협동적 필터링을 이용한 K-최근접 이웃 수강 과목 추천 시스템

손기락**, 김소현**
한국외국어대학교 컴퓨터및정보통신공학부*, 한국외국어대학교 교육대학원 전자계산교육전공**

요 약

협동적 필터링은 사용자가 좋아할 만한 항목을 예측하기 위하여 비슷한 선호도를 가지는 다른 사람들의 평가 항목에 근거하여 추천하는 방법이다. 이러한 협동적 필터링 기법은 오늘날과 같이 대규모의 정보가 효과적으로 축적되고 이용 가능하게 된 정보화된 사회에서는 현명한 의사결정을 하도록 도와주는 역할을 한다. 본 논문에서는 대학생들이 수강과목의 취사선택을 용이하게 할 수 있도록 수강과목 추천 시스템을 설계하고 구현하였으며 실험적으로 평가하였다. 먼저, 학생들은 과거 자신이 수강하였던 과목에 대한 과목 선호도를 데이터베이스에 입력한다. 과목 선호도의 패턴이 유사한 학생들은 유사 그룹으로 간주된다. 성향이 유사한 사용자를 찾기 위해 일반적으로 사용되고 있는 피어슨 상관계수에 의한 유사도를 이용하였다. 수강 과목을 예측하려는 학생과 가장 유사한 패턴을 보이는 K 명의 학생들의 수강 과목에서 가장 높은 선호도를 보이는 과목들의 순서화된 리스트를 추천 과목으로 제시한다. 설문 조사를 통한 실험 데이터를 이용하였으며 평균 절대 에러를 사용하여 제안한 방법의 정확도를 평가하였다.

키워드: 협동적 필터링, 과목 추천

K-Nearest Neighbor Course Recommender System using Collaborative Filtering

Kirack Sohn*, So Hyun Kim**
Hankuk University of Foreign Studies, School of Computer and Information Communication Engineering*,
Hankuk University of Foreign Studies, Graduate School of Education, Dept. of Computer Education**

ABSTRACT

Collaborative filtering is a method to predict preference items of a user based on the evaluations of items provided by others with similar preferences. Collaborative filtering helps general people make smart decisions in today's information society where information can be easily accumulated and

* 이 연구는 2007학년도 한국외국어대학교 교내학술연구비의 지원에 의하여 이루어진 것임

analyzed. We designed, implemented, and evaluated a course recommendation system experimentally. This system can help university students choose courses they prefer to. Firstly, the system needs to collect the course preferences from students and store in a database. Users showing similar preference patterns are considered into similar groups. We use Pearson correlation as a similarity measure. We select K-nearest students to predict the unknown preferences of the student and provide a ranked list of courses based on the course preferences of K-nearest students. We evaluated the accuracy of the recommendation by computing the mean absolute errors of predictions using a survey on the course preferences of students.

Keywords: collaborative filtering, course recommendation

1. 서론

인터넷의 발달로 온라인상에서 검색을 통하여 활용할 수 있는 정보의 양이 많아짐에 따라 정보의 홍수 문제가 대두되고 있다. 이러한 문제를 해결하기 위해 단지 정보를 제공하는 수동적 형태보다는 적극적으로 정보를 추천해주는 추천 엔진(recommendation engine)이 요구되고 있다. 이러한 요구에 따라 최근 전자상거래(e-commerce)를 기점으로 사용자의 구매 촉진을 위해 전자상거래 시스템을 이용하는 고객들로부터 얻어진 구매정보를 기초로 고객이 좋아할 만한 제품을 예측하여 고객에게 정보를 제공하는 추천시스템이 개발되고 있다[6][7][8].

이러한 추천시스템의 핵심은 추천방법을 표현한 추천 알고리즘에 있다. 다양한 형태의 추천 알고리즘이 있지만 추천 시스템에 가장 많이 사용되는 방법으로 협동적 필터링(collaborative filtering)을 이용한 알고리즘이 있다. 이 같은 협동적 필터링이 이용된 추천시스템은 추천결과가 상당히 우수하여 가장 많이 사용되는 방법으로 Amazon.com, CDnow.com 등 상업적으로 성공한 여러 전자상거래 사이트에서 적용되고 있다[7]. 협동적 필터링에는 사용자 기반 협동적 필터링(user-based collaborative filtering)과 아이템 기반 협동적 필터링(item-based collaborative filtering)이 있다.

첫 번째로 사용자 기반 협동적 필터링[5][7]은 사용자가 좋아할 만한 항목을 예측하기 위하여 비슷한 선호도를 가지는 다른 사용자들의 평가 항목에 근거하여 추천하는 방법이므로 높은 예측능력과 추천능력을 가지는 장점이 있다. 특정 사용자와 비슷한 선호도를 가지는 이웃들을 선정하는 기법에는 클러스터링(clustering), K-최근접 이웃(k-nearest neighbor), 베이지안 네트워크(bayesian network)와 같은 여러 가지 방법이 있으나 대부분의 경우 K-최근접 이웃 방법을 이용한다[5]. K-최근접 이웃 방식의 추천엔진은 다수의 평가자들 중에 평가에 참여할 사용자를 선택하고 또한 참여자에 대한 가중치를 부여하기 위해 사용자간의 유사도를 이용한다. 가장 일반적으로 정확하다고 여겨져서 사용되고 있는 유사도는 피어슨 상관계수(Pearson correlation)이다.

두 번째로 아이템 기반 협동적 필터링[5]은 대부분의 사람들이 과거에 자신이 좋아했던 항목과 비슷한 항목이면 좋아하는 경향이 있고 반대로 싫어했었던 항목과 비슷한 항목이면 싫어하는 경향이 있다는 점을 기반으로 하고 있다. 이 필터링 방법은 사용자가 선호도를 입력한 기존의 항목들과 예측하고자 하는 항목과의 유사도를 계산하여 사용자의 선호도를 예측하는 방법이다. 즉, 예측하고자 하는 항목과 비슷한 항목들에 대하여 사용자가 높은 평가를 하였다면 그 항목도 높게 평가할 것이라고 예측하고, 낮은 평가를 하였다면 그 항목도 낮게 평가를 할 것이라고 예측하는 것이다. 아이템 기반 협동적 필터링 방법은 항목들 간의 유사도를 계산하기 위하여 두 항목에 모두 선호도를 입력한 사용자들의 선호도를 사용한다. 그러나 사용자들 간의 유사

도가 전혀 고려되지 않기 때문에 만약 특정 사용자와 전혀 선호도가 비슷하지 않은 사용자들의 평가를 기반으로 한다면 항목들 간의 상관관계의 정확도가 떨어지며 수강과목을 추천하는 시스템에서는 사용자간 선호도가 고려되어야 하기 때문에 본 논문에서는 사용자 기반 협동적 필터링 방법을 이용한다.

본 논문의 구성은 다음과 같다. 2절에서는 관련 연구로서 협동적 필터링을 알아보고 기존 수강신청 시스템의 문제점에 대해 논하고, 3절에서는 협동적 필터링 기법을 이용하여 학생들이 과목을 수강할 때 수강했던 과목을 평가한 점수를 토대로 다른 학생들에게 수강 과목을 추천하는 K-최근접 수강신청 시스템에 대해 기술하고 4절에서는 시스템의 설계 및 구현에 대해 기술하고, 5절에서는 평균 절대 에러를 측정하여 해당 시스템의 성능을 분석한다. 마지막으로 6절에서는 결론과 향후 연구과제에 대하여 기술한다.

2. 관련 연구

2.1 기존의 협동적 필터링

협동적 필터링은 가장 우수한 성능을 보이는 추천시스템의 핵심기술이다. 협동적 필터링의 개념을 처음 소개한 Tapestry와 Tapestry의 단점을 극복한 최초의 자동화된 협동적 필터링(automated collaborative filtering) 알고리즘인 GroupLens에 대해서 알아본다.

2.1.1 Tapestry

Tapestry[1][9]는 Xerox Palo Alto Research Center에서 개발한 문서 필터링 시스템이다. Tapestry로부터 협동적 필터링의 개념이 유래되었다. Tapestry는 사용자가 그들이 읽은 문서에 주석(annotation)을 다는 것을 허락한다. 이렇게 문서에 대한 주석이 달려 있으므로 다른 Tapestry 사용자는 문서를 검색할 키워드 매칭을 통한 검색뿐만 아니라 다른 사용자의 문서에 대한 주석을 통해서도 검색할 수 있다. 이러한 주석은 형식이 없는 형태(free text)일 수도 있고 “like it”과 “hate it”으로 Tapestry에 의해 정해진 형태로 문서에 주석을 부여할 수도 있다. Tapestry는 클라이언트/서버 구조를 가지며, 문서를 필터링하기 앞서 주석 저장소로부터 다른 사용자의 주석을 참고해서 각 사용자의 리틀박스에 문서를 저장하게 된다. 이러한 방법은 TQL이라는 질의어를 사용자가 직접 만들어서 Tapestry 서버에 제공하여 이루어진다.

그러나 Tapestry는 다음과 같은 문제점을 가지고 있다. 첫째, 사용자가 문서에 대한 주석을 달지 않을 경우 대처방법이 없다. 만약 많은 사용자들이 문서에 대한 주석을 달지 않는다면 필터링 질의에 의존해서 키워드 매칭의 단순한 필터링만을 수행한다. 둘째, 사용자가 TQL이라는 질의를 직접 만들어야 한다. 이것은 사용자가 새로운 분야에 대한 필터링을 하고자 하는 경우에 TQL문 자체에 오류를 포함할 수 있는 가능성이 내재되어 있다.

이와 같은 문제점이 있지만 Tapestry는 협동적 필터링 개념을 처음으로 도입한 시스템이며 현재에도 대부분의 협동적 필터링 시스템들은 Tapestry를 바탕으로 연구가 진행되고 있다.

2.1.2 GroupLens

GroupLens[1]는 Tapestry의 문제점을 해결하면서 일반적으로 받아들여지는 성능을 보이는 Netnews(usenet news)에서 개별화된 추천을 제공하는 협동적 필터링이 적용된 시스템이다. GroupLens는 문서에 대한 선호를 숫자로 나타내며 사용자 프로파일에 이 정보를 포함시켜 서버에 저장한다.

GroupLens는 두 가지 평가 방법을 갖는다. 첫 번째는 사용자의 직접적인 평가에 의한 방법, 두 번째는 사용자의 직접적인 평가가 없는 경우 다른 사용자의 프로파일을 기반으로 한 상관관계(correlation)에 의해 문서에 대한 평가를 예측하는 방법이다.

GroupLens에서는 Tapestry에서 문제가 되었던 사용자가 읽은 문서에 대한 평가를 하지 않은 경우

는 같은 관심을 가지는 사람들의 평가를 기반으로 해결하려는 방법을 시도하였다.

2.2 기존의 수강 관련 시스템의 한계와 대안

기존의 수강 관련 시스템은 수강 신청이나 평가에 치중되어 학생들이 수강 신청을 할 때 얻을 수 있는 기초적인 자료는 수강 신청 시에 얻는 강의계획서, 선배들의 조언 등이 대부분이어서 수강 과목 추천의 형태가 수동적이고 추천의 범위가 제한적이었음을 알 수 있다. 학생들이 수강 과목 이후에 하는 과목 평가도 학교만의 자료로 이용되어 학생들에게는 정보가 공유되지 않았다.

따라서 학생들도 과목에 대한 정보를 공유하고 자동으로 추천해줄 수 있는 시스템의 필요성이 절실하며 특히, 자신과 성향이 같은 학생의 과목을 추천받는 사용자 기반 협동적 필터링을 이용하여 수강과목추천시스템을 구현한다면 매우 유용할 것이다.

3. K-최근접 이웃을 이용한 수강 과목 추천

수강과목 추천 시스템의 사용자 기반 협동적 필터링과 관련하여 사용자(이웃)를 선정하는 기법인 K-최근접 이웃 방법의 원리와 동작을 알아본다. K-최근접 이웃(K-nearest neighbor)은 문서와 문서 관련도(relevance feedback)를 이용하여 문서를 분류하는 방법에 자주 사용되었으며, 기억기반 추론(MBR: Memory Based Reasoning)이라고도 한다[3].

새로운 문서에 대한 범주를 결정할 때, 학습문서에서 그 문서와 가장 가까운 K개의 문서들을 추출하여, K개 문서가 속하는 범주를 이용하여 새로운 문서에 할당하게 된다. 이 때, 주어진 문서와 각 이웃문서의 관련도가 이웃문서의 범주를 가질 가중치로 합해진다.

유사성에 기반한 검색을 효율적으로 지원하는 K-최근접 이웃 방식은 평가에 참여할 사용자를 선택하고 참여자에 대한 가중치를 부여하기 위해 사용자간의 유사도를 이용한다.

추천 엔진의 동작 특성에서 예측하려고 하는 사용자와 유사한 K명의 사용자를 선별하는 단계는 시간이 가장 많이 걸리는 작업이다. 즉 K개의 이웃을 찾아내는 작업이다. K개의 이웃을 찾아내기 위해서는 데이터베이스에 있는 모든 사용자 정보를 스캔하여 유사도를 계산하여야 하기 때문에 시간이 가장 많이 걸리는 작업이다. 모든 사용자가 예측하려는 항목의 등급 결정에 참여한다고 하여 정확도가 향상되지 않는다. 실험을 통해서 가장 가까운 사용자의 수가 50명 이상이면 정확도가 더 개선되지 않음이 나타나고 있다[3].

알고리즘의 동작 설명을 위해 지금부터는 예측을 하려는 사용자를 활성 사용자(active user), 예측하려는 항목 즉, 활성 사용자가 아직 등급을 매기지 않은 항목을 활성 항목(active item)이라 부른다. 추천 엔진의 동작은 아래와 같은 3단계로 나누어진다.

Step 1. 모든 사용자에 대해 활성사용자와의 유사도에 따라 가중치를 부여한다.

Step 2. 예측에 참여할 사용자의 집단 즉, 활성 사용자와 유사도의 절대 값이 가장 큰 K명의 사용자를 선별한다. 유사도가 양수로 비슷한 사용자는 긍정적으로 영향을 미치고(positive combination), 반대의 성향을 가진 이웃, 즉 유사도가 음수로 가장 큰 사용자는 부정적으로 영향을 미친다(negative combination).

Step 3. 활성 항목에 대해 선택된 참여자의 등급의 가중치 조합(weighted combination)으로부터 예측치를 계산한다.

유사도로 피어슨 상관계수를 사용할 경우, 피어슨 상관 계수가 양수라는 것은 활성 사용자와 참여자의 성향이 비슷하다는 것을 나타낸다. 피어슨 상관계수가 음수라는 것은 참여자는 활성 사용자와 반대되는 행동을 하는 것으로 판단된다. 예를 들어, 수학 과목을 좋아하는 학생이 국어 과목을 싫어하는 경향이 있으며, 반대로 국어 과목을 좋아하는 학생은 수학과목을 싫어하는 경향이 있다. 이 경우 수

학 과목을 좋아하는 학생에 대한 활성 항목의 값을 결정할 때 수학 과목을 좋아하는 학생의 과목 선호도는 긍정적으로 영향을 미치고 국어 과목을 좋아하는 학생의 선호도는 부정적으로 영향을 미치는 것으로 판단하게 된다.

이웃에 근거한 알고리즘(neighbor-based algorithm)은 예측하려고 하는 사용자의 이웃에 가중치를 부여하고, 그 이웃이 주어진 항목에 대한 등급에 대한 가중치 평균(weighted average)을 예측값으로 사용한다. 활성 사용자를 a라 하고, 활성 항목을 i라고 하자. 활성 사용자 a의 항목 i에 대한 예측 값 $P_{a,i}$은 아래와 같은 식에 의해 계산된다. $r_{u,i}$는 사용자 u가 항목 i에 대해 매긴 등급이다. $\overline{r_u}$는 사용자 u의 등급 평균을 나타낸다. n은 예측 결정에 참여하는 이웃의 수를 나타낸다. <표 1>에서 user2의 국어과목에 대한 선호도를 예측하는 경우, u는 'user2'이고 i는 '국어'가 된다. $\overline{r_a}$는user2의 등급 평균 3이다. 사용자마다 선호도를 나타내는 방법이 다르기 때문에 활성 사용자의 평균을 이용할 필요가 있다. 즉 어떤 사용자는 높은 선호도 값을 이용하여 선호도의 차이를 나타내고 어떤 사용자는 넓은 범위의 선호도 값을 이용하는 경우가 있다. $r_{u,i}$는 사용자 a의 선호도 예측에 참여하는 사용자 u가 항목 i에 대한 선호도이다. <표 1>에서 'user1'은 'user2'의 '국어' 과목 선호도 예측에 참여하기 때문에 $r_{u,i}$는 4이다. 물론 이 값이 'user1'의 선호도 평균에서 얼마나 차이를 보이는가가 의미있기 때문에 아래 식에서처럼 user1의 선호도 평균을 빼야 한다. $W_{a,u}$는 활성 사용자와 예측에 참여하는 사용자 사이의 유사도이다.

$$P_{a,i} = \overline{r_a} + \frac{\sum_{u=1}^{n}(r_{u,i} - \overline{r_u}) * W_{a,u}}{\sum_{u=1}^{n} W_{a,u}}$$

a : 예측하려는 사용자

i : 예측하려는 항목

$r_{u,i}$: 사용자 u가 항목 i 에 대해 매긴 등급

$\overline{r_u}$: 사용자 u의 등급 평균

n : 예측 결정에 참여하는 이웃의 수

여기서, $W_{a,u}$는 활성 사용자 a 와 사용자 u 와의 유사도 가중치를 나타내는데, 가장 흔히 사용되는 피어슨 상관계수는 다음과 같다.

$$W_{a,u} = \frac{\sum_{i=1}^{m}(r_{a,i} - \overline{r_a}) * (r_{u,i} - \overline{r_u})}{\sigma_a * \sigma_u}$$

i : 예측하려는 사용자 a와 사용자 u가 함께 등급을 매긴 항목

σ_a : 사용자 a의 표준 편차

σ_u : 사용자 u의 표준 편차

피어슨 상관관계의 특성은 두 사용자가 함께 매긴 등급의 항목이 일치하면 즉, 두 사용자의 선호도가 일치하면, 유사도는 1이고, 항목에 대한 등급이 정반대이면 즉 선호도가 정반대이면 유사도는 -1이다.

<표 1> 과목 평가에 대한 예

	수학	과학	영어	국어	사회
user1	3	4		4	
user2	4	2		?	3
user3	4	3	3		3
user4		5	4	2	2
user5	2	5		3	5
user6	5	3	4	3	4

예를 들어, 각 학생들이 위의 표와 같이 5등급에 따라 과목들에 대해 평가를 내렸다고 가정해 보자. 여기서 등급 5는 가장 좋은 등급이고, 등급 1은 가장 낮은 등급이며, 공백은 아직 매기지 않은 등급을 의미한다. 이 경우 활성 사용자 'user2'의 활성 항목 '국어' 에 대해 매길 등급에 대해 예측하려 한다. 일단 이 경우에 모든 사용자가 모두 등급 결정에 참여한다고 가정한다. 'user2'과 다른 학생들의 유사도

를 먼저 계산해 보면 아래와 같다.

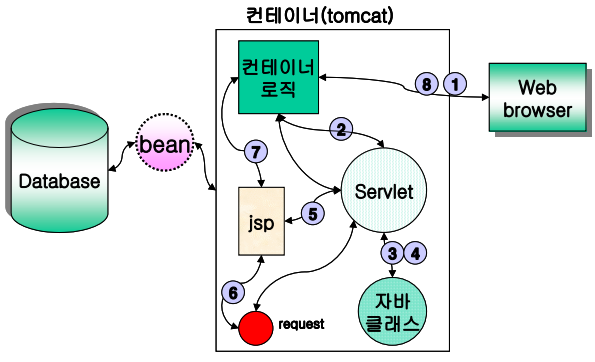
user2 - user1 : -1
user2 - user3 : 0.86
user2 - user4 : -1
user2 - user5 : -0.86
user2 - user6 : 1

위의 공식 $P_{a,i}$에 유사도를 대입해보면 국어 과목에 대해 3.20 정도의 등급을 예측한다.

4. 설계 및 구현

4.1 구성도

본 논문에서는 수강과목의 평가와 추천이 이뤄지는 시스템을 구성하기 위해 자바와 JSP를 프로그래밍 언어로 사용하여 웹 페이지를 구현하였다. 해당 시스템 구조는 다음과 같다. 프로그램의 유지보수를 용이하게 하고 가독성을 높이기 위해 재현 로직(presentation logic)과 업무 로직(business logic)을 분리하였다. 재현 로직을 JSP를 이용하였으며 업무 로직은 자바 클래스로 구현하였다. 데이터베이스 좁속은 자바 빈즈로 구현하였다. 따라서 업무 로직의 변화나 재현 로직의 변화가 프로그램의 미치는 영향을 최소화 하였다.



(그림 1) 시스템 아키텍처

다음은 위 (그림 1)에 표시된 클라이언트와 서버간의 데이터 흐름에 대한 설명이다..

① 사용자가 정보를 컨테이너로 보낸다.
② 컨테이너는 URL을 분석하여 담당 서블릿을 찾아 요청을 넘긴다.
③ 서블릿은 해당 자바 클래스를 호출한다.
④ 자바 클래스는 결과를 서블릿으로 넘겨주고 서블릿은 이 정보를 request 객체에 저장한다.
⑤ JSP에 이 Request 객체를 포워딩(forward)한다.
⑥ JSP는 서블릿이 넣어 놓은 정보를 Request 객체에서 추출한다.
⑦ JSP는 여기에 바탕하여 HTML 페이지를 작성한다.
⑧ 컨테이너는 이 페이지를 사용자에게 넘겨준다.

4.2 시스템 개발 환경

본 논문에서 시스템 개발을 위해 사용한 개발 환경은 <표 2>와 같다.

<표 2> 개발 환경

구 분	사 양
운영체제	Window XP
어플리케이션 서버	Jakarta-Tomcat 5.0.28
DBMS	MySQL 4.1.19-win32
JDK	J2SDK1.4.2_11
저작언어	Java (JSP, Servlet), HTML
웹 브라우저	Internet Explorer 6.0

5. 실험 및 결과 분석

수강 과목 선호도에 대한 실제 데이터가 존재하지 않기 때문에 제안한 방식의 유용성을 검증하기가 어렵다. 본 논문에서는 실험적 평가를 위해 서울 소재 모대학교 정보공학대학 학생들을 대상으로 하여 (그림 2)와 같이 수강 과목 설문조사를 토대로 자신이 수강하였던 과목에 대하여 선호도를 기술하게 하였다. 이 데이터 집합은 총 20명의 사용자가 12개의 과목에 대해 0 ~ 5 점 사이의 과목 선호도를 부여하고 선호도의 이유에 관해서도 작성할 수 있도록

구성하였다. 본 논문에서는 예측 값의 정확성을 평가하기 위해 평균 절대 에러(MAE: Mean Absolute Error)를 사용하였으며 아래의 식에 나타낸 것과 같이 구할 수 있다.

$$E = \frac{\sum_{i=1}^{N} | \xi_i |}{N}$$

수강평가 관련 설문조사서

본 설문은 무기명으로 하며 수강 평가가 필요한 관련 논문에 자료로 활용할 것입니다.
참여에 진심으로 감사드립니다.
설문 방법 >> 6-4학년 대상 과목에 해당하는 설문조사입니다.
해당 과목의 점수를 (1-5점 : ① 별로였다~⑤ 아주 괜찮았다.) 체크.
과목에 해당하는 교수님이 바뀌셨어도 자신이 들었던 때에 해당하는 강의 평가를 해 주시면 되겠습니다.
설문조사자>>

과목명	점수	평가 이유(안 써도 됨)
소프트웨어공학	1	그룹별 프로젝트가 비중심화.
컴퓨터조직론	5	실습이 도움이 되었다.
운영체제	5	
알고리즘	4	
컴파일러		안 들었음
데이터통신	4	
시스템프로그래밍	5	
데이터베이스	8	
컴퓨터네트워크		안 들었음
인공지능		안 들었음
자료구조론	4	
무선인터넷		안 들었음

설문에 응해주셔서 감사합니다 ^^

(그림 2) 수강 과목 관련 설문조사서 양식

위의 식에서 N은 총 예측 회수이며 ξ_i는 예측 값과 실제 값의 오차를 나타내고, i 는 각 예측 단계를 나타낸다.

이미 데이터베이스에는 매겨진 평가 값이 존재하지만 점수를 평가하지 않은 것으로 가정하고 그 값을 가린 다음 다른 데이터를 이용하여 점수를 예측한다. 이 예측값과 기존에 매겨진 평가값의 차이의 절대값이 정확도 에러이다. 예측을 하기 위해 전체 N개의 데이터 가운데 (N-1)개를 이용하여 남은 1개를 예측하는 one-hold-out cross-validation을 이용하였다.

<표 3> 실험 결과

test data sets	one-hold-out cross-validation error
240 data	0.9521

테스트 결과 위와 같이 0.9521 정도의 값을 나타내므로 피어슨 상관계수를 사용하는 것이 정확도가 우수함을 알 수 있다. 본 실험을 위해 사용된 데이터 집합은 소규모의 학생을 대상으로 설문 조사를 통해 이루어졌다. 실제 시스템을 현장에 적용하여 대용량의 데이터가 수집될 경우 정확도는 더 높아질 것으로 예상한다.

6. 결론 및 향후 연구과제

본 논문에서는 추천 시스템의 예측 능력을 향상시키기 위하여 유사한 선호도를 가지는 사용자들의 평가에 근거하여 항목들 간의 유사도를 구하여 특정 항목에 대한 사용자의 선호도를 예측하여 추천해 주는 기법을 제안하였다. 기존의 추천 시스템 연구는 전자 상거래를 기준으로만 이루어져 왔지만 본 연구에서는 교육과 연관한 추천 시스템의 활용 정도와 정확성을 예측하여 보았다. 사용자와 유사한 선호도를 가지는 이웃을 선정하는 기법으로 K-최대근접 이웃 방법을 적용하여 성능을 시험한 결과 성향이 비슷한 사람끼리 잘 맞아 예측 정확도가 비교적 우수하였다.

실제로 수강 과목에 대한 학생들의 선호도 데이터가 없기 때문에 설문조사에 근거한 소규모 데이터를 기반으로 추천 방법의 정확도를 분석하였다. 성능을 평가하는 데에는 데이터 수 자체가 미약하여 여러 방법을 통한 데이터 수집이 과제로 남았다. 실제 시스템을 운영하여 많은 학생들의 데이터가 축적되어야만 제대로 된 성능 평가가 가능하고 제안된 방식의 유효성을 검증할 수 있을 것으로 기대한다.

현재 대학교의 수강 과목은 전공 필수, 교양 필수와 같이 필수 과목이거나 전공 선택과 같이 좁은 영역에서의 선택인 경우가 많다. 이런 경우에는 수

강 과목의 추천이 큰 의미를 가지지 않을 것이다. 반면 교양 선택이나 자유 선택과 같이 선택의 폭이 넓은 범위를 지닌 경우에는 학생들이 다양한 과목에 대한 정보를 얻기가 힘들다. 이와 같이 선택의 폭이 넓은 경우 또는 과목에 대한 정보가 부족한 경우 수강 과목 추천 시스템이 유효할 것으로 판단된다.

본 연구에서는 사용자의 선호도를 예측할 때 과목에 대한 선호도만을 가지고 수행하였으나 향후에는 과목의 여러 가지 속성(담당 교수명 등)에 대한 선호도를 이용한다면 보다 신뢰성 있고 향상된 결과를 기대할 수 있을 것이다.

참고문헌

[1] 김진상, “협동적 필터링을 위한 동시출현빈도 사용의 제한 피어슨 알고리즘,” 명지대 대학원 석사학위논문, 2002.

[2] 김형일, “협동적 필터링을 위한 데이터 블러링 기법,” 동국대 대학원 석사학위논문, 2004.

[3] 김혜재, “K-최근접 이웃 추천 엔진에서의 벡터 유사도 사용에 대한 실험적 분석”, 한국외국어대 대학원 석사학위논문, 2002.

[4] 오재영, “전자상거래에서 연관규칙을 이용한 추천 시스템 설계,” 명지대 대학원 석사학위논문, 2004.

[5] 박지선, “A Predictive Algorithm Using 2-way Collaborative Filtering for Recommender Systems,” 연세대학교 컴퓨터과학과 석사학위논문, 2000.

[6] N. Belkin, B. Croft, "Information filtering and information retrieval: two sides of the same coin?," Communications of the ACM, Vol.35, No.2, 1992.

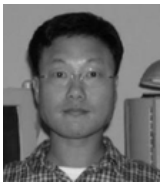
[7] Badrul M. Sarwar, George Karypis, Joseph A.Konstan, John T. Riedle, Application of Dimensionality Reduction in Recommender System-A Case Study, ACM WebKDD 2000 Web Mining for E-Commerce Workshop, 2000.

[8] W. Hill, L. Stead, M. Rosenstein and G. Furnas, "Recommending and Evaluation Choices in a virtual Community of Use," In Proceedings of CHI'95, 1995.

[9] B. Sheth, A learning approach to personalized information filtering, MIT, 1994.

저자소개

손 기 락



1984 서울대학교 계산통계학과 학사

1986 서울대학교 계산통계학과 석사

1993 미국, Univ. of California, Santa Cruz, 전산학 박사

1994~1996년 전자통신연구원 선임연구원

1996~현재 한국외국어대학교 컴퓨터및정보통신공학부 교수

<관심분야> 데이터베이스, 데이터마이닝

김 소 현

2004 덕성여자대학교 컴퓨터과학부 인터넷정보공학과 학사

2006 한국외국어대학교 교육대학원 전자계산교육학과 석사

<관심분야> 컴퓨터교육, 데이터마이닝