협력적 여과 시스템에서 귀납 추리를 이용한 순위 결정

(Ranking by Inductive Inference in Collaborative Filtering Systems)

고 수 정 †

(Su-Jeong Ko)

요 약 협력적 여과 시스템은 새로운 사용자의 행위를 파악하고 사용자가 흥미로워할 아이템을 추천해주기 위해서 사용자들에 대한 새로운 정보를 필요로 한다. 이러한 정보를 획득하기 위하여 협력적 여과 시스템은 기존 데이터를 기반으로 학습을 하고, 그 결과에 따라 사용자에 대한 새로운 정보를 찾아낼 수 있다. 본 논문에서는 사용자에 대한 새로운 정보를 획득하기 위한 방법으로 귀납적 추리 방법을 제안하고, 추리된 사용자의 정보를 이용하여 아이템의 순위를 결정한다. 제안된 방법에서는 귀납적 기계 학습 방법인 NMF를 이용하여 사용자를 학습시켜서 모든 사용자들을 그룹으로 군집시키고, 각 그룹으로부터 카이제곱을 이용하여 그룹의 특징을 추출한다. 다음으로, 귀납 추리 방법의 하나인 베이지언 확률모델을 이용하여 새로운 사용자가 입력한 평가값과 각 그룹의 특징을 기반으로 사용자를 적합한 그룹으로 분류한다. 마지막으로, 사용자가 결측한 아이템을 대상으로 로치오(Rocchio) 알고리즘을 적용하여 아이템의 순위를 결정한다.

키워드 : 귀납 추리, 순위 결정, 카이제곱, 베이지언 확률 모델, 로치오 알고리즘, 협력적 여과 시스템

Abstract Collaborative filtering systems grasp behaviors for a new user and need new information for the user in order to recommend interesting items to the user. For the purpose of acquiring the information the collaborative filtering systems learn behaviors for users based on the previous data and can obtain new information from the results. In this paper, we propose an inductive inference method to obtain new information for users and rank items by using the new information in the proposed method. The proposed method clusters users into groups by learning users through NMF among inductive machine learning methods and selects the group features from the groups by using chi-square. Then, the method classifies a new user into a group by using the bayesian probability model as one of inductive inference methods based on the rating values for the new user and the features of groups. Finally, the method decides the ranks of items by applying the Rocchio algorithm to items with the missing values.

Key words : inductive inference, ranking items, chi-square, bayesian probability model, Rocchio algorithm, collaborative filtering system

1. 서 론

협력적 여과 시스템은 사용자들이 아이템에 대해 평가한 값을 이용하여 추천하는 방법을 사용하고 있다. 협력적 여과는 보통 두 가지의 응용 시나리오를 채택하고 있다. 첫번째, GroupLens[1]와 같이 사용자가 아이템에 대해 잠재적으로 가지고 있을 아이템에 대한 평가값을 예측하는 방법[2,3]이다. 두번째 경우는 아이템을 Top_N으로 정렬하고, 정렬된 아이템의 목록을 이용하여 추천하는 방법이다[4]. 대부분의 전자상거래 시스템에서는 내용없이 추천을 하기 때문에 결측치 예측 시스템보다는 정확도가 높은 순위정렬 방식인 Top_N 추천

· 이 연구는 인덕대학 연구비에 의해 수행되었음

† 종신회원 : 인덕대학 컴퓨터소프트웨어 교수
sjko@induk.ac.kr

논문접수 : 2010년 3월 30일
심사완료 : 2010년 7월 14일

정보과학회논문지 : 소프트웨어 및 응용 제37권 제9호(2010.9)

방식을 사용하고 있다[4]. Top_N 추천 방법을 사용하는 기존의 협력적 여과 방법으로는 EigenRank 방법[4], 사용자 프로파일과 질의어의 문맥을 이용하는 방법[5] 등이 있다. EigenRank 방법은 Kendall 순위 계수를 사용하여 아이템 순위간의 상관 관계를 결정하고, 연관 피드백을 사용하여 아이템의 순위를 정렬한다. 다음으로, 질의어 문맥을 이용하는 방법은 기계 학습을 사용하여 사용자 프로파일을 학습하고, 프로파일로부터 질의어를 분류한다. 분류된 질의어 집합으로부터 질의어의 문맥을 유추할 수 있으며 아이템의 순위는 사용자 프로파일과 분류된 분류집합으로부터 추출한다.

반면, 협력적 여과 시스템은 새로운 사용자의 행위를 파악하고 그 새로운 사용자에게 결측된 아이템의 값을 예측하거나 사용자가 흥미로워 할 아이템을 추천해주기 위해서 사용자들에 대한 새로운 정보를 필요로 한다. 이러한 정보를 획득하기 위하여 기존의 데이터를 기반으로 학습을 하고, 그 결과에 따라 사용자에 대한 새로운 정보를 획득할 수 있다. 이와 같이 사용자에 대한 새로운 정보를 획득하기 위한 방법으로, 귀납적 학습 방법[6]을 사용하는 여러 방법이 있다. 규칙 기반으로 귀납적 추천을 하는 방법[7]은 사용자를 분류할 비슷한 사용자들의 그룹인 α 그룹을 찾고, α 그룹의 자료를 기반으로 기초 자료를 보완함으로써 새로운 사용자가 누락한 불완전한 평가값을 규칙 기반에 의하여 귀납적으로 유추한다. [8]은 RACOFI(Rule-Applying Collaborative Filtering) 시스템을 제안하여, 협력적 여과 방법을 이용하여 평가값을 예측하고, 유추 방법에 의해 예측된 평가값을 보완한다. [9]는 확률적인 유추방법으로 협력적 여과 시스템의 사용자를 모델링하여 사용자에게 제품을 추천하는 방법을 사용한다.

본 논문에서는 협력적 여과 모델에서 귀납 추리를 이용하여 아이템의 추천 순위를 결정하는 방법을 제안한다. 제안된 방법에서는 NMF(Non-negative Matrix Factorization)를 사용하여 사용자를 학습시킴으로써 모든 사용자들을 그룹으로 분류하고, 카이제곱을 이용하여 각 그룹의 특징을 추출한다. 다음으로, 새로운 사용자에게 아이템을 추천하기 위해 새로운 사용자가 입력한 평가값과 각 그룹의 특징을 기반으로 귀납 추리 방법의 하나인 베이지언 확률 모델을 이용하여 새로운 사용자를 적합한 그룹으로 분류한다. 마지막으로, 사용자가 결측한 아이템을 대상으로 로치오(Rocchio) 알고리즘을 적용하여 아이템의 추천 순위를 결정한다.

귀납 추리를 이용한 아이템 순위 추천 방법은 순위 추천를 위해 귀납적 방법을 적용한 방법이기 때문에 여러 가지 관점에서 기존의 협력적 여과 방법이 갖는 단점들을 보완할 수 있다. 구체적으로 기술하면 다음과 같다.

첫째, 오프라인 상에서 사용자를 군집시키고 그룹의 특징을 추출하고, 온라인 상에서 아이템의 순위를 추리하는 방법을 사용함으로써 사용자가 증가될 경우 발생하는 기존의 모델 기반 협력적 여과 방법[2]의 문제점을 보완한다.

둘째, 순위를 결정하여 추천함으로써 규칙 기반의 귀납적 추천[7], RACOFI[8], 그리고 확률적인 방법을 이용하는 방법[9]과 같이 기존의 귀납적 학습이 갖는 평가값 예측 기반의 오류로 인해 발생하는 성능 저하 문제를 보완한다.

셋째, [4]와 같이 사용자간의 유사도만을 이용하지 않고 NMF를 이용한 군집 방법을 사용하며, 군집의 특징 추출에 사용자 평균값만을 사용하는 방법[8]과 달리 카이제곱 통계량을 이용하여 군집된 그룹의 특징을 보다 정확도 높게 추출한다.

마지막으로, 학습된 결과를 기반으로 새로운 사용자를 그룹으로 분류하고 분류된 그룹 내에서 로치오 알고리즘을 이용하여 아이템의 순위를 귀납적으로 추론함으로써 기존의 순위를 결정하는 방법[4,5]보다 추천의 정확도를 높인다.

본 논문의 구성은 다음과 같다. 2장에서는 전체 시스템 구성도를 보이며, 3장에서는 사용자-아이템 행렬을 학습하여 사용자를 군집시키고, 그룹의 특징을 추출하는 방법을 기술한다. 또한, 4장에서는 귀납 추리를 이용하여 아이템의 추천 순위를 결정하는 방법을 제안하고, 5장에서는 성능 평가 방법과 평가 결과를 기술한다. 마지막으로 6장에서는 결론을 제시한다.

2. 전체 구성도

그림 1은 협력적 여과 시스템에서 귀납 추리를 이용한 아이템 순위 결정 시스템의 구성도를 보인다. 구성도는 귀납 학습 단계와 귀납 추리 단계로 구분된다.

귀납 학습 단계에서는 사용자-아이템 행렬의 사용자를 군집시키고, 군집된 각 그룹들의 특징을 추출한다. 사용자 군집을 위하여 기계 학습과 데이터 마이닝 분야 등 많은 분야에 적용되고 있는 NMF[10-12]를 사용한다. 또한, 그룹의 특징을 추출하기 위해서는 정보 검색 분야에서 문서의 특징을 추출하는 용도로 효과적인 방법인 카이제곱 통계량[13]을 이용한다.

귀납 추리 단계에서는 학습 단계에서 학습한 결과를 기반으로, 새로운 사용자가 결측하여 평가한 아이템에 대한 순위를 결정한다. 이와 같이 아이템의 순위를 결정하기 위해 우선적으로, 귀납적 기계 학습을 통해 생성된 그룹의 특징을 이용하여 새로운 사용자의 그룹을 결정한다. 다수의 사용자를 적당한 그룹으로 분류하기 위해 기계 학습 분야에서 많이 사용되고 있는 알고리즘으로

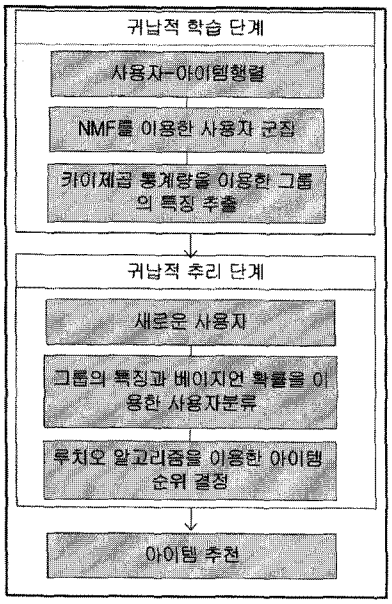

그림 1 귀납 추리를 이용한 아이템 순위 결정 시스템을 위한 구성도

는 규칙 기반 모델[14]과 귀납적 학습 모델이 있다. 규칙 기반 모델은 범주간의 규칙을 전문가가 찾아주거나 학습을 통한 규칙을 발견하여 사용자를 분류하는 방법이며, 귀납적 학습 모델은 사용자로부터 특징을 추출하여 이를 기반으로 확률을 사용하여 분류하는 방법이다. 제안된 방법에서는 협력적 여과 시스템의 사용자-아이템 행렬의 데이터가 대용량성이며 많은 경우의 수가 존재하므로, 평가값에 대한 경우의 수를 확률 모델에 적용하는 귀납적 학습 모델인 베이지언 확률 모델[15,16]을 이용한다.

마지막으로, 베이지언 확률을 이용하여 분류된 사용자 정보와 그룹의 특징을 기반으로 정보 검색 분야에서 많이 사용되는 질의 확장 방법인 루치오 알고리즘[17]을 협력적 여과 시스템에 적용하여 아이템의 순위를 결정한다.

이와 같이 귀납 학습 단계와 귀납 추리 단계를 통해 추출한 아이템의 순위를 이용하여 아이템을 추천함으로써 사용자는 추천 목록에서 가장 상위 아이템부터 하위 아이템을 찾아보며 흥미가 있는 아이템을 선택할 수 있다.

3. 귀납적 학습을 통한 그룹의 특징 추출

본 장에서는 아이템에 대한 순위를 유추하기 위하여 우선적으로 사용자-아이템 행렬을 귀납적으로 학습하여 사용자를 군집시키고, 군집된 그룹의 특징을 추출하는 방법을 기술한다.

3.1 NMF를 이용한 사용자 군집

NMF는 객체를 각 객체에 대한 부분정보의 조합으로 인식하는 정보에 착안하여 객체정보를 기초 자질(base feature)과 부호 자질(encoding feature)로 구분하여 표현한다[11]. 따라서, 이러한 정보를 협력적 여과 시스템에 적용한다면 최소한의 인수만으로도 사용자의 선호도에 영향을 주고, 각 인수가 사용자와 아이템에 어떻게 적용되는가를 결정함에 의해 아이템에 대한 사용자의 선호도를 결정할 수 있다는 장점을 갖는다[18,19]. 또한, NMF는 여러 분야의 연구에 사용된 방법으로, 이 방법이 사용자 군집에 사용되었을 때 속도면에서는 K-means와 동등한 성능을 나타내고, 정확도면에서도 높다는 결과가 나와서 그 우수함이 증명된 방법이기도 하다[10]. NMF를 문서 군집에 적용하는 방법[20]에서는 문서를 의미를 갖는 벡터로서 표현하였다. 본 논문에서 제안된 방법에서는 문서를 사용자-아이템 행렬의 사용자에 적용시켜 사용자를 의미 관계를 갖는 그룹으로 군집시킨다.

식 (1)은 사용자 군집에 사용할 사용자-아이템 행렬(M)을 기저 벡터(base vector)의 집합(B)과 은닉 벡터(hidden vector)의 집합(H)으로 분해한다. 그리고, 식 (2)는 행렬(M)에 대한 양의 인수를 찾기 위한 목적으로 사용한다. 식 (2)는 B와 H의 곱과 M과의 차인 A를 최소화하는 비음수 행렬 B와 H를 찾을 때까지 반복되어 계산된다.

$$M = BH \tag{1}$$

$$A = \frac{1}{2} \| M - BH \| \tag{2}$$

식 (2)의 목적 함수 A의 값이 수렴 허용오차보다 작아지기 위해서는 B와 H의 원소값을 반복적으로 갱신하는 작업이 필요하며, 이를 위해 식 (3)의 계산식을 이용한다.

$$b_{ij} \leftarrow b_{ij} \frac{(MH^T)_{ij}}{(BHH^T)_{ij}}, h_{ij} \leftarrow h_{ij} \frac{(B^TM)_{ij}}{(B^TBH)_{ij}} \tag{3}$$

식 (3)에서 B^T는 행렬 B의 전치행렬이며 H^T는 행렬 H의 전치행렬을 의미한다. 또한, b_{ij}와 h_{ij}는 각각 행렬 B와 행렬 H의 요소이다.

다음으로, 분해된 기저벡터의 집합(B)을 정규화하여 1의 크기로 만든 뒤, 그 가중치를 은닉벡터(H)에 가중한다.

표 1은 NMF를 설명하기 위하여 10명의 사용자로 구성된 사용자-아이템 행렬에 NMF를 적용한 결과이다. 표 1에 나타난 값은 속성의 크기를 3으로 정의하고 계산된 사용자의 의미 속성 가중치를 의미한다.

표 1을 기반으로 사용자-아이템 행렬의 사용자를 3개의 의미 속성에 대한 3차원 가중치 벡터로써 표현할 수 있다. 이들 사용자간의 유사도는 각각의 가중치 벡터의

표 1 가중치를 갖는 은닉벡터 - 사용자의 의미 속성 가중치

	u1	U2	u3	u4	U5	u6	U7	u8	u9	u10
의미 속성1	0.752	0.165	0.148	0	0.747	0.651	0.548	0.593	0.006	0.303
의미 속성2	0.168	0.649	0.771	0.002	0.272	0.000	0.394	0.432	0.905	0.369
의미 속성3	0.131	0.186	0.082	0.98	0	0.381	0.125	0.023	0.029	0.398

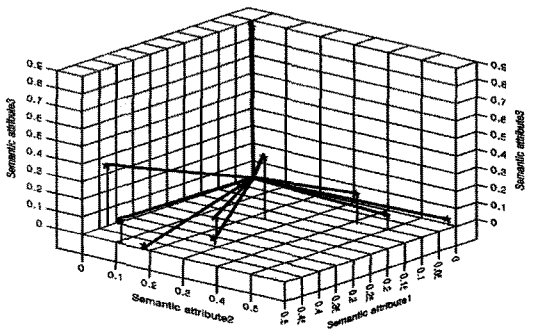

그림 2 NMF를 이용한 사용자 군집

코사인 유사도를 계산함으로써 구할 수 있다. 그림 2는 표 1의 사용자들을 벡터로 표현하여 가시화시킨 결과를 나타낸다. 벡터간의 거리를 기반으로 0.5이상의 코사인 유사도를 갖는 사용자들을 같은 군집으로 군집시킨다.

3.2 카이제곱을 이용한 군집의 특징 추출

정보 검색 분야에서는 문서의 특징을 추출하기 위하여 많은 연구가 제안되었으며[13], 제안된 방법 중 카이제곱 통계량과 정보 획득량의 방법이 가장 효과적이었다는 연구 결과를 기술하였다. 본 논문에서 제안한 방법에서는 그룹의 특징을 추출하기 위하여 정보 획득량보다는 계산량이 적은 카이제곱 통계량을 이용한다.

임의의 변수는 수치변수와 범주변수, 총 두 가지 종류로 구분할 수 있다. 범주변수는 그 범주 내에서 결과를 찾아 저장하며, 수치변수는 수치적인 형태로 데이터를 저장한다. 예를 들어, "당신의 전공이 무엇이나?" 또는 "당신은 차가 있는가?"라는 질문에는 각각 "전자계산학", "아니오"라는 답과 같이 분류적인 결과의 답을 산출한다. 그러나 "당신의 키는 얼마입니까?" 또는 "당신의 G.P.A는 얼마입니까?"라는 질문의 답에는 수치적인 정답을 만들 수 있다. 카이제곱은 범주변수의 일종으로, 비독립적인 범주들에 속하는 데이터값이 얼마나 서로 서로 다르게 분배되었는가의 정도를 평가하고, 종속적인 그룹 사이에 속한 범주데이터의 수를 계산한다[21]. 일반적으로, 대용량 데이터는 기존의 조사나 실험에 의한 자료에서 발생하는 변수의 수와는 비교할 수 없을 정도로 많은 변수를 발생시킨다. 때로, 데이터의 용량이 컴퓨터의 처리 능력을 넘어서는 경우도 있고, 모형화에 소요되는 시간이 너무 많이 커서 이를 기다릴 수 없는 경우도 있다. 이러한 경우, 카이제곱 통계량은 쓸모 있는 변수들을 빠르게 선택하여 줌으로써 시간을 단축시킬 수 있다. 따라서, 카이제곱 통계량을 협력적 여과 시스템에 적용시켰을 때 여과 시스템의 대용량 데이터로 인해 발생하는 문제를 해결할 수 있다.

표 2 아이템i_j 와 그룹k와의 공기 빈도

아이템 \ 그룹	G_k 에 할당됨	G_k 에 할당되지 않음
적합	A	B
비적합	C	D

카이제곱 통계량을 협력적 여과 시스템에 적용하기 위하여 우선적으로, j번째 아이템i_j 와 그룹k(G_k)와의 공기의 빈도를 표 2와 같이 정의한다. 협력적 여과 시스템의 사용자들은 그룹1~그룹t 중 하나의 그룹으로 분류되며, 각각의 그룹은 $\{G_1, G_2, ..., G_k, ..., G_t\}$와 같이 표기한다. 공기의 빈도는 사용자가 아이템i_j 에 대해 평가한 평가값을 기반으로 한다. 또한, 아이템i_j 가 그룹k에 적합한지 아닌지의 기준은 평가값이 0과 1사이이므로 평가값이 0.5보다 클 경우 사용자가 아이템을 선호한다고 판단하여 적합하다고 결정하고, 그렇지 않을 경우 적합하지 않다고 결정한다.

식 (4)는 아이템i_j 의 그룹k에서의 카이제곱 통계량 값을 정의한다[13].

$$\chi^2(i,G_k)=\frac{m_i\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (4)$$

식 (4)에서 m_i 은 A+B+C+D의 결과로 사용자-아이템 행렬의 사용자들 중 아이템i_j 에 대해 평가한 사용자수로 정의한다. A는 그룹k에 속한 사용자 중에서 아이템i_j 를 평균이상으로 평가한 평가값의 합, B는 그룹k에 속하지 않는 사용자 중에서 아이템i_j 를 평균이상으로 평가한 평가값의 합이다. 그리고, C는 그룹k에 속한 사용자 중에서 아이템i_j 을 평균이하로 평가한 사용자의 수로부터 아이템을 평균이하로 평가한 평가값의 합을 뺀 값을 나타낸다. 마지막으로, D는 그룹k에 속하지 않는 사용자들 중 아이템i_j 를 평균이하로 평가한 사용자의 수로부터 평균이하로 평가한 평가값을 뺀 값을 나타낸다.

이와 같이 정의한 A, B, C, D를 계산하기 위한 식은 식 (5), 식 (6), 식 (7), 그리고 식 (8)과 같다.

$$A = \sum_{u_i \in G_k \,\&\, r_{ui} > 0.5} r_{ui} \quad (5)$$

$$B = \sum_{u_i \notin G_k \,\&\, r_{ui} > 0.5} r_{ui} \quad (6)$$

$$C = l - \sum_{u_i \in G_k \,\&\, r_{ui} < 0.5} r_{ui} \quad (7)$$

$$D = q - \sum_{u_i \notin G_k \,\&\, r_{ui} < 0.5} r_{ui} \quad (8)$$

식 (5), 식 (6), 식 (7), 그리고 식 (8)에서 G_k는 그룹 k를 나타내며, u_i는 사용자u_i를, 그리고 r_{ui}는 사용자u_i가 아이템i_j에 대해 평가한 값을 나타낸다. 또한, 식 (7)에서 l은 그룹k에 속한 사용자 중에서 아이템i_j를 평균 이하로 평가한 사용자의 수를 나타내고, 식 (8)에서 q는 그룹k에 속하지 않은 사용자 중에서 아이템i_j를 평균이하로 평가한 사용자의 수를 나타낸다.

반면, 협력적 여과 시스템에서 사용자-아이템 행렬은 일부 아이템에 대한 평가값이 매우 희박하다는 특성을 갖는다. 그래서, 0~0.2의 평가값을 갖는 아이템의 경우와 평가값의 대부분이 결측된 경우 모두 식 (4)에 대입하면 분모가 0인 경우가 발생한다. 이와 같은 문제점을 해결하기 위해, 이러한 경우 식 (4)의 카이제곱 통계량을 0으로 계산한다.

표 3은 카이제곱 통계량을 계산하기 위해 Each Movie 데이터 집합에서 무작위로 추출한 13명의 사용자들이 평가한 0~1사이의 평가값과 NMF를 사용하여 분류된 그룹의 번호를 나타낸다.

표 3과 같이 군집된 사용자들을 대상으로 식 (4)를 적용하여 카이제곱 통계량을 계산하였을 경우, 표 4와 같은 값으로 결과값이 나온다.

표 4와 같이 계산된 카이제곱을 기반으로 그룹의 특징을 추출할 수 있다. 그룹k의 특징은 벡터의 형태로 정의하며, 표 4에 기술된 카이제곱 통계량을 가중치로 정한다. 식 (9)는 그룹k의 특징벡터를 식으로 정의한 결과이다.

$$\overrightarrow{f_{G_k}} = \{\chi^2(i_1, G_k), \chi^2(i_2, G_k), ..., \chi^2(i_n, G_k)\} \quad (9)$$

4. 귀납 추리를 이용한 선호도 순위 결정

추천 시스템에서 사용자는 추천된 목록 중 가장 상위의 아이템을 선호한다. 사용자는 그 목록에서 가장 상위 아이템부터 하위 아이템을 찾아보며 흥미가 있는 아이템을 선택할 수 있다. 이와 같은 Top_N 추천 방식을 제공하기 위하여 본 장에서는 아이템의 순위를 결정하는 방법을 기술한다.

4.1 베이지언 확률 모델을 이용한 사용자 분류

베이지언 확률 모델은 새로운 사용자가 그룹으로 분류될 확률을 계산하기 위하여 사용자가 평가한 아이템과 그룹과의 공기 확률값을 사용한다[16]. 식 (10)은 각 그룹k에 대한 사용자 u_i의 확률값 $\Pr(G_k \mid u_i)$을 Bayes 규칙을 이용하여 정의한다[22].

$$\Pr(G_k \mid u_i) = \Pr(G_k) \times \frac{\Pr(u_i \mid G_k)}{\Pr(u_i)} \quad (10)$$

협력적 여과 시스템에서 아이템간의 발생은 독립적이므로, 식 (11)의 Naïve Bayes 분류자를 사용하여 사용자u_i를 확률값이 가장 높은 값을 갖는 그룹으로 분류한다.

$$g^o(u_i) = \arg\max_{G_k} p(G_k) \prod_{i \in H} \Pr(i \mid G_k)^{n(i_j, u_i)} \quad (11)$$

$p(G_k)$는 최대 우도 측정을 이용하여

표 3 카이제곱 통계량을 계산하기 위한 사용자-아이템 행렬

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}	그룹
u_1		0.6	0.4	0.8	0.2	0.2	0.4	0.8	0.2	0.6	0.8	0.4	0.8	2
u_2	0.8	0.4	0.4	0.8	0.8	0	1	0.6	0.6	1	0		0.6	2
u_3	1	0.8	0.8	0.8	0.8			0.8	0.6	0.8	0	1	1	2
u_4	0.2	0	0.4	0.4		0	0.6	0	0.4	0	0.4	0.4		3
u_5	0.6		0.4	0.8	0.4	0	0.4	1	0	0.8	0.8	1	0.4	1
u_6	0.8	0	0.2	1	0	0	0.6	0.4		1	1	0.4	1	1
u_7	0.6	0.6	0.6	0.8	0	0.2	0.6	0.8	0.6	0.8		0.8		1
u_8	1	0.8	0.8	0.8	0.2	0.4	0.8	0.8	0	0.8	1	1		1
u_9	1	0.8	0.8	0		0.2	0.6	0	0.8	0.8	0.8	1	0.4	2
u_{10}	0.6	0.6	1	0.8		0	0.8	0.2		0.8	0.8	0.6	0.8	3

표 4 아이템에 대한 카이제곱 통계량

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}
그룹 1	0.804	0.024	0.08	0.23	3.778	0	0.064	1.36	1.022	0.919	2.52	0.258	0.211
그룹 2	0.716	0.754	0.006	0.442	3.778	0	1.225	0.275	2.25	0.661	1.714	0.011	0.025
그룹 3	4.661	0.69	0.056	0.714	0	0	1.105	2.5	0.68	4.955	0.08	0.667	0.467

$p(G_k) = \frac{|G_k|}{\sum_{r=1}^{t}|G_r|}$ 로 정의한다. 아이템의 집합, I = {i_1,i_2,i_3,…,i_n}와 같이 정의하고, 사용자u_i가 아이템에 대하여 평가한 값의 집합을 H라고 하였을 때, 식 (12)는 H를 정의한다. 식 (12)는 사용자u_i가 모든 아이템에 대해 평가한 값 중에서 결측치를 H로부터 제외시킴을 의미한다.

$$H = \{r_{ui} \mid 0 < i \le n, i \in N, r_{ui} \neq \phi\} \quad (12)$$

식 (12)에서 N은 자연수의 집합을 의미하고, n은 아이템의 수를 나타낸다. 그리고, r_{ui}는 사용자u_i가 아이템 i_j에 대해 평가한 값을 나타낸다.

식 (11)에서 그룹에서의 각 아이템 발생 확률은 3.2절에서 정의한 식 (4)로 정의한다. 또한, 식 (11)의 $n(i_j,u_i)$는 사용자u_i가 아이템에 대해 평가한 수를 나타내며, 사용자가 아이템에 대해 평가한 값 r_{ui}로 대체한다. 이와 같은 이론을 기반으로 식 (11)을 식 (13)으로 변경한다.

$$g^o(u_i) = \underset{G_k}{\operatorname{argmax}} p(G_k) \prod_{i \in H} \chi^2(i, G_k)^{r_{ui}} \quad (13)$$

반면, 식 (13)에 의한 결과는 평가값이 0일 경우나 사용자-아이템 행렬에 결측치가 있을 경우 모두 평가값이 0으로 지정될 뿐 아니라 행렬의 요소가 모두 0~1이하이기 때문에 그 결과가 상당히 작은 값으로 나타난다. 또한, 사용자가 아이템에 대해 평가를 많이 할수록 결과값이 작아지는 결과도 나타난다. 따라서 식 (13)의 결과에 로그를 적용하는 것이 바람직하나 $p(G_k)$, r_{ui}, 그리고 $\chi^2(i,G_k)$의 값은 0의 값도 존재하고, 대부분 1보다 작은 값이므로 로그식에 그대로 적용할 수 없다. 이에 대한 보완적인 방법으로 각각의 값에 1을 더하여 로그를 적용시켜서 $p'(G_k) = p(G_k) + 1$, $r'_{ui} = r_{ui} + 1$, 그리고 $\chi^{2'}(i,G_k) = \chi^2(i,G_k) + 1$로 변경하고, 그 결과를 식 (14)와 같이 적용하여 식 (13)을 재정의한다.

$$g^o(u_i) = \underset{G_k}{\operatorname{argmax}} (\log p'(G_k) + \sum_{i \in H} r'_{ui} \log \chi^{2'}(i, G_k)) \quad (14)$$

식 (14)를 이용하여 각 사용자가 각 그룹에 속할 확률을 정렬하여 확률이 가장 크게 나타난 그룹으로 사용자를 분류한다.

4.2 로치오 알고리즘을 이용한 선호도 순위 결정

4.1절에서와 같이 그룹으로 분류된 사용자에게 아이템을 추천하기 위하여 아이템들의 순위를 결정한다. 정보검색 분야에서는 사용자가 제시한 질의어를 확장하여 최종적으로 질의를 재생성함으로써 입력하지 않은 질의어에 대해서도 해당하는 문서를 검색할 수 있도록 하는 질의 확장 방법이 있다[23]. 이와 같은 방법은 재현율을 높이므로 사용자의 만족도를 높이는 방법으로 정보검색 분야에서 많이 사용되어왔던 방법이다. 제안된 방법에서는 질의어 확장 원리를 아이템 순위를 결정하는 방법에 적용한다. 3.2절에서 계산한 군집별 특징 벡터의 가중치는 사용자-아이템 행렬의 모든 사용자를 대상으로 계산된 결과이므로, 모든 사용자가 아닌 성향이 비슷한 사용자만을 대상으로 질의어 확장 원리를 기반으로 아이템의 순위를 결정하는 과정이 필요하다.

Salton과 Buckley[17]는 질의 용어 가중치 재산정 및 질의 용어 확장 방법을 이용한 실험을 하였다. 실험의 대상은 표준 벡터 피드백 방법인 Ide Dec-Hi, Ide Regular, 그리고 로치오 알고리즘에 기반한 방법이며, 이 실험에서는 문서 벡터와 초기 질의 벡터를 병합하는 방법을 사용하였다. 그 결과, 적합하지 않은 문서에서 용어가 발생할 경우 가중치를 줄이고 적합문서에서 용어가 발생할 경우에는 가중치를 증가하는 방법으로 용어의 가중치를 재산정할 수 있었다. 식 (15)는 로치오 알고리즘, 식 (16)은 Ide Regular 알고리즘, 그리고 식 (17)는 Ide Dec-Hi 알고리즘이다.

$$Q_1 = Q_0 + \beta \sum_{q=1}^{n1} \frac{R_q}{n_1} - \gamma \sum_{q=1}^{n2} \frac{S_q}{n_2} \quad (15)$$

$$Q_1 = Q_0 + \sum_{q=1}^{n1} R_q - \sum_{q=1}^{n2} S_q \quad (16)$$

$$Q_1 = Q_0 + \sum_{q=1}^{n1} R_q - S_q \quad (17)$$

제안된 방법에서는 식 (15), 식 (16), 그리고 식 (17)의 방법 중 아이템의 평가값 평균과 적합하지 않은 아이템의 평가값 평균을 적용할 수 있는 식 (15)의 로치오 알고리즘을 협력적 여과에 적용하여 아이템에 대한 선호도를 재산정하고 아이템들의 순위를 결정한다.

식 (15)의 로치오 알고리즘을 이용하기 위하여 Q_0의 값은 식 (14)을 이용하여 분류된 그룹의 식 (9)와 같은 형태의 특징벡터의 가중치 값으로 정한다. 또한, β, γ는 정보 검색 분야에서 적합 문서와 비적합 문서의 값을 조정할 때 사용하는 상수이나 협력적 여과에 적용할 때는 사용자가 평가한 아이템이 적합할 경우나 적합하지 않을 경우 모두 같은 가중치를 갖으므로 β, γ의 값을 모두 1로 정한다. n_1은 사용자가 분류된 그룹의 사용자들 중 해당 아이템에 대해 높은 선호도를 갖는 사용자의 수이며, n_2는 해당 아이템에 대해 선호도를 낮게 평가한 사용자 수이다. 즉, R_q는 새로운 사용자가 분류된 그룹의 사용자들 중 해당 아이템을 0.5이상으로 평가한 값을 나타내며, S_q는 0.5이하로 평가된 아이템들의 평가값을 1로부터 뺀 값을 나타낸다. 예를 들어, 아이템에 대하여 0.2로 평가했다는 것은 평가된 아이템이 사용자에게 0.8의 가중치만큼 부적합하다는 특성을 갖기 때문이다.

표 5 새로운 사용자u_a가 아이템에 평가한 예

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}
u_a		0.4		0.8	0	0.2		0.6	0.2		0.8		1

표 5는 새로운 사용자u_a가 아이템에 대해 평가한 예이다. 사용자u_a는 아이템i_1, 아이템i_3, 아이템i_7, 아이템i_{10}, 아이템i_{12}에 대하여 그 평가값을 결측하였다.

사용자u_a가 평가한 값과 표 4의 그룹별 아이템 가중치를 기반으로 식 (14)에 적용하였을 경우, 이 사용자는 그룹1로 분류된다. 식 (9)에 나타난 형태의 그룹1 특징 벡터는 $\vec{f}_{g1}$ = {0.804,0.024,0.08,0.23,3.778,0,0.064,1..36,1.022,0.919,2.52,0258,0.211} 와 같으며, 이를 기반으로 식 (15)에 적용한다. 그 결과, {아이템i_1, 아이템i_3, 아이템i_7, 아이템i_{10}, 아이템i_{12}}의 가중치 값은 {0.804, 0.08, 0.064, 0.919, 0.258}로 계산된다. 따라서, 결측된 아이템은 {아이템i_{10}, 아이템i_1, 아이템i_{12}, 아이템i_7, 아이템i_3}의 순서로 정렬되어 추천된다.

5. 성능 평가

본 논문에서 제안한 귀납 추리를 이용한 순위 결정 방법(Ranking_induction)은 기초 자료가 없는 상태에서 규칙 기반으로 새로운 추천을 하는 방법(Rule_induction)[7], 협력적 여과와 추론 규칙을 병합하는 방법(C_RACOFI)[8], 그리고 확률을 이용하여 사용자 선호도를 귀납 추론하는 방법(Probabilistic_I)[9]과 이웃의 수를 변경시켜 가면서 순위 예측의 정확도를 평가하였다. 또한, 순위 예측을 목적으로 하는 협력적 여과 방법 중 평가값으로부터 사용자 선호도를 모델링함으로써 아이템 순위를 결정하는 방법(EigenRank)[4]과 다른 사용자뿐 아니라 현재 사용자의 연과 피드백을 사용하여 아이템의 순위를 다시 정렬하는 방법(Re-Ranking_C)[5]과도 사용자의 수를 변경시켜 가며 그 성능을 평가하였다.

5.1 평가 집합

본 논문에서 제안된 방법은 성능 평가를 위해 72,000명의 사용자와 1,628의 아이템을 가진 Each Movie 데이터 집합과 73,421의 사용자와 100개의 유머를 아이템으로 갖는 Jester 데이터 집합[24]을 사용하였다.

5.1.1 Each Movie 데이터 집합

실험을 위해, Each Movie 데이터로부터 40개의 아이템보다 더 많이 평가된 아이템을 갖는 11,000명의 사용자 만을 추출하였다. 10,000명의 사용자는 훈련집합으로 사용하였으며, 1,000명의 사용자는 테스트집합으로 사용하였다.

5.1.2 Jester 데이터 집합

Jester 데이터 집합은 -10.00~10.00사이의 값을 포함하며 99는 평가하지 않은 값임을 나타낸다. 한 사용자당 하나의 행을 나타내며, {5, 7, 8, 13, 15, 16, 17, 18, 19, 20}의 열에 대해서는 대부분의 사용자가 아이템에 대하여 평가를 하였다. 평가를 위해 Jester 데이터 집합으로부터 5,000명의 사용자를 무작위로 추출하고, 또한 평가 밀도가 높은 열 {5, 7, 8, 13, 15, 16, 17, 18, 19, 20}에 해당하는 아이템을 추출하였다. 4,000명의 사용자는 훈련집합으로 사용하였으며, 1,000명의 사용자는 테스트집합으로 사용하였다.

5.2 평가 척도

전통적인 평가 기반의 협력적 여과 알고리즘은 사용자 평가값을 얼마나 정확도 있게 예측하였는가를 평가한다. 이와 같은 평가 기반의 협력적 여과 알고리즘의 성능 평가에 사용되는 성능 평가 측정인자는 Mean Absolute Error(MAE)와 Root Mean Square Error (RMSE)로, 실제 평가값과 예측한 평가값 사이의 차이를 기반으로 성능을 평가한다[3]. 그러나 본 논문에서 제안한 방법은 평가값의 예측보다는 아이템 순위 조정 결과의 정확도를 향상시키는 데 있으므로 NDCG(Normalized Discounted Cumulative Gain)척도를 사용하고자 한다[4,25]. NDCG는 정보검색분야에서 검색된 문서의 적합성 여부를 판단하기 보다는 오히려 순위를 부여하고 그 부여된 순위가 정확도 있게 부여되었는가를 평가하기 위해 사용되는 데 많이 사용되고 있다[26]. 이와 같은 측정 척도를 협력적 여과 시스템에 적용시키기 위하여 사용자들이 평가한 평가값은 순위가 부여된 적합성의 판단으로써 제공된다.

NDCG 척도는 협력적 여과 시스템의 순위가 부여된 아이템 목록에서 상위의 Top_N을 평가한다. 사용자들의 집합을 Q로 정의하고, $R(u_i,p)$는 본 논문에서 제안한 방법에 의해 결측치를 예측하고 정렬한 후, 정렬된 p번째 위치의 예측값을 나타낸다. 사용자들의 Q집합에 관한 k번째 위치에서의 NDCG는 식 (18)로 정의된다.

$$NDCG\ (Q,k) = \frac{1}{Q}\sum_{u_i \in Q} Z_{u_i} \sum_{p=1}^{k} \frac{2^{R(u_i,p)}-1}{\log(1+p)} \quad (18)$$

식 (18)에서 Z_{ui}는 DCG의 값을 정규화하기 위한 정규화 인자로, NDCG의 값이 0~1사이의 값을 갖도록 하기 위한 인자이다. 이를 위하여, Z_{ui}는 식 (19)로 정의한다.

$$Z_{u_i} = \frac{1}{T_{u_i}} = \frac{1}{\sum_{p=1}^{k} \frac{2^{T(u_i,p)}-1}{\log(1+p)}} \quad (19)$$

식 (19)에서 $T(u_i,p)$는 예측값이 아닌 테스트 집합의 사용자 u_i가 평가한 실제 평가값으로, 정렬한 후 정렬된 p번째 위치의 예측값을 나타내며, T_{ui}는 이상적인 DCG를 나타낸다.

NDCG는 0부터 1의 값으로 클수록 보다 효율적인 순

위 성능을 나타낸다. NDCG는 log(1+p)의 수식을 사용함에 따라 순위가 상위에 속할수록 평가값에 민감하다는 특성을 갖는다. 이러한 특성은 추천 시스템의 순위의 질을 평가하는 데 바람직함을 시사한다.

5.3 성능 평가 결과

그림 3은 귀납 추리를 이용하는 방법과의 성능을 평가하기 위해 본 논문에서 제안한 Rangking_induction 방법, Rule_induction 방법, C_RACOFI방법, 그리고 Probabilistic_I 방법을 이웃의 수를 변경시켜 가면서 계산한 NDCG의 측정 결과를 보인다.

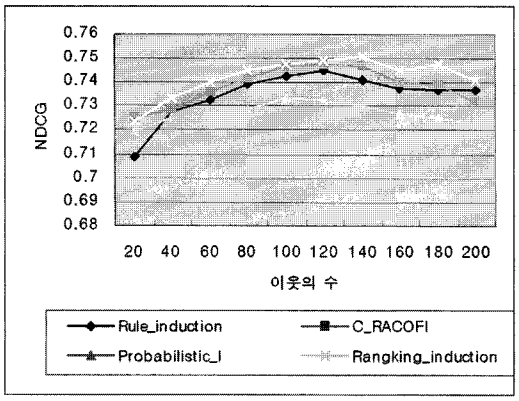

그림 3 귀납 추리를 이용하는 방법들의 NDCG

그림 3에서 Ranking_induction의 방법은 C_RACOFI와 Rule_induction의 방법보다 NDCG의 값이 약간 높으며, Probabilistic_I의 방법보다 현저히 높은 성능을 보인다. C_RACOFI 방법은 두 단계로 구분하여 추천하는 방법을 사용하고 있는 데, 첫 단계에서는 다른 사용자들이 평가한 평가값의 평균값만을 이용하여 결측치를 예측하고, 두번째 단계에서 유추에 의한 방법으로 예측값을 보완하는 방법을 사용한다. 이와 같은 방법은 첫번째 단계에서 평균값만을 이용하기 때문에 정확도가 낮으며 두 단계로 구분하여 진행하므로 속도도 낮다. Rule_induction 방법은 사용자를 가장 비슷한 흥미를 갖는 단체 α로 분류할 때의 정확도가 추천에 큰 영향을 미친다. 그러나 그 분류의 정확도가 낮을 뿐 아니라 단순히 아이템의 발생 빈도를 이용하기 때문에 제안한 방법에 비하여 성능이 낮다. Probabilistic_I 방법은 추천의 전처리로 사용자 프로파일을 구성해야 하는데 이를 위한 시간이 많이 소요되며, 확률적인 방법만을 사용하기 때문에 정확도가 낮다. 제안된 방법은 사용자 프로파일이 아니고 그룹 프로파일을 구성하며, 이 또한 오프라인 상에서 구성하기 때문에 Probabilistic_I의 방법보다 시간이 적게 소요된다.

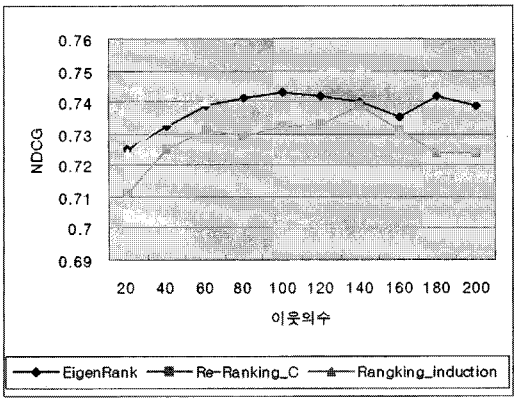

그림 4 순위 결정을 이용하는 방법들의 NDCG(Each Movie 데이터 집합)

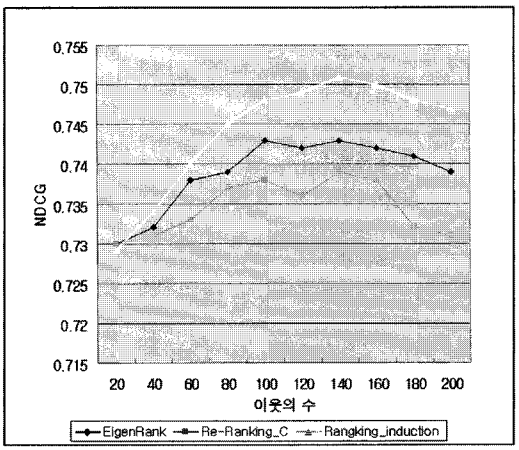

그림 5 순위 결정을 이용하는 방법들의 NDCG(Jester 데이터 집합)

그림 4와 그림 5는 순위를 결정하는 협력적 여과 방법들간의 성능을 평가하기 위해, Ranking_induction, EigenRank, 그리고 Re-Ranking_C과의 NDCG의 값을 이웃의 수를 변동시켜가며 Each Movie 데이터 집합과 Jester 데이터 집합을 대상으로 각각 측정한 결과를 보인다.

그림 4와 그림 5에서 Ranking_induction 방법은 EigenRank와 Re-Ranking_C의 방법보다 높은 성능을 보인다. EigenRank 방법은 새로운 사용자에 대한 초기 자료가 부족하였을 경우 사용자간의 유사도를 정확하게 계산하지 못하기 때문에 그 초기 평가 문제로 인하여 정확도가 낮다. Re-Ranking_C의 방법은 내용 기반에서 질의어의 전후 관계를 이용할 경우 정확도가 높아질 수 있으나 본 논문에서는 평가 기반의 자료를 사용하므로 정확도가 낮다는 결과를 갖는다. 따라서 평가 기반의 방

법에서도 질의어의 전후 관계를 이용할 수 있는 연구가 필요하다. 또한, Each Movie 데이터 집합을 평가한 경우와 Jester 집합을 대상으로 평가한 결과, 평가를 위해 수집한 Jester 데이터 집합이 희박성이 적은 아이템으로 구성되었고 사용자가 아이템에 대해 평가한 값이 정밀하므로, 그 평가 결과의 곡선이 Each Movie 데이터 집합를 사용하여 평가한 결과의 곡선보다 유연함을 보인다.

6. 결 론

본 논문에서 제안된 방법은 귀납적 학습 방법을 통하여 생성된 정보를 가지고 사용자의 행위를 추론하였으며, 추론된 사용자의 행위를 이용하여 아이템의 순위를 추리하는 협력적 여과 방법을 제안하였다. 성능 평가 결과, 제안한 방법은 기존의 협력적 여과 시스템이 갖는 단점을 보완할 수 있었으며, 귀납적 추론만을 사용하는 방법과 순위만을 추론하는 방법보다도 각각 높은 정확도를 보였다.

향후, 카이제곱 통계량을 바탕으로 군집의 특징 추출을 활용하는 방법에 대한 보다 정확도 높은 연구가 필요하다.

참 고 문 헌

[1] MovieLens collaborative filtering data set, "Http://www.cs.umn.edu/Research/GroupLens/index.html," GROUPLENS RESEARCH PROJECT, 2000.

[2] J. Pessiot, T. Truong, N. Usunier, M. Amini, and P. Gallinari, "Learning to rank for collaborative filtering," *Proceedings of the 9th International Conference on Enterprise Information Systems (ICEIS 2007)*, 2007.

[3] Breese, J. S., Heckerman, D., and Kardie, C., "Empirical analysis of predictive algorithms for collaborative filtering," *Proceedings of the fourteenth Conference on Uncertainty I Artificial Intelligence*, 1998.

[4] N. N. Liu and Q. Yang, "EigenRank: A Ranking-Oriented Approach to Collaborative Filtering," *Proceedings of ACM Conference on Research and Development in Information Retrieval (SIGIR'08)*, 2008.

[5] Rohini U and V. Varma, "A Novel Approach for Re-Ranking of Search results using Collaborative Filtering," *Proceedings of International Conference on Computing: Theory and Applications (ICCTA'07)*, 2007.

[6] Mitchell, K., *Machine learning*, McGraw Hill, New York, 1997.

[7] A. Nguyen, N. Denos, and C. Berrut, "Improving New User Recommendations with Rule-based Induction on Cold User Data," *Proceedings of the 2007 ACM conference on Recommender systems*, 2007.

[8] D. Lemire, H. Boley, S. McGrath, and M. Ball, "Collaborative Filtering and Inference Rules for Context-Aware Learning Object Recommendation," *International Journal of Interactive Technology and Smart Education*, vol.2, no.3, 2005.

[9] A. Eckhardt, "Induction of User Preferences in Semantic Web," *Proceedings of WDS'07*, 2007.

[10] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix trifactorizations for clustering," *Proceedings of SIGKDD*, 2006.

[11] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Advances in Neural Information Processing Systems*, 2001.

[12] S. Ko, "A Hybrid Collaborative Filtering Using a Low-Dimensional Linear Model," *Journal of KIISE : Software and Applications*, vol.36, no.10, Oct. 2009.(in Korean)

[13] Y. Yang and J. O. Pederson, "A comparative study on feature selection in text categorization," *Proceedings of the 14th international conference on Machine Learning*, 1997.

[14] C. Apte, F. Damerau, and S. M. Weis, "Towards language independent automated learning of text categorization models," *Proceeding of the 17th annual international ACM-SIGIR*, 1994.

[15] D. D. Lewis, "Navie(bayes) at forty: The independence assumption in information retrieval," *European Conference on Machine Learning*, 1998.

[16] A. McCallum and K. NIgram, "A comparison of Event Models for Naïve Bayes Test Classification," *AAAI'98 workshop on Learning for Text Categorization*, 1998.

[17] G. Salton and C. Buckley, "Improving Retrieval Performance by Relevance Feedback," *Journal of the American Society for Information Science*, vol.41, no.4, 1990.

[18] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," *Advances in Neural Information Processing Systems*, 2001.

[19] M. Wu, "Collaborative Filtering via Emsembles of Matrix factorizations," *Proceedings of KDD Cup and Workshop 2007*, 2007.

[20] W. Xu, X. Liu, and Y, Gong, "Document clustering based on non-negative matrix factorization," *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 2003.

[21] D. Eck and J. Ryan, http://math.hws.edu/javamath/ryan/ChiSquare.html, *Mathbeans Project*, 2009.

[22] I. S. Dhillon, S. Mallela, and R. Kumar, "Enhanced Word Clustering for Hierarchical Text Classification," *Proceedings of 8th ACM. SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

[23] H. Jung, "The Automatic Newspaper Summarization using Information Retrieval Method," Master Thesis, Sogang University, 2007.
[24] Ken Goldberg, http://goldberg.berkeley.edu/jester-data/, University of California, 2002.
[25] H. Valizadegn, R. Jin, R. Zhang, and J. Mao, "Learning to Rank by Optimizing NDCG Measure," *Advance in Neural Information Processing Systems (NIPS 23)*, 2009.
[26] K. Järvelin, and J. Kekäläinen, "Cumulated Gain-based Evaluation of IR Techniques," *ACM. Transactions on Information Systems*, vol.20, no.4, 2002.

고 수 정
1990년 인하대학교 전자계산학과 졸업(학사). 1997년 인하대학교 전자계산교육(석사). 2002년 인하대학교 전자계산공학과(박사). 2003년~2004년 Univ. of Illinois at Urbana Champaign Post Doc. 2004년~2005년 Colorado State University Research Scientist. 2005년~현재 인덕대학 컴퓨터소프트웨어과 교수. 관심분야는 데이터마이닝, 정보검색, 기계학습, 정보보안