내용 기반 필터링을 위한 프로파일 학습에 의한 선호도 발견

Discovery of Preference through Learning Profile for Content-based Filtering

정경용*, 조선문**
상지대학교 컴퓨터정보공학부*, 배재대학교 IT**

Kyung-Yong Chung(kyjung@sangji.ac.kr)*, Sun-Moon Jo(sunmoon@pcu.ac.kr)**

요약

사용자가 정보를 효율적으로 이용할 수 있도록 제어하고 필터링하는 일을 도와주는 정보 시스템이 등장하였다. 내용 기반 필터링은 아이템의 특징을 기술하는 정보와 사용자의 기호를 가지고 있는 프로파일을 비교하여 사용자에게 필요한 정보를 추천하는 방법이다. 이는 학습 방법에 따른 정확도가 변한다는 문제점이 있다. 본 논문에서는 내용 기반 필터링을 위한 프로파일 학습에 의한 선호도 발견을 제안하였다. 문제점을 개선하기 위해서 6단계로 평가한 선호도에 따른 추정치를 부여하여 프로파일 학습을 함으로써 추천의 정확도를 향상시켰다. 제안한 방법을 MovieLens 데이터에 적용하여 실험 및 평가를 실시하였는데, 기존 연구와 비교 실험을 통해 성능을 평가하였다.

■ 중심어 :|추천 시스템|정보 검색|데이터마이닝|내용 기반 필터링|

Abstract

The information system in which users can utilize to control and to get the filtered information efficiently has appeared. Content-based filtering can reflect content information, and it provides recommendation by comparing the feature information about item and the profile of preference. This has the shortcoming of the varying accuracy of prediction depending on learning method. This paper suggests the discovery of preference through learning the profile for the content-based filtering. This study improves the accuracy of recommendation through learning the profile according to granting the preference of 6 levels to estimated value in order to solve the problem. Finally, to evaluate the performance of the proposed method, this study applies to MovieLens dataset, and it is compared with the performance of previous studies.

■ keyword : |Recommender System|Information Retrieval|Data Mining|Content Based Filtering|

1. 서 론

원투원 마케팅 사업에서 개인화 추천은 사용자의 선호도에 부합하는 아이템을 제공함으로써 이를 얻기 위한 시간과 비용을 줄여주고, 손쉽게 아이템에 접근하도록 하는 장점을 갖는다. 또한 개인화를 통해 자신과 비슷한 유형을 갖는 사용자와 교류할 수 있는 기회를 가질 수도 있다[1][2]. 개인화 추천 시스템은 정보 검색이

* 본 논문은 2007년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음 (KRF-2007-331-D00418)

접수번호 : #070614-002
접수일자 : 2007년 06월 14일
심사완료일 : 2007년 08월 02일
교신저자 : 정경용, e-mail : alice8105@dankook.ac.kr

나 정보 필터링의 일부분으로 구축된다. 추천 시스템은 사용자가 어떤 정보에 관심이 있는지, 어떤 정보가 유용한지를 프로파일을 기반으로 개인화 예측에 초점을 맞춘다. 여기서 사용되는 추천 기법은 협력적 필터링, 내용 기반 필터링, 통계 기반 필터링, 사례 기반 필터링이 있다. 본 논문에서는 아이템의 특징을 기술하는 속성 정보와 선호도를 가지고 있는 프로파일을 비교하여 필요한 정보를 추천하는 내용 기반 필터링에 대해서 기술한다. 이는 정보 검색 분야의 학습 알고리즘에 의해 적용되기가 쉽지 않고 학습 방법에 따라 추천의 정확도가 변한다는 단점을 갖는다[3]. 따라서 내용 기반 필터링의 정확도를 향상시키기 위해서 프로파일 학습에 의한 선호도 예측 방법을 제안하였다.

2. 텍스트 범주화 방법

텍스트 범주화를 처리하기 위한 기존 연구는 학습 문서에서 학습한 결과를 이용하여 문서에 적당한 범주를 할당하는 방법을 사용하였다. 또한, 최근 통계적인 분류 방법과 기계 학습에 대한 관심이 증대됨에 따라 문서 분류는 다변량 회귀 모델, 최근 인접 분류, 베이즈 확률 접근, 의사결정트리, 신경망, 기호 규칙 학습, 연역 학습 알고리즘 등으로 적용되었다[4]. 이들 중에서 텍스트 문서의 분류를 위한 연구[6]에서 높은 분류 효율을 나타내는 방법은 Naïve Bayes 분류법이다. Joachims는 Naïve Bayes 분류자를 사용하여 유즈넷 뉴스 기사 분류를 시도하여, 그 결과로 89%의 분류 정확도를 얻을 수 있음을 보였다. Lang은 순위화된 기사를 학습 집합으로 사용하여 사용자가 관심있는 기사를 예측함을 보였다[8]. McCallum[7]은 Naïve Bayes 가정을 사용한 연구를 크게 두 가지의 형태로 분류하고 실험을 통하여 비교 평가하였다.

첫 번째 형태는 문서 내의 단어의 발생과 비발생만을 고려하여 문서를 분류하는 방법으로 일반적으로 이진 독립 모델이라 명하거나 특별히 문서 분류에 있어서 다중 이형 베르눌리 모델이라고 명하였다. 이진 독립 모델에서 문서를 단어의 공간에 대한 벡터의 형태로 간주하기 때문에 주어진 사전으로부터, 공간($t\in\{1,2,\cdots,|V|\}$)의 각각의 차원은 단어 w_t에 부합한다. 문서 d_i에 대한 벡터의 차원 t를 $B_{i,t}$라고 표현한다. 이는 단어가 문서 내에 출현하면 1, 출현하지 않으면 0이라고 하면 (식 1)로부터 주어진 클래스의 문서는 문서 내에 존재하는 단어의 확률 곱으로 나타낼 수 있다.

$$p(d_i|c_j,\theta)=\sum_{t=1}^{|V|}(B_{i,t}P(w_t|c_j,\theta) + (1-B_{i,t})(1-P(w_t|c_j,\theta))) \quad \text{(식 1)}$$

레이블된 학습 문서를 $D=\{d_1,d_2,\cdots,d_{|D|}\}$로 표현하면, 확률 분포 모델의 파라미터를 학습하는 것은 클래스 조건의 단어 확률을 추정하는 것으로 설명할 수 있다. 파라미터는 $0\le\theta_{w_t|c_j}\le1$인 조건에서 $\theta_{w_t|c_j}=P(w_t|c_j,\theta)$와 같이 정의할 수 있다. 레이블에 따라 $P(c_j|d_i)\in\{0,1\}$라 하면 클래스 c_j에서 단어 w_t의 확률을 (식 2)에 의해 추정할 수 있다.

$$\theta_{w_t|c_j}=\frac{1+\sum_{i=1}^{|D|}B_{i,t}P(c_j|d_i)}{2+\sum_{i=1}^{|D|}P(c_j|d_i)} \quad \text{(식 2)}$$

클래스의 사전 파라미터 θ_{c_j}는 (식 3)의 최대 우도 측정에 의해 얻어진다.

$$\theta_{c_j}=P(c_j|\theta)=\frac{\sum_{i=1}^{|D|}P(c_j|d_i)}{|D|} \quad \text{(식 3)}$$

이러한 이진 독립 모델을 기반으로 단어의 출현 및 비출현 확률을 질의에 따른 연관과 비연관으로 간주하여 검색된 문서의 순위화와 연관 피드백에 사용한 연구가 있다[8].

두 번째 형태는 문서 내의 단어의 발생과 비발생 뿐만 아니라 단어의 출현 빈도까지 고려하는 방법으로 다항 모델이라 부른다. 다항 모델에서 문서는 사전으로부

터 유도되는 단어의 순서화된 연속으로 표현된다. 여기서 문서의 길이는 각 클래스에 대해 독립임을 가정한다. 다항 모델도 문서내의 단어 확률은 문서 내에서 단어의 문맥과 위치에 독립적이라는 Naïve Bayes 가정을 갖는다[9]. $N_{i,t}$를 단어 w_t가 문서 d_i에서 발생하는 빈도수라고 정의하면, (식 4)로부터 클래스에 대한 문서의 확률은 다음과 같다.

$$P(d_i|c_j,\theta)=P(|d_i|)|d_i|!\prod_{t=1}^{|V|}\frac{P(w_t|c_j,\theta)^{N_i t}}{N_i t!} \quad \text{(식 4)}$$

단어의 확률은 $0\le\Theta w_t|c_j\le 1$, $\sum\Theta w_t|c_j=1$인 조건에서 $\Theta w_t|c_j=P(w_t|c_j,\Theta)$와 같이 정의할 수 있다. 레이블된 학습 문서로부터 (식 5)의 파라미터를 추정할 수 있고 클래스의 사전 확률 파라미터는 (식 3)으로부터 구할 수 있다.

$$\theta_{w_t|c_j}=P(w_t|c_j,\theta_j)=\frac{1+\sum_{i=1}^{|D|}N_{i,t}P(c_j|d_i)}{|V|+\sum_{s=1}^{|V|}\sum_{i=1}^{|D|}N_{i,s}P(c_j|d_i)} \quad \text{(식 5)}$$

McCallum[7][8]은 위의 두 가지 연구를 기반으로 웹문서, 유즈넷 기사를 포함하는 문서 집합에 대해 두 가지 분류 방법을 적용하여 분류 효율을 비교하였다. 그 결과 단어의 출현 빈도와 문서의 길이를 함께 고려하는 다항 모델이 이진 독립 모델에 비해 평균 27%의 에러가 감소됨을 보였다. 텍스트 검색과 범주화에서 가장 중요한 것은 텍스트의 특징을 추출하는 방법이다. 텍스트의 특징 추출 방법에 따라 텍스트 검색과 범주화의 성능은 차이를 보인다. 특징 추출을 위해 기계 학습에서 이용하는 기존 방법은 역문서 빈도, 정보 이득, 상호 정보, 용어 연관도 등의 방법이 있다. 역문서 빈도는 역문서 빈도수를 단어의 빈도수와 같이 적용함으로써 그 문서를 대표하는 단어를 쉽게 찾을 수 있는 기법으로 (식 6)으로 표현 가능하다[8]. (식 6)은 문서 d_i에서 단어 t_k가 갖는 특징으로서 중요도 또는 가중치를 표현하며 $w_{i,k}$로 정의한다.

$$w_{i,k}=f_{i,k}\cdot\ [\log(N/DF)+1] \quad \text{(식 6)}$$

문서의 빈도 DF는 N개의 문서 중에서 단어 t_k가 존재한 문서의 개수를 의미하고 $log(N/DF)$+1은 역문서 빈도수를 의미한다. 역문서 빈도를 이용한 특징 추출은 높은 단어부터 낮은 단어로 정렬하여 상위의 빈도를 나타내는 단어만을 문서의 특징으로 추출한다. 역문서 빈도는 특징 표현을 위한 어휘 감소에 있어서 가장 간단한 방법이나 특징을 예상하는데 있어서 한계를 나타내는 단점을 갖는다[8].

3. 프로파일 학습에 의한 내용 기반 필터링

사용자는 특정 저자 또는 제목을 검색하여 훈련 집합을 선택하고 평가한다. 이는 데이터베이스를 전부 검색하거나 무작위로 아이템을 선택하는 것을 피할 수 있다. 사용자는 선택한 각각의 아이템에 대해 1-6까지 1씩 증가하면서 총 6단계의 선호도를 평가할 수 있다. 본 연구에서 사용하는 웹문서는 웹 로봇 에이전트에 의해 추출된 아이템이다[8]. 본 논문에서는 역문서 빈도를 이용한 Naïve Bayes 분류자를 적용하기 위해서는 웹문서를 형태소 분석하여 특징을 추출하고 그 결과 중에서 명사만을 사용한다. 추출된 모든 명사에 대해서 역문서 빈도를 가중치로 사용한다. 역문서 빈도는 높은 단어부터 낮은 단어로 정렬하여 상위의 빈도를 나타내는 단어만을 문서의 특징으로 추출한다. 그리고 훈련 집합의 선호도는 모델의 매개변수를 추정할 때, 훈련 문서에 가중치를 부여하기 위해서 사용한다.

3.1 역문서 빈도에 의한 가중치

단어 빈도수 방법은 많은 웹문서 중에서 웹문서를 대표할 수 있는 특징을 추출하기 위해서 단어 빈도수를 사용한다. 여기서 단어 빈도수가 높은 것이 웹문서를 정확히 대표하는 단어가 된다고 할 수 없다. 실제의 경

우, 웹문서를 대표하는 핵심어는 빈도수가 크게 높지 않기 때문이다. 이러한 단어 빈도수의 문제점을 개선하기 위하여 여러 웹문서에서 많은 빈도를 나타내는 단어는 일반적인 단어로서 웹문서의 대표성과는 관련성이 떨어진다고 간주하는 방법이 역문서 빈도수이다. 이는 웹문서에서 대표하는 단어를 쉽게 찾을 수 있는 기법으로 (식 7)로 표현이 가능하다. (식 7)은 웹문서 d_m에서 i번째 단어가 갖는 특징으로서의 가중치 또는 중요도로 표현하며 $W_{m,i}$로 정의한다.

$$W_{m,i} = TF \cdot IDF = TF_{m,i} \cdot [\log_2(N) - \log_2(DF_i) + 1] \quad \text{(식 7)}$$

$W_{m,i}$는 m번째 웹문서에서 i번째 단어의 가중치를 나타내며, $TF_{m,i}$는 m번째 웹문서에서 i번째 단어가 나타난 횟수를 나타낸다. N은 전체 훈련 문서에 존재하는 문서의 개수를 나타내고, DF_i는 i번째 단어가 나타난 문서의 수를 나타낸다. (식 7)에서 TF 항목은 웹문서에서 빈도수가 높은 단어에 가중치를 많이 주고, IDF 항목은 반대로 여러 웹문서에 나타나는 단어의 가중치를 감소시킨다. 이와 같이 가중치를 주는 방법은 정보 이론에 따라 정보량이 많은 단어에 많은 가중치를 주면서도 간단한 계산으로 가능하다는 장점이 있다. 정보 이론에서 엔트로피의 개념을 사용하여 (식 8)과 같이 정의할 수 있다. 본 연구에서는 계산상의 경제성으로 인하여 (식 7)의 방법을 사용한다.

$$W_{m,i} = \frac{\log(freq_{m,i})}{-\sum((\frac{freq_{m,i}}{\sum freq_{m,i}}) \times \log(\frac{freq_{m,i}}{\sum freq_{m,i}}))} \quad \text{(식 8)}$$

3.2 Naïve Bayes 분류자를 이용한 프로파일 학습

단어의 위치와 출현 빈도를 고려해 파라미터의 확률을 추정하는 다항 모델을 기반으로 프로파일을 학습하고 웹문서를 분류한다. 제안하는 분류자는 문서에 출현하는 모든 단어의 확률을 곱해 최대가 되는 클래스에 해당 문서를 분류하는 기존 방법과 달리 웹문서의 위치에 따른 주제어를 추출하여 웹문서를 대표할 수 있는 주제어에 가중치를 부여한다. 그리고 각각의 단어의 출현 빈도를 고려하여 추정치를 부여함으로 정확도를 높이는 변형된 베이지안 분류자를 사용한다. 여기서 단어의 위치와 출현 빈도를 이용하는 단일 단어 위치 벡터 모델의 형태를 사용한다[3][4][10].

Naïve Bayes 분류자는 학습 단계와 분류 단계를 통하여 훈련 문서에 나타나는 모든 단어를 특징으로 분류한다. 데이터베이스에서 훈련 문서(D)의 특징이 $\{word_1, word_2, \cdots, word_t\}$라고 하였을 경우 $\{class_0, class_1\}$ 중 하나의 클래스로 훈련 문서(D)를 분류한다. 내용 정보 데이터베이스의 훈련 문서(D)에 각각의 클래스의 확률은 (식 9)를 이용한다[8].

$$p(class_i|D) = \frac{P(class_i)}{P(D)} \prod_{i=1}^{|D|} P(word_i \mid class_j) \quad \text{(식 9)}$$

(식 9)에서 $word_i$는 문서에서 i번째 단어이고, $|D|$는 단어를 포함하는 훈련 문서의 길이이다. 모든 훈련 문서에 대해 $P(D)$는 상수이므로 확률의 순위만을 고려하여 분류할 경우에는 상수는 의미없게 된다. 훈련 문서에 포함될 확률 비율 $P(class_1|D)/P(class_0|D)$에 의해 분류함으로써 순위가 정해진다. 확률 비율에서 $class_1$은 긍정적으로 평가받는 클래스이고, $class_0$는 부정적으로 평가받는 클래스이다. 확률 비율이 1보다 크면 긍정적인 클래스이고 그렇지 않으면 부정적인 클래스로 분류된다.

단어의 위치와 출현 빈도를 고려한 단일 단어 위치 벡터의 경우에 pos_m은 문서의 벡터에서 m번째 위치를 나타낸다고 할 때, 각 단어의 확률 $P(word_i|class_j,pos_m)$은 (식 10)을 이용한다[5].

$$P(word_i|class_j, pos_m) = \sum_{e=1}^{N} freq \cdot classw_{ej} / \sum_{e=1}^{N} classw_{ej}|d_m| \quad \text{(식 10)}$$

단어의 확률은 훈련 문서 집합에서 추정치가 부여된다. N개의 훈련 문서 집합에서 평가한 선호도($1 \le r_{i,j} \le 6$)를 기반으로 추정치가 부여된다. 긍정적인 클래스 추정치는 $classw_{e1}=(r_{i,j}-1)/5$로 나타내고, 부정적인 클래스 추정치는 $classw_{e0}=1-classw_{e1}$로 정의한다. 단어 $word_m$가 아이템 F에서 m번째 위치에서 n번 나타날 경우 긍정적인 클래스 추정치에는 $freq.classw_{e0}$를 적용하고, 부정적인 클래스 추정치에는 $freq.classw_{e1}$를 적용한다. 여기서 $freq$는 아이템 F가 m번째 위치에서 n번 나타난 횟수이다. 아이템 F에 대한 사후 확률은 (식 11)을 이용한다. $|pos|$는 위치의 수이고, $word_{m,i}$는 m번째 위치의 i번째 단어이다. W_{mi}는 역문서 빈도를 이용한 가중치이다.

$$p(class_j|F) \quad \text{(식 11)}$$
$$= \frac{P(class_j)}{P(F)} \prod_{m=1}^{|pos|} \prod_{i=1}^{|d_{i,m}|} W_{mi} \cdot \; P(word_{mi}|class_j, pos_m)$$

역문서 빈도를 이용한 가중치를 Naïve Bayes 분류자에 첨가하여 문서 분류에 사용한다. Naïve Bayes 분류자를 이용한 기존의 방식은 문서 내에 출현하는 단어의 확률 곱으로 분류 확률을 계산하기 때문에 문서의 오분류에 영향을 줄 수 있는 단어, 즉 잡음까지 확률 계산에 포함하게 된다. 따라서 기존의 방법에서 분류 효율의 저하를 유도하였다.

3.3 프로파일 학습에 의한 선호도

프로파일을 학습하는 3가지 방법에 대해서 실험을 통하여 각각의 방법에 대한 정확도를 비교 평가하였다. 첫 번째 프로파일 학습 방법은 이진 분류 문제를 위한 단순한 이진 Naïve Bayes 분류자를 이용하는 방법이다. 6단계의 선호도에서 4-6으로 평가받은 아이템은 긍정적인 클래스, 1-3으로 평가받은 아이템은 부정적인 클래스로 간주하여 프로파일을 학습하는 방법이다. 일반적으로 사용자는 부정적으로 평가하기 보다는 긍정적으로 평가될 가능성 때문에 프로파일 학습에 의한 예측을 확률적 이진 분류로서 재계산하는 것이다. 두 번째 프로파일 학습 방법은 훈련 문서에 대해서 6단계로 평가한 선호도를 6개의 클래스로 다루는 방법이다. 이는 6개의 클래스에 대해서 각각의 사후 확률을 계산한 후 예측을 위한 (식 12)로 정의한 기대값(E)을 계산한다. $P(i)$는 클래스 i에서 프로파일 학습에 의한 사후 확률이다.

$$E = \sum_{i=1}^{6} i \times p(i) \quad \text{(식 12)}$$

(식 12)에서 6개의 클래스에 대한 기대값으로 첫 번째 방법에서 제시한 이진 분류를 할 수 있다. 세 번째 프로파일 학습 방법은 두 번째 방법에서 6단계로 평가한 선호도에 추정치를 부여하는 방법이다. 아이템에 대해서 1-6까지의 6단계 선호도($r_{i,j}$)를 [0,1]의 값의 범위를 갖는 추정치로 정규화한다[10].

3.4 적합성 피드백을 이용한 추천

훈련 문서에서 프로파일이 학습되면, 긍정적으로 평가되어지는 클래스의 사후 확률을 기반으로 예측한다. 확률을 기준으로 상위 등급의 아이템을 정렬한 후 추천 리스트를 제공한다. 추천 리스트에 대해서 비추천 아이템이 있을 경우 정보 검색 분야의 적합성 피드백을 이용하여 정보를 수정한다. 최적의 추천 리스트를 만들기 위해서 적합성 피드백은 반복해서 수행해야 한다. 사용자가 적합과 부적합으로 판정하여 결과를 입력하면 (식 13)에 의해 새로운 추천 리스트를 형성하게 된다[8].

추천된 적합 아이템의 수를 R, 부적합 아이템의 수를 N이라고 하고 각각의 추천 아이템 집합을 D_R, D_N이라고 하면 수정된 질의는 Q'는 (식 13)으로 정의한다. 선호도 $D_i(=(d_{i1},d_{i2}, ...,d_{it}))$와 추천을 하기 위해 사용된 추천 벡터 $Q(=(q_1,q_2, ...,q_t,))$, Q'는 벡터 형태이다.

$$Q' = \alpha(Q) + \beta\left(\frac{1}{R}\sum_{i \in D_R} D_i\right) - \lambda\left(\frac{1}{N}\sum_{i \in D_N} D_i\right) \quad \text{(식 13)}$$

α, β, λ는 상수로서 각 항의 비중을 나타내며 Q'는 추천 벡터 Q에 적합 아이템의 평균 벡터와 부적합 아이템의 평균 벡터의 차이를 더해준 값이 된다. 적합 아이템에 포함된 추천 벡터에 가중치를 더해주고 부적합 아이

템에 포함된 추천 벡터에 가중치를 빼줌으로써 적합성 피드백을 하는 것이다.

4. 성능평가

4.1 실험 환경 및 실험 데이터

본 논문에서 제안한 방법은 MS Visual C++ 6.0으로 구현되었으며 실험 환경은 Pentium-4 1.6 Ghz, 512MB RAM 환경에서 수행되었다. 본 논문에서 사용하는 MovieLens 데이터[9]는 미국 미네소타 대학에서 개인화 추천을 연구하기 위해 아이템에 대해서 선호도를 평가한 데이터이다. 이는 기존 연구[5][8]에서 사용하였던 EachMovie 데이터를 사용한 것보다 정확한 성능평가를 할 수 있다. [그림 1]은 아이템에 대한 선호도의 평가 분포이다. 대부분의 아이템에 대해서 0.6 이상의 선호도를 평가한 것을 볼 수 있다. 이는 사용자가 부정적으로 평가하기 보다는 긍정적으로 평가됨을 알 수 있다.

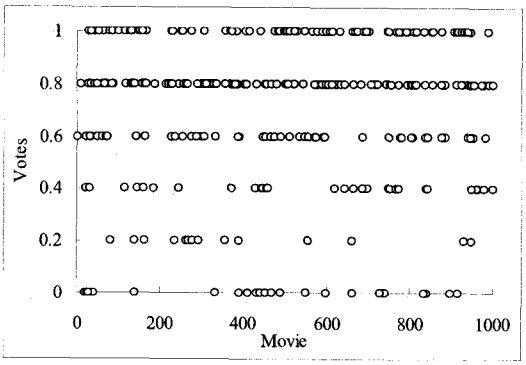


그림 1. 아이템에 대한 선호도의 평가 분포

성능 평가를 하기 위해서 MovieLens 데이터를 전처리하여 사용하였고 계산상의 편의와 메모리 절약을 위해 0-1까지 0.2간격으로 표현된 데이터에 5를 곱하여 0-5까지 1간격으로 변환하여 사용하였다. 최소 100회 이상 평가한 사용자 4,798명을 추출하여 이 가운데 1,000명을 기존 사용자 집합으로 두고 나머지 사용자 중에 무작위로 테스트 사용자 100명을 선택하여 총 1,628개의 아이템 중 테스트 사용자가 평가한 임의의 10개 아이템에 대해서 선호도를 예측하고 실제 선호도와 비교, 평가하였다.

4.2 분석 및 성능 평가

제안한 방법의 성능을 평가하기 위해 MAE를 사용한다. 제안된 내용 기반 필터링에서 프로파일 학습에 의한 선호도 예측을 평가하기 위해, 이진 분류 문제를 위한 단순한 이진 Naïve Bayes 분류자를 이용하는 방법(S_B_Bayes), 6단계의 선호도에 따른 6개의 클래스로 나누어 Naïve Bayes 분류자를 이용하는 방법(6_R_Bayes), 6단계의 선호도에 따른 6개의 클래스에 추정치를 부여하여 Naïve Bayes 분류자를 이용하는 방법(W_6_R_Bayes)을 훈련 집합의 크기를 변화시키면서 성능을 비교하였다.

표 1. 훈련 집합의 크기에 따른 MAE

훈련집합 크기	MAE		
	S_B_Bayes	6_R_Bayes	W_6_R_Bayes
10	1.320	1.292	1.299
25	1.282	1.246	1.179
50	1.210	1.205	1.095
80	1.204	1.201	1.093
100	1.188	1.179	1.067
150	1.191	1.131	1.065
200	1.191	1.127	1.061
250	1.192	1.094	1.059
300	1.101	1.073	1.052
400	1.107	1.085	1.052
500	1.128	1.070	1.051

[표 1]은 훈련 집합의 크기를 변화시킴에 따른 S_B_Bayes, 6_R_Bayes, W_6_R_Bayes의 MAE를 나타낸다. 훈련 집합의 크기가 커짐에 따라 6_R_Bayes, W_6_R_Bayes의 MAE는 낮아짐을 보인다. S_B_Bayes는 훈련 집합의 크기가 커짐에 따라 다소 낮아지기는 하나 큰 차이가 없음을 보인다. 그리고 제안된 방법은 훈련 집합의 크기에 관계없이 정확도가 유지되는 것을 볼 수 있다. MovieLens 데이터에서 10회의 동일한 실험을 하여, W_6_R_Bayes 방법이 6_R_Bayes 방법보다는 8.7%, S_B_Bayes 방법보다는 5.2%의 높은

성능 차이를 보인다. 본 연구에서는 문서를 대표할 만한 주제어를 사전에서 추출하여 문서의 주제에 관련없는 단어를 제거하거나 낮은 가중치를 부여하여 분류 오류를 감소하므로 정확도를 향상시켰다.

5. 결론

내용 기반 필터링에서 프로파일 학습 방법에 따라 추천의 정확도가 변한다는 단점을 개선하기 위해서 웹문서를 형태소 분석에 의한 단일 단어 위치 벡터 모델의 형태로 변경하고, 6단계의 선호도에 따른 6개의 클래스에 추정치를 부여하여 Naïve Bayes 분류자에 의한 프로파일 학습을 수행함으로써 추천의 정확도를 향상시켰다. 프로파일 학습에 의한 선호도 추천의 성능을 분석하기 위해 단순한 이진 Naïve Bayes 분류자를 이용하는 방법, 6단계의 선호도에 따른 6개의 클래스로 나누어 Naïve Bayes 분류자를 이용하는 방법, 6단계의 선호도에 따른 6개의 클래스에 추정치를 부여하여 Naïve Bayes 분류자를 이용하는 방법과 MAE를 이용하여 비교하였다. 비교 결과, 제안된 6단계의 선호도에 따른 6개의 클래스에 추정치를 부여하여 Naïve Bayes 분류자를 이용하는 방법의 성능이 우수함으로 보였다. 향후 연구로 기계 학습 방법으로 추천의 효율성과 확장성을 높일 수 있는 방안을 모색해야 할 것으로 보인다.

참 고 문 헌

[1] A. Ansari, S. Essegaier, and R. Rkohli, "Internet Recommendation Systems," Jour. of Marketing Research, Vol.37, pp.363-375, 2000.

[2] D. M. Pennock and E. Horvitz, "Collaborative Filtering by Personality Diagnosis: A Hybrid Memory and Model based Approach," Proc. of the Conf. on Uncertainty in AI, 2000.

[3] S. J. Ko and J. H. Lee, "User Preference Mining through Collaborative Filtering and Content Based Filtering in Recommender System," LNCS 2455, pp.244-253, 2002.

[4] D. D. Lewis and M. Ringuette, "Comparison of Two Learning Algorithms for Text Categorization," Proc. of the Symposium on Document Analysis and Information Retrieval, 1994.

[5] K. Y. Jung, "User Preference through Bayesian Categorization for Recommendation," LNAI 4099, pp.112-119, 2006.

[6] Y. H. Li and A. K. Jain, "Classification of Text Documents," Jour. of the Computer, Vol.41, No.8, pp.537-546, 1998.

[7] A. McCallum and K. Nigram, "A Comparison of Event Models for Naïve Bayes Text Classification," Proc. of the AAAI-98 Workshop on Learning for Text Categorization, 1998.

[8] 정경용, "*혼합 필터링과 연관 이웃 마이닝을 이용한 개인화 아이템 추천 기법*", 인하대학교 박사학위논문, 2005.

[9] http://www.cs.umn.edu/research/GroupLens/

[10] P. Melville, R. J. Mooney, and R. Nagarajan, "Content-Boosted Collaborative Filtering for Improved Recommendations," Proc. of the National Conf. on Artificial Intelligence, pp.187-192, 2002.

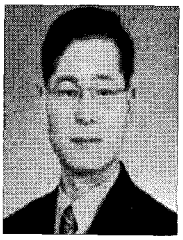
저 자 소 개

정 경 용(Kyung-Yong Chung) 정회원

- 2000년 2월 : 인하대학교 전자계산공학과(공학사)
- 2002년 2월 : 인하대학교 컴퓨터 정보공학과(공학석사)
- 2005년 8월 : 인하대학교 컴퓨터 정보공학과(공학박사)
- 2006년 3월 ~ 현재 : 상지대학교 컴퓨터정보공학부 교수

<관심분야> : 데이터마이닝, 지능시스템, 인공지능, 감성공학, 인지학습, 추천시스템, 임베이드 시스템

조 선 문(Sun-Moon Jo) **정회원**



- 2007년 : 인하대학교 컴퓨터정보공학과(공학박사)
- 2001년 ~ 2005년 : (주)세븐시스템 연구기획팀 팀장
- 2006년 ~ 현재 : 한국지식정보기술학회 논문 편집위원
- 2006년 ~ 현재 : 배재대학교 IT 교수

<관심분야> : XML, PL, 정보보안, 임베디드 시스템