

1982년 인하대학교 전자계산학과 졸업(학사)
1985년 Queens College/City University of New York (MA in computer Science)
1991년 City University of New York (Ph. D in Computer Science)
1991～현재 인하대학교 컴퓨터정보공학과 교수
2006～현재 BK21 지능형 유비쿼터스 물류 기술 연구 사업단장
2008년 1월～2008년 12월 한국 지능정보시스템학회 회장
관심분야 : 인공지능, 온톨로지, CSP, 전자상거래, etc.
E-mail : gsjo@inha.ac.kr