시간스키마 기법 2단계 클러스터링 적용 추천시스템의 성능 향상

김룡[O*], 부종수*, 홍종규**, 박원익**, 김영국*
충남대학교 컴퓨터과학과*, 충남대학교 컴퓨터공학**
{ryong[O], jkhong, wonik78}@cs.cnu.ac.kr, {bujongsu, ykim}@cnu.ac.kr

Two-step Clustring Method Using Time Schema for Performance Improvement in Recommender System

Kim Ryong[O*], Bu Jong-Su*, Hong Jong-Kyu**, Park Won-Ik**, Young-Kuk Kim*
Dept. of Computer Science, Chungnam National University*
Dept. of Computer Engineering, Chungnam National University**

요 약

기존의 추천 시스템들은 사용자 수가 증가함에 따라 추천시간이 증가하는 확장성(Scalability) 문제가 있으며, 새로운 고객의 경우 선호도 정보가 부족하여 추천 정확도가 저하되는 희박성(Saparsity) 문제가 있다.

본 논문에서는 고객의 기본 프로파일 정보 중 가장 변별력이 있는 성과 나이에 대한 그룹을 생성하고 클러스터링 함으로써 집단 내 선호 상품을 우선적으로 추천하는 1단계 클러스터링 방법을 사용하여 새로운 고객의 희박성 문제를 해결 했으며, 추천결과에 따른 피드백을 받아 시간 흐름에 따른 선호 경향을 클러스터링 하는 시간스키마 방법을 적용한 2단계 클러스터링 방법을 사용함으로써 확장성 문제를 해결함은 물론 예측 정확도를 높일 수 있는 방법을 제안한다.

1. 서 론

하루가 다르게 증가하고 있는 방대한 정보는 경제적인 측면에서 그 관리의 효율성 또한 중요하다는 사실을 일깨워 주고 있다. 과거에는 정보 제공자 입장에서의 소극적인 관리와 사용자 입장에서의 적극적인 검색이 있었다면, 현재는 그 양상이 반대로 변해야 한다는 것을 모두 인정한다. 이것은 정보 검색 분야에서 뿐만 아니라 경제활동의 중추로 부상하는 전자상거래 분야에서 더욱 중요하다.

기존의 추천시스템들이 가지는 대부분의 문제점은 사용자 수가 증가함에 따라 추천시간이 증가하는 확장성(Scalability) 문제와 새로운 고객의 경우 상품에 대한 선호도 정보가 부족할 경우 추천 정확도가 저하되는 희박성(Saparsity) 문제가 있다. 이런 문제점들을 해결하기 위한 많은 연구와 실험이 이어져 왔으나 아직도 개선의 여지가 남아 있는 상황이다.[1]

본 논문에서는 시간스키마 적용 2단계 클러스터링 기법을 이용하여 확장성 문제와 희박성 문제를 해결하려했다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 살펴보고 3장에서는 논문에서 제안한 시간스키마를 이용한 2단계 클러스터링 기법에 대해 소개하고자한다. 4장에서는 설계 및 구현에 관하여 알아본다. 5장에서는 실험 데이터에 대한 내용과 제안방법에 대한 성능 평가를 수행한다. 마지막으로 결론 및 향후 연구 방향에 대해 기술한다.

2. 관련 연구

2.1 협력적 필터링 추천 알고리즘

협력적 필터링은 크게 두 부류의 알고리즘으로 나누어지는데,하나는 메모리기반 알고리즘(memory-based algorithms)이고 다른 하나는 모델기반 알고리즘(model-based algorithms)이다.

메모리기반 알고리즘은 예측을 하기 위해 전체 사용자 데이터베이스를 관리한다. 다시 말하면 모든 아이템에 대한 모든 사용자의 선호도 데이터베이스를 유지하며, 전체 데이터베이스를 통해 계산을 수행하여 예측을 생성 한다.

모델기반 알고리즘은 모델을 추정하거나 학습하기 위해 사용자 데이터베이스를 사용하며, 그렇게 생성된 모델은 예측을 위해 이용한다. 모델 중심 알고리즘은 사용자의 선호도를 사용자, 아이템, 그리고 평가 내용을 갖는 기술적인 모델로 재구성한다. 그리고 추천은 구성된 모델에 의뢰함으로써 이루어진다.

메모리기반 알고리즘은 모델기반 알고리즘보다 간단하며, 실제 상황에서 잘 작동하고 계속해서 새로운 데이터를 쉽게 추가할 수 있다. 그러나 단점으로 데이터베이스의 크기가 커짐에 따라 처리 시간이 증가한다는 것이다.[2]

모델기반 알고리즘의 경우, 일단 모델이 생성되면 예측은 빨리 계산되어진다. 그러나 데이터로부터 모델을 구축하는 것은 많은 시간이 소요되며 새로운 데이터를 추가하기 위해서는 모델을 전체적으로 재구성해야 하는 단점이 있다.

2.2 GroupLens 프로젝트

미네소타 대학의 GroupLens 프로젝트는 같은 취향이나 취미를 가진 사람들의 정보를 이용해 추천할 때 도움을 주는 시스템으로 어떤 사람의 아이템에 대한 관심도를 예측하기 위하여 다른 고객들의 평가를 모아 이용하는 분산 시스템이면서, 일반

* 본 논문은 한국산업기술평가원이 지정한 지역협력연구센터(RRC)인 충남대학교 소프트웨어연구센터의 지원으로 수행된 과제의 결과입니다.

적인 정보에 적합하도록 만들어진 필터링 기술이다.[3]

GroupLens는 인터넷 뉴스를 추천하는 시스템으로 소개된 이래 아마존, 리바이스, CDNOW 등과 같은 사이트들에서 여러 형태로 널리 사용되고 있다.

GroupLens는 두 가지 평가 방법을 이용한다. 첫째 Target User와 다른 고객들의 평가와 가장 유사한지 연관성을 계산하는 것이고, 둘째 그 유사한 고객들의 평가를 근간으로 새로운 아이템에 대한 평가를 예측하는 방법이다. 여기서 피어슨의 [-1,1]의 값을 갖는 상관계수식을 이용하여 상관관계를 구하는데 값에 따라 1로 접근할수록 양의 상관관계(Target User와 다른 회원은 유사한 성향을 가진다.)를 가지고 -1로 접근할수록 음의 상관관계를 가지며, 0으로 접근할수록 서로간의 상관관계가 없다는 의미로 해석된다.

3. 문제 해결 방안

3.1 새로운 고객문제 해결을 위한 1단계 클러스터링

처음 방문한 고객은 그 고객의 프로파일 정보 이외에 어떤 선호경향도 파악할 수 없으므로 정확한 서비스를 제공할 수 없다. 따라서 기존의 협력적 필터링 시스템은 예측하기 전에 전형적으로 새로운 고객이 초기 정보 수집단계에서 요청 항목에 대해 평가하도록 요구한다. 그리고 그 평가를 기반으로 코사인 함수나 피어슨 상관계수식을 적용하여 사용자간의 유사도를 구한 다음, 아이템에 대한 선호도를 예측하는 방식을 사용한다. 그러나 모든 사용자간의 유사도 계산을 기본으로 하는 방식은 데이터의 희박성(sparseness)으로 인해 실제로 적용하는데 중요한 문제점을 가지고 있다.

본 논문에서 제안하는 1단계 클러스터링 과정에서 이 문제를 해결한다. 아이템에 대한 어떤 평가정보도 없는 새로운 고객에게 어느 정도 성향에 맞는 아이템을 제시하기 위하여 가장 변별력이 있는 성과 나이에 대한 1단계 클러스터링을 수행한다. 그리고 그룹 프로파일에서 높이 평가된 아이템 순으로 우선적으로 추천하는 것이다. 이렇게 하면 다른 사용자들과 차별화할 수 있고, 전체 사용자에 대한 유사도를 계산하지 않아도 되므로 연산시간도 줄일 수 있다.

3.2 시간적 윈도우 스키마를 적용한 2단계 클러스터링

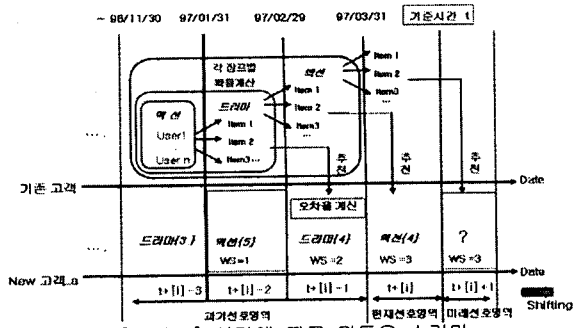
기존 논문에서 해결되지 않은 가장 큰 문제점 중의 하나가 시간에 따른 사용자의 선호 경향 파악이다. 시간적 흐름 성향을 반영하는 사전확률 방법을 이용하여 앞으로의 사용자 선호도를 표현하기 위해서는 많은 고객의 이용 히스토리가 필요하고 이에 따른 자료 수집 비용이 많이 들어간다는 문제점이 있다. 따라서 기존 논문에서는 보통 정적인 방법을 이용하여 전체 데이터에 대한 상관관계를 계산하는 방식을 이용함으로써 시간적인 성향변화에 대한 분석이 고려되지 않았다.

시간적 윈도우 스키마를 적용함에 있어 염두에 두어야 할 점은 무엇 보다 최근에 수집된 데이터가 향후 선호도 확률을 계산하는데 적절히 이용되지 못하고 있다는 점이다. 사용자의 과거 이용 데이터의 크기가 새로운 데이터 크기에 비해 월등히 크기 때문에 주로 과거 데이터에 의해 선호도가 좌우되는 문제점이 있다. 특히 선호도가 시간에 따라 자주 변하는 경향을 가지는 고객들에게는 과거 데이터 보다는 최근 데이터가 미래 선호도를 예측하는데 더 많은 영향을 줄 것이다.[4]

본 논문에서는 시간스키마에 따른 2단계 클러스터링 생성 기법 적용으로 평가시점별 사용자의 선호도 테이블을 갱신함으로써 사용자의 최근 선호경향을 반영하였다.

시간스키마 적용을 위해서 사용자들의 히스토리 데이터를 시간 축(월)을 기준으로 여러 개의 작은 집합으로 나눈 다음 2단계 클러스터링을 전 시점에서 같은 장르를 본 고객에 대해서만 윈도우 사이즈 안에서 동적으로 수행한다. 그리고 다음 추천시점에서 그 고객들이 평가한 장르들 중 대표 선호 장르를 순서대로 정렬하여 보여 준다.

즉 과거 평가 데이터를 이용하여 추천하는 것이 아니라 최근 윈도우 사이즈 만큼의 데이터를 이용하여 향후 선호 아이템을 예측하여 추천해 줌으로써 시간에 따른 고객 선호 경향 반영에 큰 역할을 한다.



[그림 1] 시간에 따른 윈도우 스키마

[그림 1]은 윈도우 사이즈가 3인 시점에서의 선호 영역과 2단계 클러스터링 수행 과정

3.3 시간적 윈도우 스키마 시나리오

윈도우 사이즈는 클러스터링 되는 영역 사이즈로 과거선호영역이 된다. 새로운 고객이 평가하는 장르 패턴을 적용하여 시점별로 클러스터링 되는데 본 논문에서는 최대 3으로 하였으며, 3을 넘으면 오른쪽으로 이동(Shifting) 되어 윈도우 사이즈는 항상 3을 유지한다.

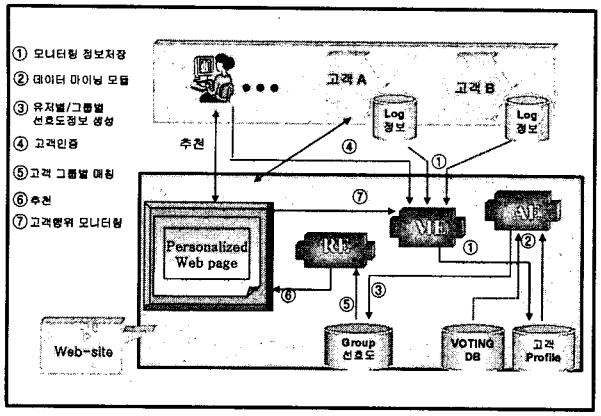
t는 기준시간으로 훈련 집합과 실험 집합을 나누는 기준점이 된다. t+[i]+n은 시간성향 반영을 위해 한 달 간격으로 시간 축을 나누었다. [장르i]+n에서 n은 월 간격으로 시간 축을 구분했기 때문에 한 달 동안 장르 [i]를 평가한 회수가 된다.

[그림 1]에서 과거 선호영역은 현시점을 기준으로 윈도우 사이즈만큼의 전 평가 데이터 영역을 의미하고, 현재 선호영역은 평가하는 현재 시점을, 미래선호 영역은 예측하고자 하는 바로 앞 시점의 선호 영역을 의미한다. t+[i]-2 시점에서의 새로운 고객이 평가한 대표 장르는 액션이고, t+[i]-1 시점에서는 드라마, t+[i] 시점에서는 액션이다.

4. 목표 시스템 설계 및 구현

[그림 2]는 본 논문에서 제안한 2단계 클러스터링 기법을 적용할 목표 시스템 환경구조이다. 실험에 사용한 웹사이트는 Dynamic Learning VOD(Video On Demand) 시스템으로 고객이 원하는 비디오를 찾을 수 있도록 도와주는 추천 사이트 이다. 추천 사이트는 고객들에게 비디오를 대여해주고, 그 대여 정보를 데이터베이스에 저장하며, 피드백을 줄 때마다 동적으로 선호도 정보를 갱신하여 최신 선호 경향을 반영하여 준다.

시스템 구성 엔진모듈은 데이터 처리 모듈인 분석엔진(AE) 그리고 고객행위를 모니터링하는 모니터링 엔진(ME) 핵심 모듈인 추천엔진(RE)으로 나누어 볼 수 있다.[5]

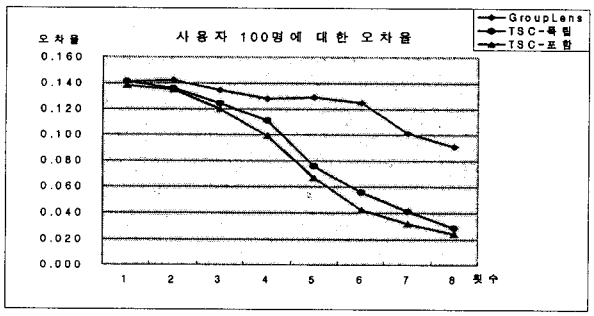


[그림 2] VOD시스템 아키텍쳐

5. 실험 및 성능평가

5.1 실험방법

본 실험에서는 2단계 클러스터링 방법이 성과 나이 그룹에 독립적이냐 또는 포함관계에 있느냐에 따라 기존 피어슨 상관관계를 이용한 추천과 연산시간과 예측정확도면에서 어느 정도 향상되었는지 평가한다. 그리고 장르를 10개에서 76개로 세분화하였을 경우 추천리스트를 비교함으로써 좀 더 개인화된 맞춤정보 제공을 위해서는 아이템 속성이 반영되어야 함을 평가한다. [그림 3]은 실험 결과로 기존 피어슨 상관계수식을 이용한 추천 예측정확도 평가와 본 논문에서 제안한 2단계 클러스터링 방법의 예측정확도 평가 결과이다.



[그림 3] 사용자 100명에 대한 MAE 그래프

5.2 예측정확도 평가

협력 필터링을 이용한 추천 시스템의 실험에서는 일반적으로 예측 정확도 평가 방법으로는 MAE(Mean Absolute Error)를 사용한다. MAE는 실험에서 발생한 평균 절대 오차 값을 말하며, 전체 예측 회수에 대해 발생한 평균 예측 오차를 의미하고, 아래 [식1]과 같이 계산된다.

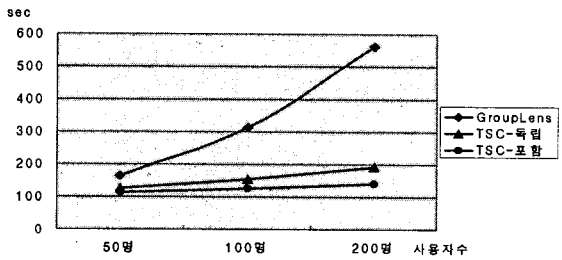
$$E = \frac{\sum | P - v |}{n}$$

[식 1] 예측 정확도 평가

[식 1]에서 P는 사용자 상품 선호도 예측 값이며 v는 사용자 실제 평가 값이다. 예측 값과 실제 값의 차이를 구하여 그 차이를 누적하여 평균을 구함으로써 평균 얼마정도의 차이가 있는지 알 수 있는 식이다. MAE가 작을수록 추천 시스템의 예측 정확도가 높음을 의미한다.

6. 결론 및 향후 연구과제

본 논문에서는 추천 시스템에서 가장 보편적으로 보이고 있는 협력적 필터링(Collaborative Filtering) 방법 하에서 이러한 문제들을 해결하기 위한 방안을 제시하였다. 대부분의 협력적 필터링 시스템들은 사용자간의 유사도를 구하는데 코사인함수나 피어슨 상관계수식을 이용하므로 아이템수가 많아질수록 사용자가 아이템에 관련된 정보를 얻는데 어느 정도 한계가 있기 때문에 두 사용자간에 선호도를 표시할 확률은 적어지게 되고, 상관관계를 비교 할 아이템 수는 증가하게 된다.



[그림 4] 사용자 증가에 따른 선호도 계산 시간

본 논문에서는 [그림 4]와 같이 시간에 따른 성향 변화를 고려하고 2단계에 걸쳐 클러스터링 하는 기법을 사용하여 새로운 사용자에 대한 예측 정확도 및 시스템 속도 면에서 선호도 계산시간을 단축시킴으로써 효율성을 개선하였다.

향후 연구과제로는 사용자의 선호도를 보다 잘 표현할 수 있는 데이터를 선택 연구하여 보다 높은 성능을 보장할 수 있는 시스템을 구축하는 것이 필요하고 본질적으로 협력추천시스템이 안고 있는 결여 데이터를 해결 혹은 완화할 수 있는 방안들이 결합되었을 때 본 모델도 더 나은 성능을 기대할 수 있을 것이다. 또한, 추천대상자(Target User)와 유사한 히스토리를 가지는 특정 N명의 이웃이 가진 정보를 바탕으로 추천이 이루어지므로 국소적 추천에 머물게 되고, 나머지 이웃들에게서 이끌어낼 수 있는 전역적 추천을 놓칠 수 있으므로 그에 대한 방안의 연구가 필요하다.

참고 문헌

[1] kai Yu, Anton Schwaighofer, Volker Tresp, Xiaowei Xu and Hans-peter Kriegel, "Probabilistic Memory-Based Collaborative Filtering", IEEE Transactions on knowledge and data Engineering, vol. no.1 January 2004.

[2] 이정원 외6인, "데이타마이닝 알고리즘의 분류 및 분석", 한국정보과학회 논문지D-데이터베이스, 제28권 제3호 pp. 279~300, 2001. 09

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

[4] S. Kang, J. Lim and M. Kim, "Modeling the User Preference on Broadcasting Contents Using Bayesian Belief Network Presentation," VCIP, Vol. 5308, pp. 958-967, Jan. 2002.

[5] 박성준, 김영국, 김용, "B2B e-Marketplace에서 웹 에이전트 기반 추천 시스템", 한국정보과학회, VOL. 31 NO. 01 pp. 754~756, 2004. 04