내용 기반 협력적 여과 시스템에서 사용자 프로파일을 이용한 자동 선호도 평가

(Automatic Preference Rating using User Profile in Content-based Collaborative Filtering System)

고 수 정 † 최 성 용 †† 임 기 욱 ††† 이 정 현 †††
(Su-Jeong Ko) (Sung-Yong Choi) (Kee-Wook Rim) (Jung-Hyun Lee)

요 약 협력적 여과 시스템은 {사용자-문서}의 행렬을 기반으로 사용자에게 웹 문서를 추천하는 데 있어서 효율적인 시스템이다. 그러나 협력적 여과 시스템은 초기 평가 문제와 희박성으로 인하여 추천의 정확도가 저하된다는 단점을 갖는다. 본 논문에서는 협력적 여과 시스템의 희박성과 초기 평가 문제를 해결하기 위하여 사용자 프로파일을 생성시킴으로써 자동으로 선호도를 평가하는 방법을 제안한다. 본 논문에서 사용하는 프로파일은 협력적 여과 시스템에서의 {사용자-문서} 행렬을 기반으로 생성된 사용자 프로파일에 내용 기반 여과 시스템에서 연관 피드백을 이용하여 생성한 사용자 프로파일을 상호정보의 방법에 의해 병합함으로써 생성한 내용 기반 협력적 사용자 프로파일이다. 생성한 내용 기반 협력적 사용자 프로파일을 정규화시키고, 정규화한 프로파일을 협력적 여과 시스템의 {사용자-문서} 행렬에 반영함으로써 자동으로 선호도를 평가한다. 제안된 방법은 사용자가 웹 문서에 대해서 선호도를 평가한 데이타베이스에서 평가되었으며, 기존의 방법보다 보다 효율적임을 증명한다.

키워드 : 협력적 사용자 프로파일, 자동 선호도 평가, 내용 기반 여과 시스템

Abstract Collaborative filtering systems based on {user-document} matrix are effective in recommending web documents to user. But they have a shortcoming of decreasing the accuracy of recommendations by the first rater problem and the sparsity. This paper proposes the automatic preference rating method that generates user profile to solve the shortcoming. The profile in this paper is content-based collaborative user profile. The content-based collaborative user profile is generated by combining a content-based user profile with a collaborative user profile by mutual information method. Collaborative user profile is based on {user-document} matrix in collaborative filtering system, thus, content-based user profile is generated by relevance feedback in content-based filtering systems. After normalizing combined content-based collaborative user profiles, it automatically rates user preference by reflecting normalized profile in {user-document} matrix of collaborative filtering systems. We evaluated our method on a large database of user ratings for web document and it was certified that was more efficient than existent methods.

Key words : collaborative user profile, automatic preference rating, content-based filtering system

1. 서 론

협력적 여과 시스템에서 입력 데이타는 사용자와 문서의 2차원 행렬로 구성되는데 행(Row)은 사용자, 열(Column)은 문서 목록, 행렬의 값은 사용자가 문서에 대해 평가한 선호도를 나타낸다[1]. 이러한 협력적 여과 시스템은 웹 문서의 내용을 전혀 고려하지 않고, 단지 웹 문서에 대한 사용자의 흥미 정도에만 중점을 두고 추천을 하므로 초기 평가 문제의 한계점을 갖는다. 또한, 많은 수의 웹 문서를 대상으로 사용자가 선호도를 평가하였을 경우, {사용자-문서} 행렬의 희박성으로 인하여 예측의 정확도가 저하되는 단점도 갖는다[2,3]. 협력적 여과 시스템에서의 {사용자-문서} 행렬이 희박해

† 비 회 원 : Colorado State University
sjko@uiuc.edu
†† 비 회 원 : 인하대학교 전자계산공학과
sychoi@nlsun.inha.ac.kr
††† 종신회원 : 선문대학교 산업공학과 교수
rim@omega.sunmoon.ac.kr
†††† 종신회원 : 인하대학교 컴퓨터공학부 교수
jhlee@inha.ac.kr
논문접수 : 2002년 11월 20일
심사완료 : 2003년 12월 26일

지는 또 한가지의 이유는 사용자가 웹 문서에 대해 전체적으로 평가한 것이 아니라 부분적으로 평가함으로 인해 발생하는 결측치 때문이기도 하다[4]. 이러한 한계점을 해결하기 위하여 협력적 여과 시스템과 내용 기반 여과 시스템을 병합하는 여러 연구가 있다[5-14]. LSI[10]와 SVD[13]에서는 희박성 문제를 해결하기 위하여 행렬의 차원 수를 감소시키는 방법을 사용하였으나 초기 평가 문제는 해결하지 못하였다. [5,7-9,11]에서는 초기 평가 문제를 해결하였으나 희박성 문제는 해결하지 못하였다. [2,6,12]에서는 초기 평가 문제와 희박성 문제를 모두 해결하려고는 하였으나 한계를 나타냈으며, 결측치를 해결하지 못하는 단점도 갖는다. [14]에서는 내용 기반 사용자 프로파일을 기반으로 초기 평가 문제를 해결하기 위하여 문서의 내용으로부터 특징을 추출하여 프로파일을 생성하는 방법을 사용한다. 이 방법은 문서의 내용을 추천에 반영함에 따라 초기 평가 문제로 인한 문제점은 해결할 수 있었으나 결측치 문제와 단어 의미 중의성 문제를 해결하지 못함으로 인하여 추천의 정확도에 있어 커다란 장점을 나타내지 못하였다. 본 논문에서는 기존의 방법에서와 같이 문서에 대한 특징을 추출하고 이를 기반으로 생성한 내용 기반 프로파일만을 사용하는 방법이 아니고, 협력적 사용자 프로파일과 내용 기반 사용자 프로파일을 병합하고, 그 결과를 {사용자-문서} 행렬에 반영함으로써 자동으로 선호도를 평가하는 방법을 제안한다. 여기에서, "협력적 사용자"는 협력적 여과 시스템에 나타나는 사용자를 의미하며, "내용 기반 사용자"는 내용 기반 여과 시스템에 나타나는 사용자를 나타낸다. 프로파일을 생성하기 위한 전처리로 협력적 여과 시스템과 내용 기반 여과 시스템에 나타나는 문서로부터 특징을 추출한다. 특징 추출을 위한 방법으로는 대표 단어를 갖는 연관 단어 모델을 이용한다[16]. 추출된 특징을 기반으로 협력적 사용자 프로파일과 내용 기반 사용자 프로파일을 생성하고, 상호정보를 이용하여 이들을 병합한다. 병합한 결과를 {사용자-문서} 행렬의 형식에 맞게 반영함으로써 결측치로 인해 발생하는 협력적 여과 시스템의 희박성과 초기 평가 문제를 해결한다.

2. 협력적 여과 시스템에서의 자동 선호도 평가를 위한 전체 구성도

그림 1은 프로파일을 이용한 협력적 여과 시스템에서의 자동 선호도 평가를 위한 전체 구성도를 나타낸다. 그림 1은 Block1으로부터 Block4로 구성한다.

Block1은 내용 기반 사용자가 검색한 문서를 기반으로 내용 기반 사용자 프로파일을 생성하는 단계이다. 이를 위하여, 내용 기반 사용자가 연관 피드백으로 검색한

그림 1 자동 선호도 평가를 위한 전체 구성도

문서로부터 대표 단어를 갖는 연관 단어 마이닝의 방법에 의해 특징을 추출하며, 그 결과를 기반으로 내용 기반 사용자 프로파일을 생성한다. 또한, 추출한 특징의 수와 비례하도록 가중치가 부여된 내용 기반 사용자 프로파일을 생성하여, C_BU_i로 정의한다.

Block2는 협력적 여과 시스템의 {사용자-문서} 행렬을 기반으로 협력적 여과 사용자 프로파일을 생성하는 단계이다. {사용자-문서} 행렬을 $R=\{r_{ij}\}$로 정의하고, 행렬 R에 속한 모든 문서들로부터 대표 단어를 갖는 연관 단어 마이닝의 방법에 의해 특징을 추출한다. 협력적 사용자 프로파일은 사용자가 선호도를 평가한 문서로부터 추출한 연관 단어로 구성한다. 생성한 협력적 사용자 프로파일의 연관 단어에 가중치를 부여하기 위하여 사용자가 평가한 초기의 선호도를 초기 가중치로 설정하고, 초기 가중치에 연관 단어의 빈도를 곱함으로써 최종 가중치를 부여한다. 이와 같은 절차에 따라 생성한 가중치가 부여된 협력적 사용자 프로파일을 CU_i로 정의한다.

Block3은 가중치가 부여된 내용 기반 사용자 프로파일 C_BU_i와 가중치가 부여된 협력적 사용자 프로파일 CU_i를 병합하고, 병합한 결과를 협력적 사용자 프로파일에 반영하여 내용 기반 협력적 사용자 프로파일 C_CU_i을 생성하는 단계이다. 또한, 자동 선호도 평가를 위해 생성한 내용 기반 협력적 사용자 프로파일을 정규

화한다.

Block4는 자동 선호도 평가를 위해 {사용자-문서} 행렬의 선호도를 수정하는 단계이다. Block3에서 정규화시킨 내용 기반 협력적 사용자 프로파일의 정보를 기반으로 {사용자-문서} 행렬의 선호도를 수정하고 보완한다. 보완된 {사용자-문서} 행렬의 선호도는 {사용자-문서} 행렬의 형식에 부합하지 않으므로, {사용자-문서} 행렬의 형식에 맞는 6단계로 변화시킨다. 그 결과, 초기에 평가되지 않았던 {사용자-문서} 행렬의 선호도를 자동으로 평가할 수 있다.

3. 가중치가 부여된 사용자 프로파일의 생성

협력적 사용자 프로파일을 기반으로 자동으로 선호도를 평가하기 위해서, 먼저 내용 기반 사용자 프로파일과 협력적 사용자 프로파일을 생성한다. 내용 기반 사용자 프로파일을 생성하는 것은 협력적 여과 시스템의 초기 평가 문제를 해결하고, 자동으로 평가한 선호도의 정확도를 높이기 위한 목적이다. 협력적 사용자 프로파일은 {사용자-문서} 행렬을 기반으로 생성한다. 이를 위하여 문서에 대한 특징 추출이 우선되어야 한다.

3.1 문서의 특징 표현

본 논문에서는 문서 표현 형태로서 단일 단어 벡터 모델[12]의 형태를 채택한다. 단일 단어 벡터 모델의 형태는 문서로부터 추출한 단어가 p개 있을 경우, 문서를 p차원의 벡터로 표현한다. 벡터 모델은 단어와 단어가 독립적이라는 것을 전제 조건으로 정한다. 단일 단어 벡터 모델의 구성 요소인 단어는 문서로부터 추출한 명사이다. 단일 단어 벡터 모델은 단어와 단어가 독립적이므로 문서내의 단어 순서와 문서 구조는 고려되지 않는다. 이러한 단점을 보완하기 위해 본 논문에서는 문서의 특징을 단일 단어 벡터 모델의 형태를 응용한 연관 단어 벡터 모델의 형태로 표현한다. 문서의 특징을 연관 단어 벡터 모델의 형태로 표현하기 위해 Apriori 알고리즘을 사용한다.

연관 단어 벡터 모델의 특징을 이용하는 방법은 문서를 연관 단어 집합으로 표현하므로 단어의 중의성 문제로 인한 사용자의 혼란을 방지할 수 있으며, 문서를 보다 상세하게 표현할 수 있다. 또한, 이 방법은 재현율의 장점도 갖는다. 그러나 문서의 특징이 단일 단어를 기준으로 하는 것이 아니고 단어의 집합인 연관 단어를 기준으로 함으로 인하여 정확도면에서 단점을 나타낸다. 즉, 사용자가 문서에 대한 선호도를 평가할 경우, 서로 다른 문서임에도 불구하고 그들을 같은 문서로 판단할 수도 있고, 서로 비슷한 문서임에도 불구하고 사용자가 다른 도메인으로 판단하여 일률적이지 않은 평가를 할 수 있다는 점이다. 또 하나의 단점은 새로운 문서가 추가될 경우 데이타베이스를 갱신해야 하며, 특징 추출을 위해 기존에 저장된 내용을 모두 검색해야 한다는 것이다.

본 논문에서는 이와 같은 정확도의 단점과 새로운 문서 추가시의 문제점을 해결하기 위하여 역문헌빈도를 이용하여 연관 단어의 대표 단어를 지정하는 방법을 사용한다. 역문헌빈도를 이용하여 문서로부터 특징을 추출하고, 역문헌빈도를 연관 단어에 속한 각 단어의 가중치로 지정한다. 다음으로, 각 연관 단어에 속한 단어 중에서 가중치가 가장 큰 단어를 대표 단어로 지정한다. 결과적으로, 문서는 역문헌빈도라는 또 다른 방법으로 추출한 특징을 포함하는 것이다. 이에 따라 연관 단어만를 특징으로 사용함에 따라 나타나는 문제점들을 해결할 수 있다. 문서의 특징을 표현하는 방법을 보다 구체적으로 설명하면 다음과 같다.

Apriori 알고리즘은 형태소 분석[17]에 의해 추출된 명사들로부터 연관 단어를 마이닝한다. 마이닝한 결과를 이용하여 각 문서를 연관 단어들의 집합, 즉 연관 단어 벡터 모델로 나타낸다. 표 1은 연관 단어 마이닝에 의해 추출한 웹문서의 특징을 나타낸다.

표 1 웹문서로부터 추출된 특징의 예

웹문서	문서의 특징
웹문서1	데이타&암호&통신망 게임&설명&제공&공략 게임&이용&기술&개발 삭제&게임&개인전&경고
웹문서2	국내&최신&기술&설치 게임&순위&이름&스포츠 게임&일정&선수&참가 **위원회&선수&선발**

표 1과 같이 추출한 연관 단어에 대한 대표 단어를 선정하기 위하여 역문헌빈도를 이용한다. 이를 위한 전처리로 문서를 형태소 분석하고, 그 결과 중에서 명사만을 추출한다. 추출된 모든 명사의 역문헌빈도[14]를 계산하기 위해 식 (1)을 이용한다.

$$W_{nk} = f_{nk} \bullet [log_2 \frac{n}{DF} + 1] \quad (1)$$

fn_k는 문서 내 모든 단어에 대한 단어 n_k의 상대빈도이며, n은 학습문서의 수이며 DF는 학습문서에서 단어 n_k가 나타난 문서의 수를 의미한다. 역문헌빈도가 높은 단어부터 낮은 단어로 정렬하여 상위의 빈도를 나타내는 단어만을 문서의 특징으로 추출한다. 이와 같은 방법으로 추출한 실험문서 $d1$의 특징은 $\{n_1, n_2, \cdots, n_k, \cdots, n_m\}$이다. 특징에 대한 가중치는 각 단어의 역문헌빈도로 정의하고, $\{W_{n1}, W_{n2}, \cdots, W_{nk}, \cdots, W_{nm}\}$라고 나타낸다.

표 1과 같이 연관 단어 마이닝의 방법에 의해 추출한 특징과 연문헌빈도를 이용하여 추출한 $\{n_1,n_2,\cdots,n_k,..,n_m\}$와의 비교를 통하여 동일한 단어에 대하여 동일한 가중치를 부여한다. 따라서 연관 단어에 속한 단어의 각각에 가중치가 부여된다. 연관 단어에 나타난 단어 중에서 가중치가 가장 큰 단어가 그 연관 단어의 대표 단어이며, 이를 강조함으로써 대표 단어임을 나타낸다. 예를 들어, 표 1에 나타난 연관 단어 중에서 '(국내&최신&기술&설치)'의 대표 단어가 '기술'이라고 하였을 경우, '기술'이라는 단어가 대표 단어임을 (국내&최신&*기술*&설치)와 같이 명시한다. 따라서, 사용자가 연관 단어와 대표 단어를 기반으로 선호도를 평가할 수 있다. 결과적으로, 서로 다른 웹 문서일지라도 같은 평가를 하는 문제점을 해결할 수 있다.

식 (2)는 p개의 연관 단어로 구성된 문서 d_j의 특징을 정의하는 식이다.

$$d_j=\{AW_{j1}, AW_{j2},\cdots, AW_{jk}, \cdots, AW_{jp}\} \quad (2)$$

식 (2)에서 AW_{j1} , AW_{j2}, AW_{jk}, AW_{jp}는 각각 문서 d_j로부터 추출된 표 1과 같은 형태의 연관 단어를 나타낸다. 예를 들어, 표 1에 나타난 웹문서1의 '*데이타*&암호&통신망'의 연관 단어는 AW_{j1}에 해당한다.

식 (2)에서 연관 단어를 가장 적절하게 추출하기 위해서 신뢰도는 85보다 크도록 지지도는 22보다 작도록 지정해야 한다[16].

3.2 내용 기반 사용자 프로파일

대부분의 사용자는 유사한 문서 혹은 동일한 문서를 연속적으로 검색하는 습관을 갖는 특성이 있다. 따라서, 이러한 특성을 이용하여 사용자가 연속적으로 검색한 문서로부터 특징을 추출하여 내용 기반 사용자 프로파일을 생성한다. 이와 같이 생성한 내용 기반 사용자 프로파일은 연관 피드백에 의한 결과로 주기적으로 수정된다[20].

내용 기반 사용자 프로파일은 3.1절과 같은 방법에 의해 추출된 특징으로 구성된다. 정확도를 높이기 위해 문서에 대한 가중치를 내용 기반 사용자 프로파일에 추가한다. 가중치는 사용자에 의해 검색된 문서로부터 추출된 모든 연관 단어에 대한 각 연관 단어의 비율로 정의한다.

본 논문에서는 식 (2)와 같은 형태의 연관 단어를 기반으로 내용 기반 사용자의 프로파일을 구성한다. 표 2는 i번째 내용 기반 사용자 프로파일(C_BU_i)의 구조를 나타낸다. 즉, 내용 기반 사용자 프로파일은 가중치가 부여된 연관 단어의 집합으로 구성된다. 표 2에서 $\{w_{i1},w_{i2},\cdots,w_{it}\}$는 각 연관 단어의 가중치를 나타내며, t는 사용자 프로파일에 포함되어 있는 모든 종류의 연관 단어의 수이다. $\{AW_{i1}$, $AW_{i2},\cdots$, $AW_{ik},\cdots$, $Aw_{it}\}$는 식 (2)와 같은 형태로써 문서로부터 추출한 연관 단어이다.

표 2 내용 기반 사용자 프로파일의 구조

User ID	가중치	연관 단어	...	가중치	연관 단어	...	가중치	연관 단어
C_Bu_i (사용자: c_bu_i, $1\le j\le p$)	w_{i1}	AW_{i1}		w_{ij}	AW_{ij}	...	w_{ip}	AW_{it}

3.3 협력적 사용자 프로파일

웹 문서 기반의 협력적 여과 시스템은 {사용자-문서} 행렬을 기반으로 사용자에게 문서를 추천한다. 협력적 여과 시스템에서의 사용자는 모든 문서에 대해 선호도를 평가하지 않는다. 따라서 {사용자-문서} 행렬에 결측치가 발생된다. 이러한 결측치는 {사용자-문서} 행렬을 더욱 희박하게 만드는 원인이 된다. 본 절에서는 결측치로 인한 {사용자-문서} 행렬의 희박성을 줄이기 위한 전처리로 협력적 사용자 프로파일을 생성시키는 방법을 기술한다.

3.3.1 {사용자-문서} 행렬의 구성

p개의 특징 벡터로 구성된 m개의 문서와 n명의 사용자 집합을 정의할 경우, 사용자 집합은 $U=\{cu_i\}(i=1,2,\cdots,n)$로 정의하고, 문서의 집합은 $I=\{d_j\}(j=1,2,\cdots,m)$로 정의한다. $R=\{r_{ij}\}(i=1,2,\cdots,n\ \ j=1,2,\cdots,m)$는 {사용자-문서}의 행렬이다. 행렬의 요소 r_{ij}는 문서 d_j에 대한 사용자 cu_i의 선호도를 나타낸다. 표 3은 협력적 여과에 대한 {사용자-문서}의 행렬을 보인다.

표 3 협력적 여과 시스템에서 {사용자-문서} 행렬

	d_1	d_2	d_3	d_4	...	d_j	...	d_m
cu_1	r_{11}	r_{12}	r_{13}	r_{14}	...	r_{1j}	...	r_{1m}
cu_2	r_{21}	r_{22}	r_{23}	r_{24}	...	r_{2j}	...	r_{2m}
...	...	...	...	...	...	...	...	...
cu_i	r_{i1}	r_{i2}	r_{i3}	r_{i4}	...	r_{ij}	...	r_{im}
...	...	...	...	...	...	...	...	...
cu_n	r_{n1}	r_{n2}	r_{n3}	r_{n4}	...	r_{nj}	...	r_{nm}

웹 문서 추천을 위한 협력적 여과 시스템에서의 사용자는 문서에 대한 선호도의 정도를 평가한다. 선호도의 정도는 0~1.0까지 0.2씩 증가하면서 총 6단계로 구분한다. 6단계의 선호도 단계 중에서 0.5보다 큰 선호도로 평가를 받은 문서는 사용자에게 흥미로운 문서라고 평가한다. 본 논문에서 사용하는 웹 문서는 웹 문서 수집기에 의해 수집된 컴퓨터에 관련된 문서이다. 웹 문서의 특징은 3.1절에서 기술된 방법인 대표 단어를 갖는 연관 단어 마이닝의 방법을 사용하여 추출한다. 표 3에서 r_{ij}는 식 (3)과 같은 형태로 정의한다. 즉 행렬의 요소 r_{ij}

는 선호도의 6단계와 전혀 평가를 하지 않은 경우 중 하나에 속한다.

$$r_{ij} \in \{\phi, 0, 0.2, 0.4, 0.6, 0.8, 1\} (i = 1,2,...,n)(j = 1,2,...,m) \quad (3)$$

식 (3)에서 ϕ는 협력적 여과 사용자 i가 문서 j에 대해 평가를 하지 않았음을 의미한다.

표 4는 협력적 여과 추천 시스템에서 웹 문서에 대해 사용자가 평가한 선호도의 예를 나타낸다. 문서의 특징은 3.1절의 방법에 의해 추출한 대표 단어를 갖는 연관 단어이다. 표 4에서 '?'의 의미는 선호도를 자동으로 평가해야 하는 부분임을 나타낸다. d_j는 식 (2)의 정의에 의하여 연관 단어 벡터 모델로 표현한다.

표 4 웹 문서에 대해 사용자가 평가한 선호도의 예

	d_1	···	d_j	···	d_m
cu_1	0.2	···	1	···	0.4
cu_2	?	···	0.8	···	0.6
cu_3	0.4	···	0.6	···	?
···	···	···	···	···	···
cu_n	0.4	···	?	···	?

3.3.2 협력적 사용자 프로파일의 생성

협력적 사용자 cu_i의 프로파일은 협력적 사용자가 선호도를 평가한 문서의 특징을 기반으로 생성한다. 협력적 사용자가 선호도를 낮게 평가하였을 경우 평가된 문서의 특징에 대한 가중치는 낮게 정의하고, 선호도를 높게 평가하였을 경우 특징에 대한 가중치는 높게 정의한다. 따라서 문서의 특징으로 표현한 연관 단어의 선호도는 가중치에 따라 변화된 값으로 표현된다. 협력적 사용자 cu_i가 문서 d_j에 대해 평가한 선호도를 r_{ij}로 정의한 것과 같이, 문서 d_j의 특징으로 식 (2)와 같이 추출된 p개의 연관 단어 중에서 k번째 연관 단어 AW_{ijk}의 초기 가중치는 r_{ij}의 값으로 부여한다. 여기서, 연관 단어 AW_{ijk}의 가중치는 c_w_{ijk}로 정의한다. 식 (4)는 협력적 사용자 cu_i의 사용자 프로파일을 생성하기 위하여, c_w_{ijk}에 초기 가중치를 부여한다. 초기 가중치는 최초로 부여된 가중치를 의미한다. AW_{ijk}의 초기 가중치는 연관 단어가 속하는 상품에 대해 평가한 선호도 r_{ij}의 값을 사용한다. 사용자가 직접 평가한 선호도는 평가되지 않은 상품의 선호도를 예측하기 위한 가장 정확하고도 중요한 자료이기 때문이다.

$$c\_w_{ijk} = Preference(AW_{ijk}) = r_{ij}$$
$$(\text{사용자 } cu_i,\ 1 \le j \le m,\ 1 \le k \le p) \quad (4)$$

표 5는 식 (4)의 정의에 의해 생성한 연관 단어의 초기 가중치 c_w_{ijk}를 계산하는 방법을 구체적으로 보인다. 표 5에 나타난 사용자 cu_i는 d_1, d_j, d_m 문서에 대하여 각각 0.2, 0.8, 1로 선호도를 평가한 하였다. 연관 단

표 5 cu_i의 프로파일 생성을 위한 초기 가중치

문 서	초기 가중치	연관 단어
d_1 (선호도 r_{i1}=0.2)	c_w_{i11} (0.2)	AW_{i11} **게임**&구성&선수&선발
	c_w_{i12} (0.2)	AW_{i12} 국내&최신&**기술**&설치
	···	···
	c_w_{i1k} (0.2)	AW_{i1k} 그림&인기&**서비스**&음악
	···	···
	c_w_{i1p} (0.2)	AW_{i1p} **이용**&기술&개발
d_j (선호도 r_{ij}=0.8)	c_w_{ij1} (0.8)	AW_{ij1} **이용**&기술&개발
	c_w_{ij2} (0.8)	AW_{ij2} **게임**&구성&선발&순위
	···	···
	c_w_{ijk} (0.8)	AW_{ijk} 국내&최신&**기술**&설치
	···	···
	c_w_{ijp} (0.8)	AW_{ijp} 구성&선수&**순위**
d_m (선호도 r_{im}=1.0)	c_w_{im1} (1.0)	AW_{im1} 제공&일러스트&**설명**
	c_w_{im2} (1.0)	AW_{im2} **이용**&기술&개발
	···	···
	c_w_{imk} (1.0)	AW_{imk} **게임**&순위&스포츠
	···	···
	c_w_{imp} (1.0)	AW_{imp} 그림&**데이타**&서비스&엔진

어에 포함된 각 단어 중에서 강조된 단어는 대표 단어임을 나타낸다.

표 5에서는 식 (4)의 정의에 따라 $\{AW_{i11},\cdots\}$,$\{AW_{ij1},\cdots\}$,$\{AW_{im1},\cdots\}$ 등의 연관 단어에 초기 가중치가 각각 0.2, 0.8, 1로 정의됨을 볼 수 있다.

반면, 표 5의 AW_{ij1} ,AW_{im2}, AW_{i1p}의 연관 단어는 같은 연관 단어이나 초기 가중치가 각각 0.2, 0.8, 1로 다르게 지정됨을 보인다. 협력적 사용자의 프로파일을 구성하기 위해서는 이들의 정보를 모두 반영해야 하므로 다르게 지정된 가중치를 병합시키는 과정이 필요하다. 이를 위하여, 연관 단어에 대한 가중치를 협력적 사용자 cu_i가 선호도를 평가한 모든 문서로부터 추출한 전체 연관 단어 집합을 검색하여 같은 연관 단어가 나타날 때마다 가중치에 곱한다. 따라서, 사용자가 문서에 대해 평가한 선호도에 따라 다르게 정의된 연관 단어의 가중치를 단일 가중치로 정의할 수 있다. 표 6은 표 5와 같이 부여된 연관 단어의 초기 가중치를 이와 같은 원리에 의하여 병합시키는 구체적인 방법과 예를 보인다.

예를 들어, {AW_{ij1} ,AW_{im2}, AW_{ilp}}와 같이 {AW_{il2}, AW_{ijk}}의 연관 단어도 같은 연관 단어이므로 이들의 최종 가중치 $c\_w'_{il2}$, $c\_w'_{ijk}$ 는 각각의 초기 가중치를 서로 곱한 값인 c_w_{il2}xc_w_{ijk}로 정의된다. 반면, 표 6에 나타나지 않은 AW_{il1} AW_{ilk}… 등의 연관 단어는 같은 연관 단어를 갖지 않으므로 이들의 최종 가중치 $c\_w'_{il1}$, $c\_w'_{ilk}$,…는 초기 가중치 c_w_{il1}, c_w_{ilk}, …와 같다.

표 6 연관 단어에 부여된 최종 가중치

연관 단어	연관 단어에 대한 가중치
AW_{ij1}, AW_{im2}, AW_{ilp}	$c\_w'_{ij1}$<=c_w_{ij1}xc_w_{im2}xc_w_{ilp} $c\_w'_{im2}$<=c_w_{ij1}xc_w_{im2}xc_w_{ilp} $c\_w'_{ilp}$<=c_w_{ij1}xc_w_{im2}xc_w_{ilp}
AW_{il2}, AW_{ijk}	$c\_w'_{il2}$<=c_w_{il2} xc_w_{ijk} $c\_w'_{ijk}$<=c_w_{il2} xc_w_{ijk}

식 (5)는 표 6의 원리를 기반으로 정의한 식이다. 즉, 동일한 연관 단어이나 각기 다른 값으로 정의된 연관 단어의 가중치를 병합함으로써 동일하게 가중치를 부여하기 위한 식이다. 이를 위하여, 식 (5)는 협력적 사용자 cu_i가 평가한 문서로부터 추출한 모든 연관 단어를 데이타베이스($AWDB$)에 저장하고, 모든 연관 단어(AW_{ijk})를 검색하여 같은 연관 단어($Aw_{ij'k'}$)를 찾을 때마다 이의 선호도를 곱하고, 그 결과를 $c_w'_{ijk}$로 정의한다. 식 (6)에서 $j\neq j'$ or $k\neq k'$은 동일 연관 단어, 즉 AW_{111}=AW_{111}과 같은 형태의 비교를 제외시키는 의미이다.

$$c\_w'_{ijk} = \prod_{AW_{ijk},AW_{ij'k'}\in AWDB} c\_w_{ijk}\cdot c\_w_{ij'k'} \mid (AW_{ijk} = AW_{ij'k'}) \quad (5)$$

$(1\leq j, j'\leq m,\ 1\leq k, k'\leq p) \mid j\neq j or k\neq k', 사용자:cu_i,$

식 (5)에 따라 본 논문에서는 협력적 사용자 cu_i의 프로파일 CU_i의 구조를 표 7과 같이 정의한다. 표 7은 협력적 사용자 cu_i가 선호도를 평가한 모든 문서로부터 추출한 연관 단어 집합에서 중복된 연관 단어를 제외한 연관 단어 AW_{ijk}에 가중치 $c_w'_{ijk}$를 부여함으로써 정의한다.

표 7 협력적 사용자 프로파일 CU_i의 구조

User ID	가중치	연관 단어	…	가중치	연관 단어	…	가중치	연관 단어
CU_i ($사용자:cu_i$ $1\leq j\leq m$, $1\leq k\leq p$)	$c\_w'_{il1}$	AW_{il1}	…	$c\_w'_{ijk}$	AW_{ijk}	…	$c\_w'_{imp}$	AW_{imp}

4. 프로파일의 병합을 이용한 자동 선호도 평가

3장에서는 내용 기반 사용자 프로파일과 협력적 사용자 프로파일을 생성하였다. 본 장에서는 내용 기반 사용자 프로파일과 협력적 사용자 프로파일을 병합하고, 그 결과를 {사용자-문서}의 행렬에 반영함으로써 자동으로 사용자의 선호도를 평가하는 방법을 기술한다.

4.1 협력적 사용자와 가장 유사도가 높은 내용 기반 사용자 검색

협력적 사용자와 가장 유사도가 높은 내용 기반 사용자를 검색하기 위하여 협력적 사용자 프로파일과 내용 기반 사용자 프로파일을 병합한다. 프로파일을 병합하기 위하여 프로파일 간의 유사도를 계산한다. 프로파일 간의 유사도를 평가하는 목적으로 사용하는 방법은 대량의 상품간의 비교에 있어서 많이 사용되는 상호정보를 이용한 방법[21]이다. 본 논문에서도 상호정보의 방법을 이용하여 협력적 사용자 프로파일과 내용 기반 사용자 프로파일 간의 유사도를 구한다. 정보 이론에서 상호정보은 연관 확률 분포인 $P(x)$, $P(y)$의 값을 갖는 X와 Y 간의 통계적인 의존도를 측정한다. 이와 같은 원리에 의하여 표 7의 i'번째의 협력적 사용자 프로파일 $CU_{i'}$와 표 2의 i번째의 내용 기반 사용자 프로파일 C_BU_i 간의 유사도를 식 (6)에 의해 정의한다. 식 (6)에서 $P(AW_{ij})$, $P(AW_{i'jk})$는 각각 AW_{ij}, $AW_{i'jk}$의 출현빈도를 나타낸다. 또한, $P(AW_{ij}, AW_{i'jk})$는 AW_{ij}, $AW_{i'jk}$가 동시에 출현한 빈도를 나타낸다.

$$I(C\_BU_i, CU_{i'}) = \sum_{j=1}^{t}\sum_{k=1}^{p} P(AW_{ij}, AW_{i'jk}) log \frac{P(AW_{ij}, AW_{i'jk})}{P(AW_{ij})P(AW_{i'jk})} \quad (6)$$

식 (6)에 의해서 협력적 사용자 프로파일과 내용 기반 사용자의 유사도를 계산할 수 있다. 협력적 사용자는 식 (6)에 의한 결과로부터 가장 높은 값을 나타내는 내용 기반 사용자를 프로파일 병합에 대한 대상으로 지정한다.

4.2 프로파일 병합을 통한 내용 기반의 협력적 사용자 프로파일의 생성

본 절에서는 4.1절에서 지정한 프로파일 병합의 대상이 되는 협력적 사용자와 내용 기반 사용자의 프로파일을 병합함으로써 내용 기반의 협력적 사용자 프로파일을 생성하는 방법을 기술한다.

각 프로파일을 병합하기 위하여 협력적 사용자의 가중치를 내용 기반 사용자의 가중치와 합한다. 협력적 사용자의 가중치는 더한 결과값으로 변화한다. 표 8은 표 7과 같이 정의된 협력적 사용자 프로파일의 가중치를 내용 기반 사용자 프로파일의 가중치와 합한 후에 변화된 내용 기반 협력적 사용자 프로파일 C_CU_i로 정의한

표 8 내용 기반 협력적 사용자의 프로파일 C_CU_i의 구조

User ID	가중치	연관 단어	…	가중치	연관 단어	…	가중치	연관 단어
C_CU_i (사용자 cu_i $1 \le j \le m$, $1 \le k \le p$)	$c\_c\_w'_{i11}$	AW_{i11}	…	$c\_c\_w'_{ijk}$	AW_{ijk}	…	$c\_c\_w'_{imp}$	AW_{imp}

구조를 나타낸다. {$c\_c\_w'_{i11}$,…, $c\_c\_w'_{ijk}$,…,$c\_c\_w'_{imp}$}는 내용 기반 사용자 프로파일의 구성 요소인 연관 단어에 대한 가중치와 협력적 사용자 프로파일의 구성 요소인 연관 단어에 대한 가중치를 합한 결과이다.

이와 같이 정의한 내용 기반 협력적 사용자 프로파일을 사용하여 사용자의 선호도를 자동으로 평가한다. 표 8에 나타난 가중치는 협력적 여과 시스템에서 사용하는 식 (3)와 같이 정의한 선호도의 범위에 해당하지 않는다. 따라서 내용 기반 협력적 사용자 프로파일의 가중치를 선호도 자동 평가에 사용하기 위해서는 식 (3)과 같이 0~1사이의 값으로 변화시켜야 한다. 이를 위하여 본 논문에서는 가중치의 값을 0~1사이의 값으로 변화시키기에 편리한 코사인 정규화 방법[15]을 이용한다. 표 8에 의해 표현한 내용 기반 협력적 사용자 프로파일을 내용 기반 협력적 사용자 프로파일 벡터로 표현하고자 할 경우, 유클리디언 길이(Euclidean length)로 나눈다. 식 (7)은 표 8에 의해 표현한 내용 기반 협력적 사용자 프로파일의 가중치를 유클리디언 길이로 정의한 것이다.

$$EC\_CU_i = \sqrt{c\_c\_w'^{\,2}_{i11} + ,,, c\_c\_w'^{\,2}_{ijk} + ,,, c\_c\_w'^{\,2}_{imp}}$$

(cu_i 사용자 $1 \le j \le m$, $1 \le k \le p$) (7)

표 8과 같의 정의한 내용 기반 협력적 사용자 프로파일 C_CU_i의 유클리디언 길이에 대한 각 연관 단어의 가중치를 식 (8)과 같이 정의한다. 즉, 각 연관 단어의 가중치를 유클리디언 길이로 나눈다.

$$c\_c\_w''_{ijk} = \frac{c\_c\_w'_{ijk}}{EC\_CU_i} \quad (8)$$

4.3 자동 선호도 평가와 문서 추천

본 절에서는 식 (8)에 의해 정의된 내용 기반 협력적 사용자 프로파일의 가중치를 기반으로 {사용자-문서} 행렬의 선호도를 수정하고 보완한다. {사용자-문서} 행렬에 나타난 문서로부터 추출한 연관 단어와 내용 기반 협력적 사용자 프로파일의 연관 단어와의 일치를 검사한다. 검사 결과, 두 연관 단어가 같을 경우 선호도를 식 (8)의 값에 곱하고, 다른 경우는 식 (8)의 값을 선호도로 사용함으로써 선호도를 수정하고 보완한다. 표 9는 표 4와 같이 평가한 {사용자-문서} 행렬에 선호도의 수정 및 보완 과정을 실시한 경우의 결과를 나타낸다. 표 9에서는 표 4에서 문제가 되었던 '?'의 결측치가 많이 보완되었음을 볼 수 있다.

표 9 선호도 수정 및 보완으로 변화한 {사용자-문서} 행렬

	d_1	…	d_j	…	d_m
cu_1	0.2	…	1	…	0.4
cu_2	?->0.32	…	0.8	…	0.6
cu_3	0.4	…	0.6	…	?->0.123
…	…	…	…	…	…
cu_n	0.4	…	?->0.242	…	?->0.012

선호도 수정 및 보완 과정을 거친 {사용자-문서} 행렬의 선호도를 r_{ij}'라고 정의한다. 협력적 여과 시스템의 {사용자-문서} 행렬의 선호도는 식 (3)과 같이 0에서 1 사이의 6단계로 표현되어야 한다. 그러나 선호도 수정 및 보완 과정에 따라 보완한 {사용자-문서} 행렬에서의 선호도는 표 9와 같이 0부터 1까지의 값을 가지나 식 (3)과 같은 6단계로 선호도로 구성되지 않음을 볼 수 있다. 이와 같은 선호도는 기존에 있는 협력적 여과 시스템에서의 {사용자-문서} 행렬의 요소와 조화되지 않으므로 추천에 사용할 수 없다. 그러므로 선호도 수정 및 보완 과정을 거쳐 보완된 {사용자-문서} 행렬의 선호도 r_{ij}'는 식 (3)의 정의와 같이 변경되어야 한다. 이를 위하여, 표 10에서는 r_{ij}'를 식 (3)에 정의된 0, 0.2, 0.4, 0.6, 0.8, 1.0의 값 중에서 가장 근접한 값으로 변경하는 방법을 제시한다.

표 10 6단계 중 하나의 단계로 표현한 선호도

프로파일을 기반으로 수정 및 보완된 선호도	{사용자-문서}행렬에서 사용하기 위한 선호도
0 (r_{ij}' (0.1	r_{ij} = 0.0
0.1 (r_{ij}' (0.3	r_{ij} = 0.2
0.3 (r_{ij}' (0.5	r_{ij} = 0.4
0.5 (r_{ij}' (0.7	r_{ij} = 0.6
0.7 (r_{ij}' (0.9	r_{ij} = 0.8
0.9 (r_{ij}' (1.0	r_{ij} = 1.0

{사용자-문서} 행렬은 선호도 자동 평가를 통하여 결측치가 많이 감소되었으며, 선호도가 보다 정확한 값으로 수정 및 보완되는 결과를 얻는다. 이와 같은 결과의 {사용자-문서} 행렬의 선호도를 기반으로 문서를 추천하기 위해서는 협력적 여과 기술에서 가장 대표적인 식 (9)의 피어슨 상관(Pearson Correlation)를 사용한다. 식 (9)에서 가중치 $w(a,i)$는 새로운 새로운 사용자 cu_a와 각 사용자 cu_i와의 유사도 또는 상관 관계를 표현한다. 여기서, $w(a,i)$는 피어슨 상관 계수를 사용함에 의해 구할 수 있다[22].

$$w(a,i)=\frac{\sum_j(r_{aj}-\bar{r}_a)(r_{ij}-\bar{r}_i)}{\sqrt{\sum_j(r_{aj}-\bar{r}_a)^2\sum_j(r_{ij}-\bar{r}_i)^2}} \quad (9)$$

r_{aj}는 새로운 사용자 u_a가 문서 j에 대하여 보여준 선호도이고, 문서j는 새로운 사용자 cu_a와 사용자 cu_i가 공통으로 평가를 한 문서이다. $\bar{r}_a$와 $\bar{r}_j$는 각각 새로운 사용자 cu_a와 기존의 사용자 cu_i의 선호도 평균값이다.

협력적 사용자 cu_a에게 문서를 추천하기 위해서는 식 (9)의 $w(a,i)$의 결과 중 가장 높은 값을 나타내는 사용자 cu_i의 선호도를 기반으로 문서의 예상 선호도를 계산하고, 이 선호도가 일정 임계값 0.6이상이 될 경우 추천한다. 사용자 cu_a의 특정 문서 j에 대한 선호도는 식 (10)을 이용하며, 예상 추천값 p_{aj}는 다음과 같다.

$$p_{aj}=\bar{r}_a+\frac{\sum_{i=1}^{n}w(a,i)(r_{ij}-\bar{r}_i)}{\sum_{i=1}^{n}w(a,i)} \quad (10)$$

p_{aj}는 새로운 사용자 cu_a가 문서j에 대한 예측한 선호도 값이고, $\bar{r}_a$는 새로운 사용자 cu_a의 선호도 평균이다. $w(a,i)$는 새로운 사용자 cu_a와 사용자 cu_i의 유사도 가중치이고, n은 새로운 사용자 cu_a와 다른 사용자들 간의 유사도가 0이 아닌 사용자 수이다.

5. 성능 평가

협력적 여과 추천을 위한 데이타베이스는 200명의 사용자와 1600개의 서로 다른 웹 문서로 구성한다. 사용자는 1600개의 웹 문서에 대해 적어도 10개의 평가를 한 사용자들이다. 내용 여과 기반의 추천을 위한 데이타베이스는 1600개의 서로 다른 웹 문서로 구성한다. 1600개의 웹 문서는 웹 문서 수집기에 의해서 컴퓨터 분야의 URL로부터 수집된 문서이다. 1600개의 훈련문서는 수작업으로 8개의 컴퓨터 분야로 분류한다. 여기서 8개의 클래스는 {게임, 그래픽, 뉴스와 미디어, 반도체, 보안, 인터넷, 전자출판, 하드웨어}의 레이블이다. 8개의 클래스로 분류한 기준은 알타비스타, 야후 등의 기존의 검색 엔진이 컴퓨터 분야의 주제를 대상으로 분류한 통계에 따른 것이다. 실험 문서는 한국어의 기본 문법 체제를 기반으로 하여야 하며, 내용의 구성이 한국어 문장으로 구성되어 있어야 한다. 또한 몇 몇 단어만으로 구성되어 거의 빈문서와 같은 문서는 오류 문서이므로 실험의 대상에서 제외되어야 한다. 200명의 사용자 중 100명의 사용자는 훈련을 위해 사용하며, 나머지 100명은 테스트를 위해 사용한다.

추천의 성능을 평가하기 위해 본 논문에서는 [22]에 의해 제안된 MAE(Mean Absolute Error)와 순위 스코어 측정(Rank scoring metric)을 사용한다. MAE는 단일 문서의 추천 시스템을 평가하는 데 사용하며, 순위 스코어 측정은 순위가 있는 문서의 목록을 추천하는 시스템의 성능을 평가하는 데 사용한다. MAE에서 예측의 정확도는 실제로 사용자가 평가한 값과 예측된 값의 차이에 대한 절대값의 평균을 나타내며 식 (11)에 의해 정의된다.

$$s_a=\frac{1}{m_a}\sum_{j\in p_a}|p_{a,j}-v_{a,j}| \quad (11)$$

식 (11)에서 $p_{a,j}$는 예측된 선호도이며 $v_{a,j}$는 실제로 사용자가 평가한 선호도이다. 또한 m_a는 새로운 사용자에 의해 평가된 문서의 수를 의미한다.

순위 스코어 측정은 순위가 있는 목록에 있는 문서를 사용자가 방문 또는 평가하는 가의 측정이다. 순위 스코어 측정은 문서를 선택할 확률이 목록의 하단으로 갈수록 지수적으로 감소한다는 전제에서 측정된다. 각 문서는 사용자 선호도의 가중치의 값에 따라 내림차순으로 j에 의해 정렬되어 있다고 가정한다. 식 (12)은 순위가 부여된 문서의 목록에 대한 사용자 U_a의 순위 스코어 측정에 대한 기대 이용도(Expected utility)를 계산하기 위한 식이다.

$$R_a=\sum_j\frac{\max(V_{a,j}-d,0)}{2^{(j-1)/(\alpha-1)}} \quad (12)$$

식 (12)에서 d는 문서에 대한 중간 평가값이며, α는 반감기(halflife)이다. 반감기는 사용자가 평가하거나 방문할 50-50의 기회가 있는 목록에 있는 문서의 수이다. 본 논문의 평가에서는 반감기를 5로 사용한다. 식 (13)은 순위 스코어 척도를 사용하여 새로운 사용자에 대한 예측의 정확도를 나타내는 식이다.

$$R=100\times\frac{\sum_u R_u}{\sum_u R_u^{\max}} \quad (13)$$

식 (13)에서 $R_u^{\max}$는 사용자가 평가하거나 방문한 문서가 순위가 있는 목록상에서 상위에 나타났을 경우에 측정된 순위 스코어 측정에 대한 기대 이용도의 최대값이다.

본 논문에서는 평가를 위해 제안된 프로파일을 이용한 자동 선호도 평가 방법(P_P_D), 선호도 예측을 위해 피어슨 상관계수를 이용한 기존의 메모리 기반 방법(P_M)[1], 기존의 협력적 사용자 프로파일만을 이용한 추천 방법(C_P_D)[4,14], 내용 기반 여과 방법만을 이용한 추천 방법(Con)[21]을 군집된 사용자의 수를 변화시키면서 성능을 비교하였다.

또한 제안한 방법을 1장에 기술한 협력적 여과와 내용 기반 여과를 병합한 기존 추천 시스템 중 본 논문에

표 11 내용 기반 여과와 협력적 여과를 병합한 기존의 추천 방법

제 안	방 법	본 논문에서의 표기
내용-프로파일의 행렬에 협력적 여과를 적용 (Pazzani)[6]	-사용자 프로파일을 훈련 집합으로부터 추출된 가중치가 있는 단어의 집합으로 표현 -내용-프로파일 행렬은 여러 사용자들의 프로파일의 모임.	Pazzani
사용자의 선호도를 학습 (Lee)[2]	-비슷한 사용자의 평가뿐 아니라 상품 내용에 대한 사용자의 선호도를 학습	Lee
개인화 정보 여과 에이전트 (Good et al)[9]	-새로운 사용자에 대한 예측은 새로운 사용자의 개인화 에이전트를 협력적 여과에 응용함으로써 이루어짐	Good

표 12 사용자 수의 변화에 따른 MAE와 순위 스코어

사용자수	MAE				Rank scoring			
	P_M	C_P_D	Con	P_P_D	P_M	C_P_D	Con	P_P_D
10	0.280	0.288	0.267	0.260	53	53.2	60.2	61.2
20	0.280	0.271	0.265	0.255	53.1	54.1	61.2	63.2
30	0.276	0.269	0.264	0.242	53.5	55.2	59.3	64.3
40	0.273	0.264	0.261	0.224	54.1	55.9	60.1	65.8
50	0.271	0.250	0.255	0.203	54.2	56.1	61.2	66.9
60	0.266	0.247	0.254	0.198	54.8	57.3	62.3	67.1
70	0.266	0.231	0.255	0.181	55.1	58.7	62.4	67.3
80	0.265	0.228	0.260	0.171	55.2	60.2	62.8	67.9
90	0.265	0.210	0.261	0.163	55.5	62.3	63.1	68.1
100	0.264	0.198	0.259	0.154	56.1	63	63.2	70

서 제안한 방법과 같이 프로파일을 사용하는 방법인 표 11에 나타낸 방법과 사용자가 문서에 대해 평가한 횟수를 변화시켜가면서 비교하였다.

표 12는 식 (11)와 식 (13)를 기반으로 군집된 사용자의 수를 변화시킴에 따른 본 논문에서 제안한 P_P_D, 기존의 방법인 P_M과 C_P_D, 내용 기반 여과 방법만을 이용하는 Con의 MAE와 순위 스코어를 나타낸다.

그림 2와 그림 3은 표 12를 기반으로 한 사용자의 수에 따른 MAE와 순위 측정 척도를 나타낸다. 그림 2와 그림 3은 사용자들의 수가 많아짐에 따라 P_P_D와 C_P_D의 성능은 점차 높아지나 P_M과 Con를 이용한 방법은 사용자들의 수에 크게 의존하지 않음을 보인다. 특히, P_P_D의 성능은 사용자들의 수가 많아짐에 따라 점차 향상된 결과를 보이며, 다른 방법들에 비하여 우수한 성능을 보이는 결과를 보였다.

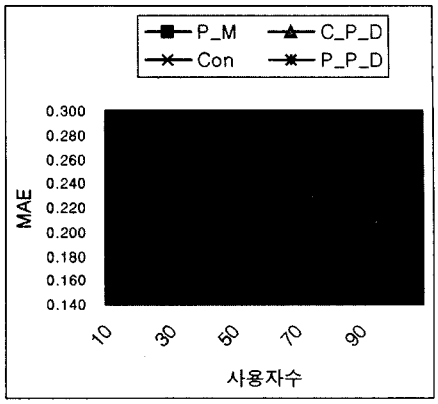

그림 2 사용자 수의 변화에 따른 MAE

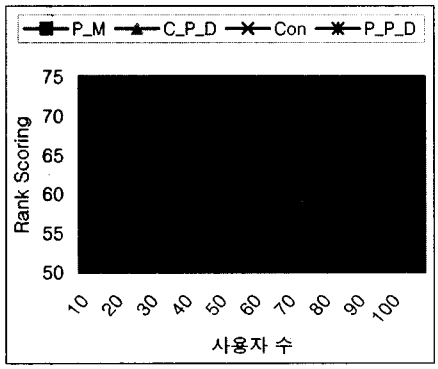

그림 3 사용자 수의 변화에 따른 순위 측정

표 13은 식 (11)와 식 (13)를 기반으로 문서에 대해 평가한 횟수를 증가시킴에 따른 제안된 협력적 여과와 내용 기반 여과를 병합한 추천 방법(P_P_D), Lee 방법, Pazzani 방법, Good 방법의 MAE와 순위 스코어를 나타낸다.

그림 4와 그림 5는 표 9를 기반으로 한 사용자가 평

표 13 n번째 평가 횟수에 따른 MAE와 순위 스코어

n번째 평가	MAE				Rank scoring			
	Pazzani	Lee	Good	P_P_D	Pazzani	Lee	Good	P_P_D
10	0.235	0.255	0.298	0.232	62.3	57.2	60.1	63.2
20	0.231	0.241	0.283	0.221	63.1	57.3	60.5	64.0
30	0.228	0.239	0.274	0.211	63.1	58.1	60.1	64.2
40	0.221	0.238	0.271	0.208	64.1	58.2	60.4	64.8
50	0.198	0.238	0.263	0.199	65.5	58.4	60.6	65.3
60	0.199	0.237	0.258	0.190	65.8	58.4	60.9	65.8
70	0.187	0.236	0.251	0.182	66.8	60.4	60.9	66.5
80	0.186	0.236	0.241	0.178	67.0	62.7	61.1	67.9
90	0.186	0.237	0.231	0.169	67.9	65.3	61.1	69.1
100	0.185	0.237	0.220	0.160	68.0	67.3	62.3	72.3

가한 횟수를 증가시킴에 따른 MAE와 순위 스코어 척도를 나타낸다.

그림 4와 그림 5에서 사용자의 프로파일을 이용하는 P_D_P와 Pazzani의 방법은 전반적으로 높은 성능을 보이며, 상대적으로 Lee와 Good의 방법은 이들 방법보다는 낮은 성능을 보인다. 부가적으로, 내용 기반 사용자의 프로파일과 협력적 사용자의 프로파일을 이용하는 P_D_P의 방법은 협력적 사용자 프로파일만을 사용하는 Pazzani의 방법보다는 높은 성능을 보인다.

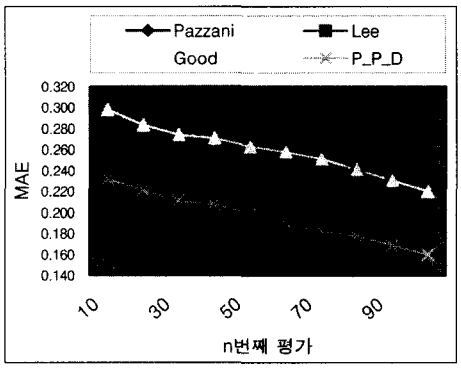

그림 4 n번째 평가에서의 MAE

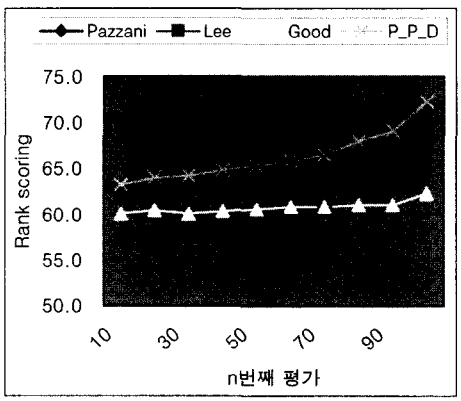

그림 5 n번째 평가에서의 순위 스코어

6. 결 론

본 논문에서는 협력적 여과 시스템의 희박성과 초기 평가 문제를 해결하기 위하여 사용자 프로파일을 생성시킴으로써 자동으로 선호도를 평가하는 방법을 제안하였다. 자동으로 사용자의 선호도를 발견하기 위해서 사용하는 프로파일은 협력적 여과 시스템에서의 {사용자-문서} 행렬을 기반으로 생성된 사용자 프로파일과 내용 기반 여과 시스템에서 연관 피드백에 의해 생성된 사용자 프로파일을 상호정보를 이용하여 병합함으로써 생성한 내용 기반 협력적 사용자 프로파일이다. 내용 기반 협력적 사용자 프로파일은 협력적 사용자 프로파일과 가장 유사한 프로파일을 소유하는 내용 기반 사용자를 검색하여 찾은 후, 두 프로파일을 병합한 결과를 기반으로 기존의 협력적 사용자 프로파일을 수정한 결과이다. 내용 기반 협력적 사용자 프로파일은 정규화 과정을 거친 뒤에 {사용자-문서} 행렬의 6단계 중 하나로 변화하여 자동으로 선호도를 평가한다.

제안된 방법의 성능을 평가하기 위해 협력적 여과 시스템에서 자동으로 선호도를 평가하는 기존의 방법과 비교하였으며, 또한 기존의 협력적 여과 시스템과 내용 기반 여과를 병합하여 프로파일을 생성하는 방법과 비교하였다. 그 결과, 각각의 방법에서 본 논문에서 제안한 방법이 기존의 방법보다 높은 성능을 보였다.

참 고 문 헌

[1] B. M. Sarwar, J. A. Konstan, Al Borchers, J. Herlocker, B. Miller, and J. Riedl, "Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System," Proceedings of the 1998 Conference on Computer Supported Cooperative Work, 1998.

[2] W. S. Lee, "Collaborative learning for recommender systems," In Proceedings of the Conference on Machine Learning, 1997.

[3] J. Delgado and N. Ishii, "Formal Models for Learning of User Preferences, a Preliminary Report," In Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-99), Stockholm, Sweden, July, 1999.

[4] 곽미라, 조동섭, "개인화된 추천 시스템의 선호도 계산을 위한 정보 필터링", 정보과학회춘계학술발표논문집, Vol. 28, No. 1, pp. 472-474, 2001.

[5] R. Raymond and J. Mooney and L. Roy, "Content-Based Book Recommending Using Learning for Text Categorization," Proceedings of the Fifth ACM Conference on Digital Libr*aries*, San Antonio, TX, pp. 195-204, June, 2000.

[6] M. J. Pazzani, "A framework for collaborative, content-based and demographic filtering," Artificial Intelligence Review, pp. 393-408, 1999.

[7] M. Balabanovic and Y. Shoham, "Fab: Content-based, collaborative recommendation," Communication of the Association of Computing Machinery, Vol. 40, No. 3, pp. 66-72, 1997.

[8] C. Basu and H. Hirsh and W. W. Cohen, "Recommendation as classification:Using social and content-based information in recommendation," In proceedings of the Fifteenth National Conference on Artificial Intelligence, pp. 714-720, Madison, WI, 1998.

[9] N. Good, J. B. Schafer and J. A. Konstan, A. Borchers, B. Sarwar, J. Herlocker, and J. Riedl, "Combining collaborative filtering with personal agents for better recommendations," In Proceedings of National Conference on Artificial Intelligence(AAAI-99), pp. 439-446, 1999.

[10] I. Soboroff and C. Nicholas, "Combining content and collaboration in text filtering," In Proceedings of the IJCAI'99 Workshop on Machine Learning in Information filtering, pp. 86-91, 1999.

[11] M. Pazzani, D. Billsus, *Learning and Revising User Profiles: The Identification of Interesting Web Sites*, Machine Learning, Kluwer Academic Publishers, pp. 313-331, 1997.

[12] S. J. Ko and J. H. Lee, "User Preference Mining through Collaborative Filtering and Content based Filtering in Recommender System," Proceedings of EC_WEB2002, LNCS2455, Springer, pp. 244-253, 2002.

[13] D. Billsus and M. J. Pazzani, "Learning collaborative information filters," In proceedings of the International Conference on Machine Learning, 1998.

[14] K. Funakoshi, T. Ohguro, "A content-based collaborative recommender system with detailed use of evaluations," Proceedings of the 4th International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies, Vol. 1, pp. 253-256, 2000.

[15] V. Rijsbergen and C. Joost, *Information Retrieval*, Butterworths, London-second edition, 1979.

[16] S. J. Ko and J. H. Lee, "Feature Selection using Association Word Mining for Classification," In Proceedings of the Conference on DEXA2001, LNCS2113, pp. 211-220, 2001.

[17] 인하대학교, 사용자 중심의 지능형 정보 검색 시스템, 최종 연구 개발 보고서, 정보통신부, 1997.

[18] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994.

[19] R. Agrawal and T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases," In Proceedings of the 1993 ACM SIGMOD Conference, Washington DC, USA, 1993.

[20] 백준호, 최준혁, 이정현, "한국어 웹 정보검색 시스템의 정확도 향상을 위한 연관 피드백 에이전트", 한국정보처리학회 논문지, 제6권, 제7호, pp. 1832-1840, 1999.

[21] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[22] John. S. Breese and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proceedings of the Conference on Uncertainty in Artificial Intelligence, Madison, WI, 1998.

고 수 정

1990년 2월 인하대학교 전자계산학과 졸업. 1997년 2월 인하대학교 교육대학원 전자계산교육 전공(석사). 2002년 2월 인하대학교 대학원 전자계산공학과 졸업(박사). 2003년 5월 2004년 4월 University of Illinois at Urbana-Champaign Research Scientist. 2004년 5월~현재 Colorado State University Research Scientist. 관심분야는 data mining, web mining, 기계학습

최 성 용

1993년 인하대학교 통계학과 졸업(이학사). 2001년 인하대학교 대학원 통계학과(이학석사). 2001년~현재 인하대학교 전자계산공학과 박사과정. 관심분야는 베이지안 학습, 신경망, 지능형 에이전트

임 기 욱

1977년 인하대학교 공과대학 전자공학과 졸업. 1987년 한양대학교 전자계산학 석사. 1994년 인하대학교 전자계산학 박사 1977년~1983년 한국전자기술연구소 선임연구원. 1983년~1988년 한국전자통신연구소 시스템소프트웨어 연구실장. 1988년~1989년 미 캘리포니아주립대학(Irvine)방문 연구원 1989년~1997년 한국전자통신연구원 시스템연구부장 주전산기(타이컴) Ⅲ,Ⅳ개발 사업책임자. 1997년~2000년 정보통신연구진흥원 정보기술 전문의원. 2000년~현재 선문대학교 교수. 관심분야는 실시간 데이타 베이스시스템, 운영체제, 컴퓨터구조

이 정 현

1977년 인하대학교 전자공학과 졸업 1980년 인하대학교 대학원 전자공학과(공학석사). 1988년 인하대학교 대학원 전자공학과(공학박사). 1979년~1981년 한국전자기술연구소 시스템 연구원. 1984년~1989년 경기대학교 전자계산학과 교수. 1989년~현재 인하대학교 컴퓨터공학부 교수. 관심분야는 자연어처리, HCI, 정보검색, 음성인식, 음성합성, 컴퓨터구조