상황정보와 공간 데이터 마이닝 기법을 이용한 추천 시스템

이배희°, 조근식
인하대학교 컴퓨터공학부
coinone°@eslab.inha.ac.kr gsjo@inha.ac.kr

Recommender System using Context Information and Spatial Data Mining

Bae-Hee Lee°, Geun-Sik Jo
School of Computer Science & Engineering, Inha University

요 약

유비쿼터스 시대를 향하여 나아가는 현대 사회에서 사람들을 위한 추천시스템은 필수 불가결한 요소 중의 하나이다. 추천 시스템 중에서 사용자의 성별, 나이, 직업 등의 인구 통계적 요소를 고려한 시스템이 주를 이루고 있지만 이러한 시스템에는 어느 정도의 한계가 있다. 추천에 있어서 사용자의 기분, 날씨, 온도 등 주변 환경의 상황이 반영되지 않고 있고 학습을 위한 데이터에 대한 신뢰도 또한 문제가 된다. 이러한 문제점을 해결하기 위해 본 논문에서는 상황정보(Context Information)와 공간 데이터 마이닝(Spatial Data Mining) 기법을 이용한 향상된 추천 시스템을 제안한다. 제안하는 시스템에서는 보다 정확한 추천을 위해 첫째, 날씨, 온도, 사용자의 기분 등의 상황정보를 고려하였다. 그리고 사용자의 유사도 측정을 통해 학습 데이터의 신뢰도를 향상시켰으며, 셋째, 의사결정 트리(Decision Tree) 기법을 이용하여 추천의 정확도를 높였다. 실험을 통하여 측정한 결과, 제안하는 추천시스템이 기존의 인구 통계적 요소만을 고려한 시스템이나 의사결정 트리만을 이용한 시스템보다 향상된 성능을 보였다.

1. 서 론

유비쿼터스 환경 속에서 사용자의 요구를 최대한 반영할 수 있는 추천 시스템에 대한 연구가 활발하다. 기존의 추천 시스템에 관한 연구를 보면 사용자의 특정 조건에 맞는 데이터만을 통과시키기 위한 필터링 기법의 시스템이 일반적 이였다. 인구통계학적으로 사용자 유형별 특징 분석을 통한 인구통계학기반의 추천 시스템, 사용자와 유사한 선호도를 가진 다른 사용자의 평가에 근거한 협업 추천 시스템 등을 예로 들 수 있다 [1]. 이러한 기존의 시스템들은 추천을 위해 사용자들의 개인정보, 구매정보 혹은 상품평가 정보와 같은 방대한 과거 정보의 수집 및 분석 결과를 필요로 한다 [2]. 과거의 개인정보를 이용한 추천 서비스와 방대한 정보 분석을 통한 추천은 여러 가지 한계를 가지고 있다. 첫째, 가령 음식점 추천 서비스를 요구할 경우에는 당시의 날씨나 온도에 따라 또는 사용자의 기분에 따라 선택하는 음식점이 달라진다. 둘째, 방대한 자료를 이용할 경우 사용자와는 전혀 다른 개인정보나 성향을 가진 사용자도 분석의 대상이 될 수 있다. 셋째, 단순한 필터링 시스템이 가지는 추천의 정확성 한계를 들 수 있다.

본 논문에서는 이러한 문제들을 해결하기 위해 상황 정보를 이용하여 추천의 성능을 향상시킨 시스템을 제안한다. 첫째, 인구통계학적 변수 외에도 날씨, 온도, 사용자의 기분 등의 상황정보를 고려하여 다양한 상황에 따른 분석을 가능하게 하였다. 둘째, 유사도 측정을 통해 사용자와 유사한 성향을 가진 데이터만을 학습데이터로 이용하였다. 이러한 전처리 과정은 데이터 신뢰도의 향상을 가져왔다. 셋째, 추천의 정확성 향상을 위해 전처리 단계를 거친 데이터를 의사결정 트리를 이용하여 분석하였다. 실험에서는 K-중첩 교차조사(k-fold cross-validation) 방법을 이용하여 학습데이터와 테스트데이터를 분류하고 평균절대오차율(Mean Absolute Error)을 측정하여 제안된 추천시스템의 성능을 증명하였다.

2. 필터링을 통한 추천 방법

2.1 인구통계학 추천

인구통계학적 추천(Demographic-based Recommendation) 시스템은 사용자의 성별, 나이, 직업 등과 같은 인구통계학적 정보에 의한 사용자 유형별 특징 분석을 통해 상품을 추천한다 [3]. 인구통계학적 추천 방식은 많은 사용자들이 유사한 흥미를 갖는 영역에서 사용자를 위한 집단별 추천은 효과적이지만, 다양한 주변 환경적 요소를 고려하기 위한 추천으로써는 비효과적일 수 있다.

2.2 내용기반 추천

내용기반 추천(Content-based Recommendation) 시스템은 용자가 평가했던 상품에 대한 특징 정보와 다른 상품에 포함된 텍스트의 특징 정보의 유사도를 이용하여 필터링 한다 [4]. 효과적인 추천을 위해 사용자의 충분한 과거 프로파일 정보를 축적해야 하며, 이 프로파일과 비교하여 높은 점수를 얻은 컨텐츠를 추천함으로써 고객이 이미 평가한 컨텐츠와 유사한 컨텐츠만을 제공하여 새로운 상품에 대한 사용자의 평가가 없는 경우에 추천이 한정되는 문제점이 있다.

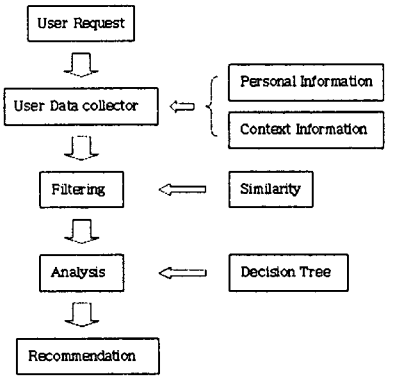
2.3 협업 필터링 추천

협업 필터링 추천(Collaborative Filtering Recommendation) 시스템은 사용자들의 상품에 대한 평가에 기초하여 유사한 성향을 갖는 사용자들을 분류하고, 이렇게 분류된 유사한 취향을 갖는 사용자로부터 상품에 대한 평가에 따라 추천하는 시스템이다 [4]. 이 시스템은 기존 사용자 부류와 유사성이 희박한 새로운 흥미를 갖는 사용자에 대한 추천의 어려움(gray-sheep

problem)은 물론, 새로운 상품에 대한 사용자들의 평가가 드물 때 해당 상품에 대한 추천의 어려움이 있다 [2].

3. 상황정보와 공간 데이터 마이닝 기법을 이용한 추천 시스템

본 논문에서 제안하는 시스템은 [그림1]과 같으며, 세 단계를 거쳐 추천이 이루어진다.



[그림1] 시스템 구조도

사용자의 요청이 들어왔을 때, 첫 번째 단계는 사용자의 개인정보와 상황정보를 수집하고, 다음 단계에서는 기존의 학습 데이터들에서 사용자의 개인정보와 유사한 그룹만을 분류해 낸다. 마지막 단계는 유사도를 통해 필터링된 데이터를 의사결정 트리 기법을 이용하여 분석한다. 위의 세 단계를 거쳐 요청한 사용자에게 추천을 한다.

3.1. 정보수집 단계

사용자가 추천을 요청하면 추천 시스템은 사용자의 개인정보와 상황정보를 수집한다. 개인정보는 사용자의 성별, 나이, 지출액, 결혼유무, 직업을 포함하고, 상황정보를 위해서는 날씨, 온도, 사용자의 기분, 날짜, 시간대를 수집한다. 본 논문에서는 시스템의 성능 테스트를 위하여 설문조사를 시행하였다. 설문에서는 사용자들의 개인정보와 상황정보를 조사하고 원하는 음식점을 선택하게 하였다. [표1]에 설문조사를 통해 얻은 데이터 중 하나를 나타내었다.

[표1] 사용자의 개인정보와 상황정보

번호	성별	나이	지출	결혼	직업	날씨	온도	기분	날짜	시간	음식점
5	1	27	50	1	1	1	7	8	8	17	3

3.2. 필터링 단계

사용자의 유사도(Similarity)를 측정하기 위해서는 유사 매트릭스 또는 여러 차원으로 유사도를 판단할 수 있는 방법이 있어야 한다 [5]. 본 논문에서는 추천 사용자로부터 추출된 개인 등록 정보와 설문을 통해 얻어진 기존 사용자들의 개인 등록 정보를 가지고 유사도를 분석한다. 사용자 유사도의 합(RIi)은 식 (1)과 같고 속성별 유사도는 식 (2)과 같다. CJi 는 직업의 유사도, Wj 는 직업에 대한 가중치를 나타낸다.

$$RIi = CSi + CAi + \ldots + CJi \quad (1)$$

$$CJi = Wj^{*}\ (1 - \frac{|\ SJ - TJ\ |}{3}) \quad (2)$$

수집된 사용자의 정보와 학습 데이터들 사이의 유사도를 식 (1)을 이용하여 계산한다. 실험을 통하여 최적화된 유사도의 임계치를 발견하고 임계치 미만의 데이터를 분석단계에서 제외시킨다. 사용자와 관련성이 적은 데이터를 제거함으로써 학습 데이터의 신뢰도와 추천의 정확성 향상을 가져온다. [표1]을 샘플데이터로 사용하여 샘플데이터와 학습데이터간의 유사도를 계산의 예를 [표2]에 나타내었다.

[표2] 유사도 계산

번호	성별	나이	지출	결혼	직업	날씨	온도	기분	날짜	시간	음식점	유사도
103	1	28	50	1	1	1	7	9	6	17	3	0.9733
63	1	24	40	1	1	1	9	6	8	16	1	0.9358
398	2	31	50	2	3	1	6	9	4	18	4	0.7683

3.3. 분석 단계

필터링 단계를 거쳐서 분류된 데이터들만을 훈련시켜 의사결정 트리를 생성한다. 의사결정 트리(Decision Tree)기법은 데이터 마이닝(Data Mining)의 분류 작업에 주로 사용되는 기법으로, 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 부류별 특성을 속성의 조합으로 나타내는 분류모형을 나무의 형태로 만드는 것이다 [6]. 의사결정 트리를 만들기 위해 가장 보편적으로 C5.0 알고리즘이 이용된다. 각 노드는 속성 값에 대한 정보이득(Information Gain)을 구한 후 정보이득이 가장 큰 속성이 최상위에 위치하게 된다. 엔트로피(Entropy)라는 척도의 감소량을 통해 정보이득을 구하고, 이를 이용하여 의사결정 트리를 구성하게 된다.

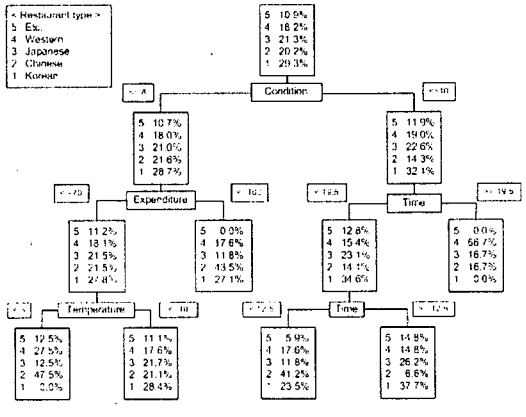
집합 S가 n개의 서로 다른 클래스에 속하고, pi를 S내에서 i번째 클래스가 차지하는 비율이라 한다면 S의 불순도 척도인 엔트로피는 식 (3)과 같이 정의된다.

$$E(S) = \sum_{i=1}^{n} - piLog_{2}pi \quad (3)$$

집합 S를 속성 A의 값에 따라 n개로 나누었을 때 얻게 되는 정보이득은 식 (4)과 같이 계산된다.

$$Gain(A) = E(S) - \sum_{k=1}^{n} E(S_k)\frac{|\ S_k\ |}{|\ S\ |} \quad (4)$$

식 (3)을 이용하여 속성 값의 엔트로피를 계산하고 식 (4)을 이용하여 정보이득이 가장 큰 속성을 최상위에 위치시켜 의사결정 트리를 생성한다. 최적 임계치(0.8 이상)인 데이터들만 분석하여 얻은 트리는 [그림2]와 같다. 트리를 이용하여 [표1]에 나타난 샘플 데이터를 가진 사용자에게 한국 음식점을 추천하게 된다.



[그림2] 의사결정 트리

4. 실험 및 결과

4.1 실험환경

본 논문의 실험을 위해서 IIS 5.0, Microsoft Active Server Page 와 MS-SQL Server을 사용해서 구현하였으며, 실험환경은 펜티엄4 2.8GHz, 512MB RAM의 시스템이었다.

학습 및 테스트에 사용된 데이터들은 2005년 4월부터 8월까지 설문을 통해 수집되었으며 설문 전에 본 연구의 주요 목적, 조사의 필요성 및 조사방법에 대하여 충분한 설명을 하였다. 회수된 설문지 중 총 500부의 설문지를 실험에 이용하였다.

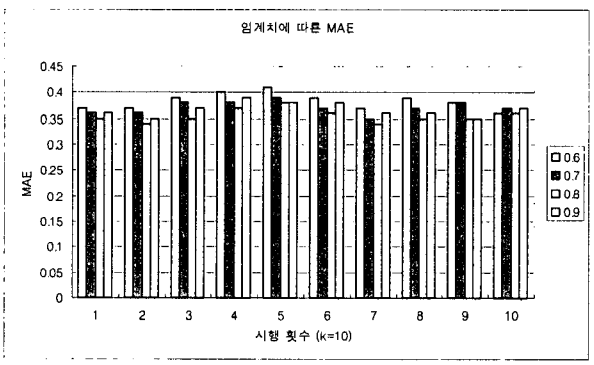
4.2 실험 평가 기준

제안하는 시스템의 정확성을 측정하기 위해서 K-중첩 교차조사(k-fold cross-validation) 방법을 이용하여 학습과 테스트를 k번 반복 수행한다. K-중첩 교차조사 방법은 데이터 셋을 k개의 상호 배반 부분 집합으로 분할한 후 i번째 집합을 테스트 집합으로 나머지를 학습 집합으로 사용한다 [7]. k번의 실험에서 평균절대오차율(Mean Absolute Error)을 측정해서 성능을 평가한다. 평균절대오차율은 예측된 평가 값이 최소화되어야 정확도가 높다고 할 수 있다. 대상집합의 실제 평가 값을 $\{r_1, \ldots, r_n\}$이라 한다면, 예측 값은 $\{p_1, \ldots, p_n\}$으로 표현하고, 오차 $E=\{e_1, \ldots, e_n\}=\{p_1-r_1, \ldots, p_n-r_n\}$ 이라면, 평균절대오차율은 아래 식 (5)와 같다 [8].

$$\overline{|E|} = \frac{\sum_{i=1}^{n} |e_i|}{N} \qquad (5)$$

4.3 실험결과 및 분석

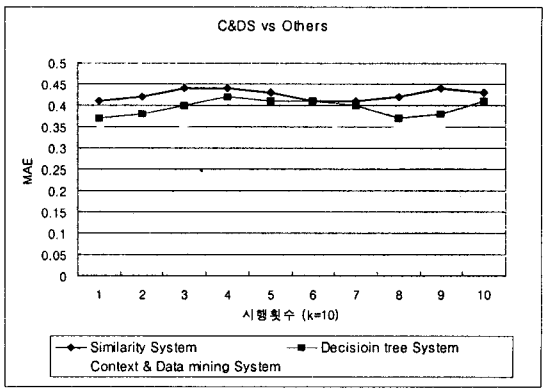
유사도 값의 최적 임계치 값을 구하기 위한 실험을 수행하였다. 샘플데이터를 학습데이터와 테스트데이터로 분류하기 위해 K-중첩 교차조사 방법을 이용하였으며 k=10 으로 설정하였으며 임계치 별로 의사결정 트리를 이용하여 평균절대오차율을 구하였다.



[그림3] 임계치에 따른 평균절대오차율 비교

그림[3]과 같이 추천을 위한 최적의 임계값은 유사도 값이 0.8일 때이다. 불필요한 학습 데이터가 많거나(임계치 0.7이상) 분석하기에 불충분한 양의 학습 데이터가 있을 경우(임계치 0.9이상) 평균절대오차율이 0.8일 때 보다 높게 평가되었다.

제안된 시스템(Context and Data mining System)의 성능을 증명하기 위해 유사도를 이용한 추천시스템(Similarity System)과 그리고 의사결정 트리 기법을 이용한 추천시스템(Decision tree System)과 성능 비교를 하였다. 제안된 시스템의 임계치는 0.8로 두고 실험하였다. 추천 시스템간의 성능 분석 결과는 [그림4]와 같고 평균절대오차율의 평균은 [표3]과 같다.



[그림4] 추천 시스템의 정확성 비교

시스템의 정확성을 나타내는 평균절대오차율의 측정치가 유사도만을 이용하거나 의사결정 트리만을 이용한 시스템보다 7.1%, 4.1% 낮게 측정되었는데, 이는 제안된 추천 시스템의 성능이 기존보다 향상되었음을 의미한다.

[표3] 추천 시스템간의 MAE

	Similarity	Decision Tree	S&D
MAE	0.425	0.395	0.354

5. 결론 및 향후 연구

과거 필터링을 이용한 추천 시스템들은 사용자의 개인정보나 상품들 간의 연관성만을 고려하는 문제점이 있었다. 본 논문은 사용자의 기분이나 주변 환경 등의 상황정보를 고려한 추천 시스템을 제안하였다. 제안하는 시스템은 정확성의 향상을 가져왔으나 상황정보에 대한 수집에 있어서 객관성이 요구된다. 또한 속성별로 주어지는 가중치에 대한 논의가 필요하며 상황정보로부터 상황을 인식하는 수준의 연구가 진행되어야 한다.

[참 고 문 헌]

[1] K. Lang, "Newsweeder: Learning to Filter News," Proceedings of the 12th International Conference on Machine Learning, Vol.7, No.4, pp.592-600, 1995.

[2] Lee, Sung-Koo, "A Recommendation System Based on Customer Preference Analysis and Filter Management," Journal of Korea Multimedia Society, 2004.

[3] B. Krulwich, "Lifestyle Finder: Intelligent User Profiling Using Large-Scale Demographic Data," Artificial Magazine, Vol.18, No.2, pp.37-45, 1997.

[4] R. Burke, "Hybrid Recommender Systems: Survey and Experiments," User Modeling and User-Adapted Interaction, Vol.12, No.4, pp.331-370, 2002.

[5] 성백균, 김상의, 박덕원, "전자상거래를 위한 사례기반의 판매지원 에이전트," 정보처리학회 제 7권, 2000.

[6] 장남식, 홍성완, 장재호, "데이터마이닝", 서울:대청미디어, 2002.

[7] J. Han, M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufman, 2001.

[8] U. Shardanand, P. Maes, "Social Information Filtering: Algorithms for Automating 'Word of Mouth'," In Proceedings of ACM, 1995.