단어 빈도와 α -cut에 의한 연관 웹문서 분류를 이용한 추천 시스템

Recommendation System using Associative Web Document Classification by Word Frequency and α -Cut

정경용*, 하원식**
상지대학교 컴퓨터정보공학부*, LG전자 MC사업부 단말연구소**

Kyung-Yong Jung(kyjung@sangji.ac.kr)*, Won-Shik Ha(sigboy@lge.com)**

요약

협력적 필터링을 개선하기 위하여 많은 기술들이 개발되고 실용화되었으나 아이템의 연관 관계를 정확하게 반영하지는 못한다. 본 논문에서는 협력적 필터링의 문제점을 보완하기 위하여 단어 빈도와 α-cut에 의한 연관 웹문서 분류를 이용한 추천 시스템을 제안한다. 제안된 방법은 형태소 분석을 통한 웹문서에서 단어를 추출하고 빈도 가중치를 계산한다. 추출된 단어를 Apriori 알고리즘을 이용해서 연관 규칙을 생성하고 신뢰도에 단어 빈도 가중치를 적용한다. 그리고 연관 규칙 하이퍼그래프 분할을 이용하여 연관 단어 간의 유사도를 계산한다. 마지막으로 유사 클래스를 기반으로 연관 웹문서를 α-cut을 이용하여 분류하고 개선된 코사인 유사도를 이용하여 유사도를 계산한다. 실험 결과 제안한 방법이 기존의 방법들보다 우수함을 확인하였다.

■ 중심어 : | 협력적 필터링 | 군집 | 연관규칙 | 추천 시스템 | 데이터마이닝 |

Abstract

Although there were some technological developments in improving the collaborative filtering, they have yet to fully reflect the actual relation of the items. In this paper, we propose the recommendation system using associative web document classification by word frequency and a-cut to address the short comings of the collaborative filtering. The proposed method extracts words from web documents through the morpheme analysis and accumulates the weight of term frequency. It makes associative rules and applies the weight of term frequency to its confidence by using Apriori algorithm. And it calculates the similarity among the words using the hypergraph partition. Lastly, it classifies related web document by using a-cut and calculates similarity by using adjusted cosine similarity. The results show that the proposed method significantly outperforms the existing methods.

■ keyword : | Collaborative Filtering | Clustering | Association Rule | Recommendation System | Data Miining |

* "본 연구는 2007년도 상지대학교 교내 연구비 지원에 의한 것입니다."

접수번호 : #070608-001
접수일자 : 2007년 06월 08일
심사완료일 : 2007년 08월 23일
교신저자 : 정경용, e-mail : kyjung@sangji.ac.kr

1. 서 론

추천은 사용자들에게 기호에 부합하는 상품이나 정보에 쉽게 접근하도록 하는 장점이 있다. 추천 시스템은 학습과 정보 필터링으로 구성되며 학습은 사용자 행위에 따라 사용자의 성향을 학습하는 것을 나타내며 내용 분석을 기반으로 한 자연어처리 분야와 밀접한 관계가 있다. 정보 필터링은 사용자에 따른 추천 정보를 제시하며 방법에 따라 협력적 필터링, 인구통계학적 필터링, 규칙 기반 필터링, 내용 기반 필터링, 컨텍스트 기반 필터링으로 나뉜다. 서로 상호 보완적인 장단점이 있기 때문에 최근에는 다수의 방법을 병합하여 사용하는 연구가 진행되고 있다. 추천 시스템의 단점은 아이템이 많아질수록 사용자가 관련된 정보를 얻는데 한계가 있기 때문에 아이템에 대해서 사용자간에 선호도를 평가할 확률은 적어지게 되고 상관관계를 비교해야할 아이템의 수는 증가하게 된다. 또한 아이템의 속성에 대한 선호도를 반영하지 못하는 문제점이 있다[1]. 그리고 아이템의 연관 관계가 반영되지 않아 최초 군집 선정시 정확성이 떨어지는 단점이 있다[2].

본 논문에서는 최초 군집 선정시 정확성을 개선하기 위하여 웹문서에서 추출한 단어들을 연관단어 형태로 만들고, 문서 분류시 문제가 되어왔던 중의성 문제를 해결하기 위하여 빈도 가중치를 적용하여 문서들을 분류한다[12]. 분류된 웹문서를 아이템이라고 간주하고 개선된 코사인 유사도를 이용하여 유사도를 계산하고 선호도 예측에 의한 아이템 추천을 한다.

2장에서는 현재까지 진행된 아이템 기반 협력적 필터링과 기존의 문서 분류 방법에 대해 살펴보고, 3장에서는 제안하는 단어 빈도와 α-cut에 의한 연관 웹문서 분류를 이용한 추천 시스템에 대해서 기술한다. 4장에서는 실험 및 평가를 제시하고 5장에서는 결론에 대하여 기술한다.

2. 아이템 기반의 협력적 필터링과 문서 분류

2.1 아이템 기반 협력적 필터링

아이템 기반 협력적 필터링은 평가한 아이템의 집합을 찾아내고 추천 후보 아이템과 유사한지 계산한다. 여기서 유사한 아이템 집합을 선택하며 동시에 대응하는 유사도를 계산한다. 유사한 아이템이 발견되면 가중치의 합으로써 예측하게 된다[4]. 아이템 유사도는 서로 다른 아이템에 대해 평가한 사용자를 분리해내고, 아이템간의 유사도는 코사인 유사도, 상관관계 유사도, 개선된 코사인 유사도로 계산한다[1]. 코사인 유사도는 아이템을 차원의 공간에 있는 벡터로 구성한다. 구성된 아이템간의 유사도는 벡터 사이의 코사인을 계산하여 측정된다. 이는 사용자 기반 협력적 필터링의 벡터 유사도와 같은 개념이다. 상관계수 기반의 유사도는 아이템간의 유사도를 피어슨 상관계수로 계산하는 방법이다. 사용자들이 공통으로 평가한 아이템을 기반으로 유사도를 계산한다. 이는 협력적 필터링의 대표적인 방법이다. 개선된 코사인 유사도는 코사인 유사도와 같은 개념이며, 기존의 방법이 서로 다른 사용자들 사이에서 나타나는 선호도의 차이를 고려하기 않는 단점을 보완한 방법이다. 기존 방법 대비 정확성 측면에서 우수하다는 장점이 있다.

2.2 기존의 문서 분류 방법

최근 문서 분류 방법으로 다변량 회기모델, 최근인접분류, 베이지안 확률접근, 의사결정트리, 신경망, 기호규칙학습, 연역학습이 제안되었다. 이는 특징 공간을 다차원으로 표현하고, 연관 단어가 아닌 단일 단어만으로 문서들을 분류하여 단어의 중의성을 반영하지 못해 오분류의 문제가 있다[12].

문서를 분류하는 기존 방법들 중 정보이득, 상호정보, 용어연관도, Naive Bayesian 분류자, K-means, SVM에 대해서 기술한다. 정보이득은 기계학습의 분야에 있어서 단어 선정의 기준으로써 자주 사용된다. 문서 안에 있는 단어의 출현과 비출현을 알아냄에 의해 카테고리 예측을 위해 획득한 정보의 비트의 수를 측정하는 방법이다. 상호 정보는 통계적인 언어에서 단어간의 유사성과 연관 관계를 확률적으로 측정하기 위해서 사용된다. 문서에서 단어가 출현한 횟수와 하나의 문장에서 출현한 빈도수를 측정한다. 용어 연관도는 정보 검색에

서 어휘 감소를 위해 사용된 방법이다. 이는 관련된 문서에서 단어가 얼마나 자주 나타나는가에 기반을 두어 단어를 평가한다. Naive Bayes 분류자는 학습과 분류 단계를 통하여 훈련 문서에 나타나는 단어를 특징으로 분류하는 방법이다. 성능이 뛰어나지만 각각의 단어에 대한 특징이 독립적임을 가정하고 있어서 중의성 문제와 단어간의 연관도를 전혀 반영하지 못하고 있다. K-means는 거리 기반 군집화 방법으로 다차원 공간상의 거리를 계산하여 군집하는 방법이다. 실행 속도는 빠르나 정확도가 낮은 단점을 불구하고 간결성 때문에 군집에 효율적으로 사용된다[6]. SVM은 많은 양의 데이터와 높은 차원의 집합을 가진 분류 작업에 특히 우수한 성능을 보이고 대량 문서의 자동분류 작업에 적합한 방법이다[5].

3. 단어 빈도와 α -cut에 의한 연관 웹문서 분류

기존의 시스템은 사용자 기반 협력적 필터링의 희박성과 확장성 문제를 개선하는데 기여하였다. 그러나 아이템간의 군집 형성 시 하나의 군집에 의존하여 최초 군집을 형성하고 이를 바탕으로 하여 추천하기 때문에 최초 군집의 정확성이 떨어진다면 정확한 추천 시스템을 기대할 수 없다. 특히 아이템간의 연관 관계를 고려하지 않거나 아이템의 속성을 반영하고 있는 빈도를 고려하지 않고 생성된 군집은 정확성을 높이는데 방해 요소로 작용할 수 있다. 기존의 문제점을 보완하기 위해서 최초 군집 형성 시 웹문서 기반의 연관 단어 빈도를 고려하였다. 사용자와 아이템의 유사도에 근거하여 협력적 필터링을 구현하였으며, 최종적으로 이를 바탕으로 추천을 수행하는 시스템을 제안하였다.

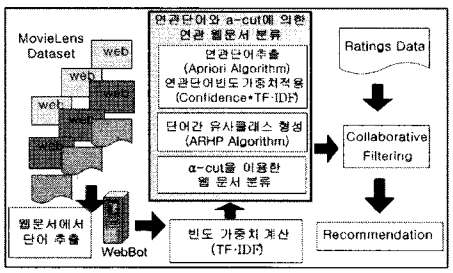

그림 1. 제안하는 추천 시스템의 구성도

제안하는 추천 시스템의 전체 구성은 [그림 1]과 같다. 본 논문에서 사용하는 MovieLens[7]은 미국 미네소타대학에서 GroupLens 연구 프로젝트를 하면서 수집된 데이터이다. 기존 연구[12]에서 사용하였던 EachMovie를 전처리한 것으로 보다 정확한 성능 평가를 할 수 있다. 적어도 20번 이상의 선호도를 가지는 사용자 943명이 1,682개의 아이템에 대한 선호도 100,000개의 래코드들로 이루어져있다. 실험 데이터에 포함된 URL의 웹 페이지에서 1,682 종류의 웹문서 내에 단어를 WebBot[2]을 이용하여 추출하였다. 추출된 단어들을 기반으로 빈도 가중치를 계산하고 Apriori 알고리즘을 이용하여 연관 규칙을 생성한다. 생성된 신뢰도에 빈도 가중치를 반영하여 연관 단어간에 유사클래스를 생성하기 위한 연관 규칙 하이퍼그래프 분할의 가중치로 사용한다. 생성된 유사 클래스들을 기반으로 웹문서들을 α-cut을 이용하여 분류한다. 분류된 웹문서들은 한편의 영화를 대표하는 아이템이라 간주하고 유사도를 계산하여 추천한다. 문서 분류시 중의성을 해결하였고 각 아이템의 주제를 잘 반영하고 있는 자주 등장하는 단어를 기반으로 하여 분류된 연관 웹문서로 추천을 함으로써 정확한 결과를 얻을 수 있다.

3.1 연관 단어간의 빈도 가중치

임의의 문서에서 특정 단어의 발생빈도가 높으면 높을수록 그 단어는 중요한 의미를 가지고 문서의 주제를 많이 반영하고 있다고 할 수 있다. 본 논문에서는 단어의 중의성 문제를 해결하기 위해 단어간의 유사도를 이용하여 웹문서를 분류한다. 단순히 단어의 연관성으로 문서를 분류한다면 그 단어의 발생빈도는 반영하지 못하게 된다. 따라서 정보검색 분야에서 사용하는 문서들 중에서 대표할 수 있는 특징을 추출하기 위해 사용하는 방법인 *TF·IDF*[13]을 사용하여 단어의 빈도 가중치를 반영한다. *TF·IDF*는 (식 1)과 같이 나타낼 수 있다. [표 1]은 각 단어별 *TF·IDF* 추정치를 나타낸다. *DF*는 단어가 나타난 문서의 수이다.

$$W_{ik} = f_{ik} \times [\log(n) - \log(DF) + 1] \qquad \text{(식 1)}$$

표 1. 각 단어별 TF·IDF 추정치

단어 \ 문서	$WDoc_1$	$WDoc_2$	$WDoc_3$	$WDoc_4$	$WDoc_5$	$WDoc_6$	…
guide	0.525	0.122	0.419	0.721	0.564	0.183	…
set	0.393	0.093	0.108	0.518	0.927	0.425	…
demand	0.422	0.286	0.527	0.253	0.714	0.115	…
network	0.775	0.583	0.288	0.152	0.623	0.577	…
space	0.111	0.719	0.624	0.711	0.252	0.421	…
…	…	…	…	…	…	…	…

연관 단어간의 빈도 가중치를 반영하기 위해 (식 1)을 이용한다. (식 2)는 연관 단어간의 빈도 가중치로서 각 연관 단어의 *TF·IDF*의 평균을 사용한다. n은 단어의 수, f_{ik}는 단어의 빈도수, WF는 단어가 나타난 웹문서의 수이다.

$$W' = \frac{1}{n}\left\{\sum^{n}(f_{ik} \times [\log(n) - \log(WF) + 1])\right\} \quad \text{(식 2)}$$

3.2 연관 규칙 생성 및 연관 단어 추출

Apriori 알고리즘은 데이터마이닝의 연관 규칙을 생성하는 방법으로 지지도를 이용하여 동시에 발생하는 아이템을 정제하고, 빈발 아이템 집합에서 생성된 규칙들은 신뢰도를 이용하여 정제하는 방식이다[8]. 여기서 후보 아이템 집합에서 각각의 지지도를 계산한 후 사용자가 정의한 지지도보다 크거나 같은 조건을 만족하는 데이터로 빈발 아이템 집합을 구성한다. 후보 아이템 집합들은 빈발 아이템 집합의 조인 연산을 통해 구성된다. 지지도는 연관 규칙을 반영하는 트랜잭션이 대용량 데이터베이스에서 얼마만큼의 비율을 차지하고 있는지를 나타내는 측정 기준으로서 통계적 중요성을 반영한 것이다. 신뢰도는 규칙이 실제로 정확한지를 판단하는 정도로서 연관 단어간의 강도를 나타내는 측정 기준(결합도)으로 사용한다. 본 논문에서는 단어의 중의성 문제를 해결하기 위하여 신뢰도를 사용한다. Apriori 알고리즘을 이용하여 연관 규칙을 생성한 후 선행단어, 후행단어, 지지도, 신뢰도, 연관 단어 수를 정리한 연관단어 지식베이스를 구축한다. [표 2]는 연관단어 지식베이스를 나타낸다.

표 2. 연관단어 지식베이스

선행단어	후행단어	신뢰도	지지도	단어수
[title][DVD][search][match]…	select	92.1%	28.6%	44
[nothing][city][another][miss]…	gridlock	95.6%	26.4%	22
[change][more][diligent]…	lifestyle	82.4%	34.8%	51
[strong][sound][beauty][zoo]…	funny	72.3%	37.6%	66

3.3 단어에 가중치를 부여한 연관 웹문서 분류

3.3.1 연관 단어간의 유사도 계산

연관 규칙을 기반으로 하여 연관 규칙 하이퍼그래프 분할 알고리즘으로 연관 단어간의 유사도를 계산한다. 데이터마이닝의 연관규칙과 하이퍼그래프 분할을 이용하여 트랜잭션 기반의 데이터베이스에서 연관된 항목들을 군집하는 방법이다[9]. 하이퍼그래프 분할을 위한 가중치로는 연관 규칙의 신뢰도에 (식 2)의 연관 단어간의 빈도 가중치를 적용한다. [표 3]은 수정된 신뢰도를 나타낸다. 하아퍼그래프는 단어들로 구성된 정점들의 집합과 빈번한 항목집합들을 나타내는 하이퍼 간선들의 집합으로 구성된다. 즉, 연관 규칙에 포함되는 항목들을 정점으로 연관 관계를 하이퍼 간선으로 매핑하는 것이다. 연관 단어간의 유사도를 이용하여 하이퍼그래프 분할로 유사 클래스를 생성한다. [표 4]는 생성된 유사 클래스를 나타낸다.

표 3. 연관 단어 빈도를 반영한 신뢰도

추출된 연관 단어	수정된 신뢰도
[title][DVD][search][match][choosing][best][voice][hike][excite][super][equal][same][good][name][tag]…	81.76%
[nothing][city][another][miss][night][woman][bad][team][dot][girl][done][day][all][dance][band][tip][hip]…	87.52%
[change][more][diligent][awareness][like][age][silver] [mellow][gold][blood][sexy][zoo][love][kiss][sex]…	76.29%
[strong][sound][beauty][zoo][somebody][clerk][speak] [long][do][make][bright][body][mouse][word][too]]…	69.52%
[record][all][learn][orange][score][total][boom][goal] [reputation][term][sum][code][stop][type][bag][bio]…	84.49%

표 4. 생성된 유사 클래스

클래스번호	클래스에 포함된 연관단어
Class 1	[moon][night][virgin][weird][speed][exit][child]···
Class 2	[city][another][bell][drink][mall][coffee][building]···
Class 3	[user][cell][score][extreme][winter][admire][tip]···
Class 4	[air][time][retire][die][useful][equipment][narrow]···

3.3.2 α -cut을 이용한 연관 웹문서 분류

α-cut은 소속 함수의 [0,1]사이의 값에서 임의의 α(0≤α≤1)값이 되는 함수에 대한 퍼지의 상태 변수의 구간을 나타낸다. 이는 퍼지 집합의 원소에 대해 집합에 속할 기준을 정의할 때 사용된다. 연관 단어 x를 원소로 하는 연관 웹문서에 대해서 임의의 [0,1]의 값을 가진 α-cut을 적용한 웹문서 집합 $WDoc_\alpha$는 (식 3)과 같이 나타낸다.

$$WDoc_\alpha = \{x | WDoc(x) \geq \alpha\} \qquad \text{(식 3)}$$

웹문서 집합 $WDoc_\alpha$는 유사 클래스에 속할 소속정도 값이 α값 이상으로 이루어진 집합이다. 본 논문에서는 웹문서 집합에서 추출한 단어간의 연관 규칙을 적용하여 웹문서를 분류함으로써 사용자의 관심도를 보다 많이 표현하고 키워드 단순 매칭에 의한 검색기법의 단점인 의미상으로 연결된 웹문서에 대한 분류를 해결한다. 유사정도에 따라 웹문서를 분류하는 방법은 웹문서에서 생성된 단어의 유사 클래스를 이용하여 동의어 문제를 해결할 수 있도록 분류한다. 따라서 사용자의 관심을 보다 효율적으로 반영하고 의미적으로 관련있는 웹문서를 동일한 카테고리로 분류함에 따라 정확한 분류를 할 수 있다. [표 5]는 클래스 번호가 1인 경우 웹문서에 포함되어 있는 단어의 가중치 평균이 0.7 이상인 웹문서만 선택하여 분류한 것을 알 수 있다. 0.7-cut을 사용한 이유는 실험을 통해서 α값이 0.7일 때 최고의 성능을 보였기 때문에 0.7-cut을 사용하였다[2].

표 5. 유사 클래스에 속한 웹문서 분류

단어 \ 웹문서	$WDoc_1$	$WDoc_2$	$WDoc_3$	$WDoc_4$	$WDoc_5$	$WDoc_6$	···
moon	0.513	0.283	0.591	0.716	0.737	0.438	···
night	0.327	0.429	0.743	0.246	0.671	0.611	···
virgin	0.221	0.294	0.418	0.167	0.917	0.794	···
weird	0.015	0.213	0.726	0.491	0.478	0.616	···
speed	0.045	0.671	0.612	0.624	0.691	0.815	···
exit	0.167	0.543	0.595	0.038	0.597	0.577	

3.4 분류된 연관 웹문서 기반의 협력적 필터링

연관 단어 빈도를 이용해서 분류된 연관 웹문서를 아이템으로 간주하여 생성된 군집 안에서 유사도를 계산한다. 사용자의 선호도를 이용하여 아이템 기반의 협력적 필터링으로 추천하게 된다.

3.4.1 아이템 유사도 계산

아이템 유사도는 서로 다른 두개의 아이템에 대해 평가한 사용자를 분리해 내고 아이템 i와 아이템 j에 대한 유사도를 계산하는 것이다. 본 논문에서는 분류된 각각의 웹문서를 아이템으로 간주하고 유사도를 정확하게 계산하기 위해 기존의 개선된 코사인 유사도를 변형한 (식 4)를 사용한다. 전체 사용자의 수를 M, 아이템의 개수를 N이라고 한다면, 아이템의 유사도를 계산하는데 최악의 경우 $O(MN^2)$이 필요하다. 하지만 사용자가 아이템에 대해 평가한 경우는 일부이기 때문에 $O(MN)$으로 근사하게 된다[4].

$$WDoc\_sim(I,J) = \frac{\sum_{u \in U}(R_{u,I} - \overline{R_I})(R_{u,J} - \overline{R_J})}{\sqrt{\sum_{u \in U}(R_{u,I} - \overline{R_u})^2}\sqrt{\sum_{u \in U}(R_{u,J} - \overline{R_u})^2}} \qquad \text{(식 4)}$$

3.4.2 선호도 예측

선호도를 예측하기 위해서는 기존의 방법인 가중치 합으로 표현되는 방법을 사용한다[1]. 아이템으로 간주되어 분류된 웹문서 I와 유사한 아이템에 대한 사용자 u의 선호도 합을 계산함으로써 예측한다. 각각의 선호도는 $WDoc_sim_{I,J}$에 의해 가중치가 적용되며 예측값

$P_{u,i}$는 (식 5)로 표현한다. N은 하나의 군집으로 만들어진 연관 웹문서의 집합이다.

$$P_{u,i} = \frac{\sum_{similar\ items,N}(WDoc\_sim_{I,N} * R_{u,N})}{\sum_{similar\ items,N}(|WDoc\_sim_{I,N}|)} \quad \text{(식 5)}$$

4. 실험 방법 및 결과

제안한 추천 방법은 MS Visual C++ 6.0으로 구현되었으며, 실험 환경은 Pentium-4 2.8 Ghz, 512MB RAM 환경에서 3가지 실험을 진행하였다. 첫 번째 실험은 기존의 문서 분류 방식과 제안한 방법과 성능을 비교 평가하였다. DF+NB 방법은 TF·IDF 기술에 Naïve Bayes 분류자를 적용한 방법이고 IG+NB 방법은 정보 이득 기술로 문서의 특징을 추출하여 Naïve Bayes 분류자로 분류하는 방법이다. WF+AC 방법은 제안하는 단어 빈도와 α-cut에 의한 연관 웹문서 분류 방법이다. 두 번째 실험은 K-mean[6], SVM[5]과 제안하는 WF+AC 방식과 사용자의 수의 증가에 따른 성능 평가를 하였다. 세 번째 실험은 추천 시스템에서 예측의 정확도의 성능을 비교 평가하였다. P+CF 방법은 피어슨 상관계수 기반의 협력적 필터링이고 ACS+ICF 방법은 개선된 코사인 유사도를 이용한 협력적 필터링이고 WF+AC+ICF 방법은 제안하는 방법을 이용한 협력적 필터링이다.

4.1 분석 및 성능 평가

성능 평가하기 위해서는 정확률과 재현율을 이용한 F-Measure 측정식[13]과 예측의 정확도를 평가하기 위해 MAE(Mean Absolute Error) 방식을 사용하였다[4][10]. 제안하는 방법에서 사용하는 α-cut의 적절한 수치를 찾아내기 위하여 α값을 변화시키면서 실험을 진행하였다. [그림 3]의 웹문서의 수에 따른 F-measure의 성능 평가에서 α값은 0.7을 사용하여 분류하는 것이 성능이 좋은 것을 알 수 있다.

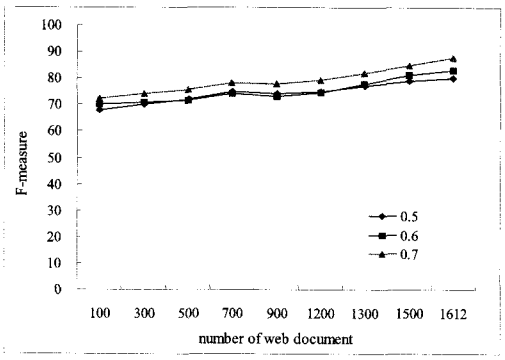

그림 3. α 값에 따른 연관 웹문서 분류 성능 평가

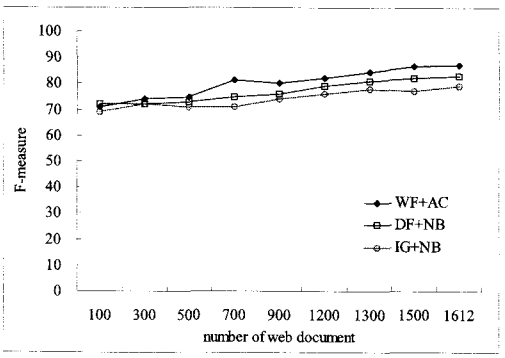

그림 4. 웹문서의 수에 따른 성능 평가

[그림 4]는 DF+NB 방법, IG+NB 방법, 실험에 의해 얻어진 α값 0.7을 사용한 WF+AC 방법과 웹문서의 수에 따른 F-measure의 성능 평가를 하였다. WF+AC 방법의 성능은 80.14%로 IG+NB 방법보다 6.03%, DF+NB 방법보다는 3.15%가 높다. 전체적으로 제안한 WF+AC 방법이 기존의 분류 방법들보다 성능이 우수함을 알 수 있다.

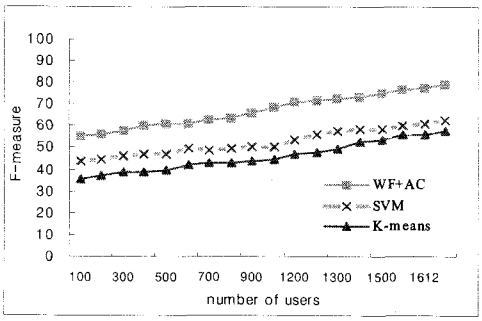

그림 5. 사용자 수에 따른 성능 평가

[그림 5]는 K-means와 SVM을 이용하여 사용자 수의 증가에 따른 F-measure의 성능 평가이다. 사용자의 수가 증가됨에 따라 모두 성능이 향상됨을 보였으며 특히 제안하는 WF+AC 방법은 K-means보다 20.1%, SVM보다는 10.3%의 높은 성능을 나타내고 SVM과 K-means는 사용자 수가 적은 경우에는 전체적으로 낮은 성능을 보였다. 이는 제안하는 방법이 사용자의 수가 적은 경우에도 높은 성능을 나타냄을 보였다.

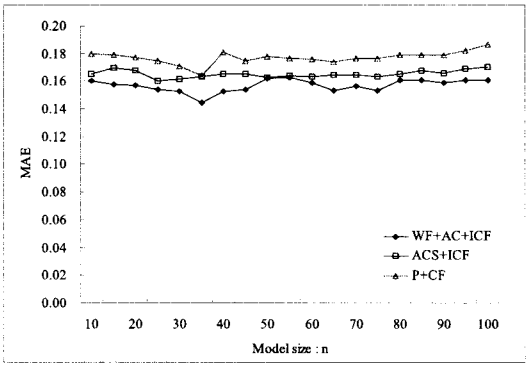

그림 6. 모델 크기에 따른 정확도 성능 평가

[그림 6]은 추천 시스템에서 P+CF 방법, ACS+ICF 방법, WF+AC+ICF 방법과 모델 크기에 따른 정확도 성능 평가이다. 여기서 10명의 유사 사용자를 참조하였으며, 각 방법의 모델 크기 n에 따라 실험을 진행하였다. 모델 크기 n은 추천을 요구한 사용자에게 추천될 유사 아이템 개수를 의미한다. 실험은 10명의 사용자가 각각 10개의 아이템에 대한 추천을 요구했을 때의 경우를 가정하였으며 동일한 사용자 리스트가 모든 방법에 적용되었다. 성능 평가를 보면 WF+AC+ICF 방법이 P+CF 방법 대비 2.09%의 MAE 감소를 나타내고, ACS+ICF 방법 대비 0.82%의 MAE 감소를 보여준다. 여기서 MAE 감소는 추천 시스템의 예측 정확도 향상을 의미한다. 제안한 방법이 모델 크기 35에서 가장 좋은 예측의 정확도 향상을 보이며, 모델 크기가 커짐에 따라 ACS+ICF 방법의 정확성에 근접함을 알 수 있다. WF+AC+ICF 방법이 P+CF 방법 보다는 정확도가 우수하지만 ACS+ICF 방법과 비교해보면 오히려 특정 구간에서의 정확도가 떨어지는 것을 볼 때 WF+AC+ICF 방법의 정확도를 항상 신뢰하기는 어려운 것으로 판단된다. 그러므로 정확도의 향상이라는 측면에서는 긍정적인 결과를 보여주지만 여러 단계를 통한 오버헤드의 문제는 차후 보완해야 할 것이다.

5. 결론

사용자 기반 협력적 필터링이 가지는 주요 단점으로 지적 되었던 희박성과 확장성에 대해 아이템 기반 협력적 필터링이 주목할 만한 성과를 거두었다. 특히 확장성 문제에 대한 성능 향상은 실시간 추천을 요구하는 추천 시스템에 많은 기여를 하였다. 그러나 기본적으로 명시적 데이터에 기반한 접근 방법이기 때문에 여전히 희박성의 문제가 남아 있으며, 아이템간의 연관 관계를 고려하지 않는다는 문제점이 있다. 기존의 추천 시스템에서 아이템간의 연관 관계를 고려하지 않아 생기는 문제점을 해결하기 위해 단어빈도를 이용하여 아이템간의 연관 관계 및 빈도 가중치로 추천의 정확도를 높이는 성과를 거두었다. 기존의 방식과는 다르게 웹문서의 분류를 바탕으로 추천 시스템을 구현하였고 웹문서 분류시 정확도를 높이기 위하여 단어빈도와 α-cut에 의한 연관 문서 분류를 이용한 것이 제안하고자 하는 주 개념이다. 분류된 웹문서를 기반으로 성능이 뛰어나다고 알려진 개선된 코사인 유사도를 사용하여 연관 문서간의 유사도를 계산하여 추천하였다. 제안한 방법에 대한 성능이 기존의 방법과 비교 실험한 결과 예측의 정확도면에서 효과적임을 알 수 있었다.

참 고 문 헌

[1] G. Linden, B. Smith, J. York, "Amazon.com Recommendations: Item-to-Item Collaborative Filtering," Internet Computing, IEEE, Vol.7, No.1, pp.76-80, 2003.

[2] K. Y. Jung and J. H. Lee, "Prediction of User Preference in Recommendation System using

Association User Clustering and Bayesian Estimated Value," LNAI 2557, pp.284-296, Springer-Verlag, 2002.

[3] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Analysis of Recommendation Algorithms for E-Commerce," Proc. of ACM E-Commerce 2000 Conference, pp.158-167, 2000.

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

[5] C. Vladimir and M. Yunqian, "Practical Selection of SVM Parameters and Noise Estimation for SVM Regression," Journal of Neural Networks, Vol.17, pp.113-126, 2004.

[6] O. Carlos, "Clustering Binary Data Streams with K-Means," Proc. of the Workshop on Research Issues on Data Mining and Knowledge Discovery, pp.10-17, 2003.

[7] http://www.cs.umn.edu/Research/GroupLens/

[8] J. Han and M. Kamber, *Data Mining: Concept and Techniques*, Morgan Kaufmann, pp.225-333, 2001.

[9] E. Han, G. Karypis, V. Kumar, and B. Mobasher, "Clustering Based On Association Rule Hypergraphs," Proc. of SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, pp.9-13, 1997.

[10] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating Collaborative Filtering Recommender Systems," ACM Transactions on Information Systems, Vol.22, No.1, pp.5-53, 2004.

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

[12] 하원식, 정경용, 이정현, "협력적 필터링을 위해 연관 단어 빈도를 이용한 웹문서 분류", 제31회 한국정보과학회 추계학술발표 논문집(I), pp.160-162, 2004.

[13] 정영미, *정보검색론*, 구미무역 출판부, pp.182-293, 2003.

저 자 소 개

정 경 용(Kyung-Yong Jung) 정회원

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

<관심분야> : 데이터마이닝, 지능시스템, 인공지능

하 원 식(Won-Shik Ha) 정회원

- 1999년 2월 : 동국대학교 컴퓨터공학과(공학사)
- 2005년 2월 : 인하대학교 컴퓨터정보공학과(공학석사)
- 2005년 3월 ~ 현재 : LG전자 단말연구소 주임연구원

<관심분야> : 데이터마이닝, 지능시스템, 인공지능