카이스퀘어 분석과 아이템기반 협력적 여과를 이용한 타겟마케팅 기법

김완섭° 이수원
숭실대학교 컴퓨터학과 인공지능연구실
wskim92@mining.ssu.ac.kr, swlee@computing.ssu.ac.kr

Target Marketing Method on Specific Item Using Chi-Square Analysis and Item-based Collaborative Filtering

Wanseop Kim° Soowon Lee
Soongsil University Computing Department

요 약

온라인 및 오프라인 상에서 추천시스템에 대한 요구가 커지고 있으며 이에 관련해 많은 연구가 이루어지고 있다. 추천시스템은 마케팅 활용의 관점에서 목표 상품에 대한 반응 가능성이 높은 고객군을 추천하는 타겟마케팅 추천시스템과 고객 개인별로 구매 가능성이 높은 상품을 추천하는 개인화 추천시스템으로 구분할 수 있다. 지금까지의 추천시스템에 관한 연구는 대부분 개인화 추천시스템의 효율 향상에 목표를 두고 있다. 그러나 기업의 타겟마케팅에 대한 요구를 적절히 지원하지 못하고 있어 타겟마케팅에 대한 연구가 필요하다.

본 연구에서는 상품별 구매 패턴을 이용하는 프로파일 기반 추천 방법을 제안하고 이 방법과 기존의 협력적 추천 방법을 결합하여 특정 상품에 반응 가능성이 높은 고객을 추천하는 방법을 제안한다. 프로파일 기반 추천에서는 카이스퀘어 검정을 사용하여 상품별로 구매 패턴에 영향을 미치는 요인을 추출하고 이를 이용하여 특징 고객군을 선별하여 전체 고객군과 특징 고객과의 엔트로피(Entropy)의 변이 정도를 예측값으로 사용한다. 실험결과, 프로파일 기반 추천과 협력적 추천을 결합하여 추천하는 방법은 한 가지 방법을 사용할 때 보다 좋은 추천 정확도를 나타내었다.

1. 서 론

온라인 및 오프라인의 쇼핑몰에서 상품과 고객의 수가 증가하면서 고객에게 적합한 상품을 여과하여 제공하는 추천시스템에 대한 요구가 갈수록 커지고 있으며 이에 관련해 많은 연구가 진행되고 있다.

본 논문에서는 효과적인 타겟 마케팅을 위해 고객의 인구통계학적 정보를 사용하여 상품별 구매 속성을 추출하여 활용하는 프로파일 기반 추천 방법을 제안한다. 또한 제안하는 프로파일 기반 추천 방법과 기존의 개인화 추천에서 사용하는 협력적 추천 방법을 결합하여 특정 상품의 마케팅에 적합한 고객군을 추출하는 방법을 제안한다. 프로파일 기반 추천에서는 카이스퀘어 검정을 사용하여 상품별로 구매에 영향을 미치는 요인을 추출하고 이를 이용하여 특징 고객군을 선별하여 전체 고객군과 특징 고객군과의 엔트로피(Entropy)의 변이 정도를 예측값으로 사용한다. 실험 결과, 프로파일 기반 추천과 협력적 추천을 결합하는 추천 방법이 상대적으로 좋은 추천 적중률을 나타내었다.

2. 관련연구

2.1 협력적 추천

대표적인 개인화 추천 기법인 협력적 추천은 사회적 여과(Socail Filtering)라고도 하며 유사한 기호를 가지는 다른 사람들의 선호도(Rating)를 고려하여 고객이 구매하지 않은 아이템을 추천한다[6]. 협력적 추천은 희소성 문제(Sparsity Problem)와 확정성 문제(Scalability Problem)로 인한 단점이 있다.

2.2 인구통계학적 추천

인구통계학적 추천은 고객들의 인구통계학적 정보, 즉 나이, 성별, 주소, 직업 등의 개인 정보를 기반으로 추천하는 방법이다. 인구통계학적 추천 방법은 고객들의 프로파일 정보가 충분히 제공되고, 프로파일 정보가 상품의 구매에 선별력이 있을 때에만 의미있는 추천이 가능한 단점이 있다.

2.3 내용기반 추천

내용 기반 추천은 추천하고자 하는 상품 또는 정보 자체의 내용과 사용자 프로필 간의 유사성을 고려하여 추천하는 방식이다. 그러나 내용기반 추천 방법은 아이템이 반드시 분류하기에 충분한 정보를 제공해야 하고, 고객들의 구매 및 선호 정보를 활용하지 못하는 단점이 있다.

3. 제안 방법

3.1 카이스퀘어 검정을 이용한 상품별 구매 속성 선택

본 연구에서는 특정 상품과 인구 통계학적 속성과의 연관성 여부를 검정하기 위해서 카이스퀘어 검정을 사용하였다. 카이스퀘어 검정을 이용하면 범주형 자료에 대해 속성 간의 상호 독립성(또는 상호 연관성) 여부를 검정할 수 있다.

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \qquad (식1)$$

카이스퀘어 독립성 검정을 상품과 프로파일간의 연관성에 적용하는 경우 범주를 {구매, 비구매}의 두 가지로 구분할 수 있으므로 자유도가 1인 카이스퀘어 분포를 따르는 식으로 볼 수 있다. 카이스퀘어 검정식을 상품에 대한 고객의 속성과 연관성 검증에 사용하는 경우 (식2)와 같은 방법으로 할 수 있다.

$$\chi^2 = \frac{(Num(Buy) - e_{buy})^2}{e_{buy}} + \frac{(Num(\sim Buy) - e_{\sim Buy})^2}{e_{\sim Buy}} \qquad (식2)$$

각 상품들은 용도에 따라 구매되는 패턴이 다르게 나타난다.

예를 들어 어떤 상품은 20대 여성인 고객들에게 주로 구매되고, 또 어떤 상품은 직업이 학생이고 남성인 고객들에 의해 주로 구매된다. 이러한 상품의 구매 패턴을 찾기 위해서 각 상품에 대하여 카이스퀘어 검정식(식2)을 사용하여 상품별로 구매에 영향을 미치는 인구통계학적 속성을 추출할 수 있다.

범주는 {구매, 비구매}의 두 가지로 구분되므로 자유도는 1이며 독립성 판별을 위한 유의수준 α의 값을 지정하여 $\chi^2_{1,\alpha}$의 값을 기준으로 속성과 상품구매의 연관성을 판별할 수 있다. 본 연구의 실험에서는 유의수준 0.05에서 상품별로 모든 속성과의 연관성 여부를 검정하였다. 채택되지 않은 속성 정보는 프로파일 기반 추천에서 사용되지 않는다. 즉, 상품의 구매에 영향을 미치지 않는 속성은 추천 시 사용되지 않는다.

예를 들어, 여성용 핸드백인 상품 A의 성별에 대한 구매 분포가 아래의 <표1>과 같다고 할 때 카이스퀘어 검정은 이 상품의 고객 프로파일에 의한 상품 구매 패턴을 파악한다.

<표1> 상품 A의 성별에 대한 구매 분포

성 별	남 성	여 성	미기입	합계
구 매	15	30	5	50
비 구 매	210	170	25	450
합 계	220	200	30	500

<표2> 상품 A의 성별에 대한 카이스퀘어 분석 결과

	관측 구매 고객수	예측 구매 고객수	관측 비구매 고객수	예측 비구매 고객수	카이스퀘어	의미
남성	15	22	205	198	2.474	비선호
여성	30	20	170	180	5.555	선호
미기입	5	3	25	27	1.481	무상관

위의 <표1>의 데이터에 대하여 카이스퀘어 분석을 적용하면 <표2>과 같은 결과를 얻을 수 있다. <표2>은 성별 속성의 모든 속성 값들에 대해 카이스퀘어 분석을 수행한 결과를 보여준다. 유의수준 0.05에서 검정할 경우, 남성인 경우 선호하지 않고, 여성인 경우 선호하는 구매 패턴을 파악할 수 있다. 미기입의 경우 카이스퀘어 값이 $\chi^2_{0.05,1}$(=3.841)보다 작으므로 상품의 구매와 상관성이 없는 것으로 볼 수 있다. 이와 같이 각 상품들의 고객 속성에 대한 구매 패턴을 분석하여 모델로 저장하고 이를 프로파일 기반 추천 시 활용하게 된다.

3.2 프로파일 기반 추천 방법

3.2.1 프로파일 기반 추천의 개념

일반적으로 인구통계학적 추천이란 고객의 나이, 성별, 주소, 직업 등과 같은 인구통계학적 정보를 이용하여 특정 상품에 대한 선호도를 예측하는 방법이다. 본 논문에서는 고객의 인구통계학적 정보를 이용하는 추천 과정을 수식화하여 추천에 적용하였다. 제안한 프로파일 기반 추천은 고객의 인구통계학적 속성 정보, 즉 고객의 프로파일에 의하여 해당 상품에 대한 선호도를 예측하는 추천 기법으로서 3.2절에서 추출한 상품별 속성 정보를 기반 모델로 사용한다. 특정 고객에게 목표 상품에 대한 선호도를 예측하고자 할 경우 고객의 프로파일 정보를 입력으로 하여 고객의 목표 상품에 대한 특징 고객군을 선정하고 전체 고객군과 선정된 특징 고객군의 구매, 비구매에 대한 엔트로피(Entropy)의 변이 정도를 계산하여 선호도를 예측하는 추천 방법을 사용하였다.

3.2.2 특징 고객군 선정

특징 고객군이란 특정 고객의 속성 값 중 목표 상품의 구매에 영향을 주는 속성들에 속한 고객들을 의미한다. 이 특징 고객군은 해당 고객의 목표 상품에 대한 선호도를 예측할 수 있는 고객군으로 인식할 수 있다. 고객별로 추출되는 속성들은 상품에 따라 다르게 적용되며, 상품별로 특징 고객군도 다르다. 이렇게 획득된 특징 고객군의 목표 상품에 대한 구매, 비구매 정보는 해당 고객에 대한 구매 성향을 대표하게 되어 그 고객의 상품에 대한 선호도를 예측할 수 있는 정보로 활용될 수 있다. 백화점 데이터에서 '이동수'라는 골프웨어 상품에 대하여 3.1의 카이스퀘어 검정을 수행하여 상품의 구매 속성을 추출한 결과는 아래의 <표2>과 같다.

<표3> '이동수' 골프웨어에 대한 카이스퀘어 분석 결과

선택된 프로파일 속성값	속성에 소속된 고객수	소속 고객 중 구매 고객수	카이-스퀘어	소속 고객군의 구매율	선호/비선호 구분
나이 50대	1420	91	69.48	0.0640	선호
나이 60대	492	41	56.34	0.0833	선호
회원타입(1)	2260	119	51.96	0.528	선호
직업(전문직)	135	16	41.22	0.1185	선호
등록점(본점	1287	95	39.81	0.0738	선호
결혼(미혼)	3229	42	26.01	0.0130	선호
주택(본인소유)	3207	134	23.41	0.0417	선호
나이(20대)	1787	17	22.02	0.0095	비선호
주소(강남구)	706	39	19.77	0.0552	선호

<표3>를 통해 이동수 골프웨어는 나이가 50대, 60대인 고객에게 선호되고, 직업이 전문직인 고객에게 선호되는 특성들을 볼 수 있다. 결혼이 미혼인 경우와 나이가 20대인 경우에는 상품의 구매 여부와 연관성이 있으나 구매율을 비교해 볼 때 해당 속성의 경우 선호하지 않는 경향이 있고 그 외에 성별, 가입형태, 주거형태 등의 속성은 이 상품의 구매와 연관성이 적음을 알 수 있다. 고객 C의 속성이 {남성, 50대, 기혼, 전문직, 부산점}와 같을 때 이 고객에게 상품 '이동수' 골프웨어를 추천하고자 할 경우 이 고객의 속성 중 상품 구매에 영향을 주는 특징 속성 집합을 추출하면 {50대, 전문직}의 속성 집합을 얻게 된다.

3.2.3 엔트로피 변이에 의한 추천

엔트로피(Entropy)란 해당 집합에서의 무질서도를 의미한다. S라는 집합이 존재하고, 이 집합에 속하는 레코드들이 Positive와 Negative의 두개의 클래스로 구분된다고 하면 집합 S의 엔트로피는 아래의 (식3)과 같이 정의할 수 있다. 여기서 p는 전체 고객 중 Positive인 레코드가 차지하는 비율을 의미한다.

$$Entropy(S) = -plog(p) - (1-p)log(1-p) \qquad (식3)$$

엔트로피는 고객군의 집합에서 고객들의 구매 여부를 기준으로 했을 때 구매와 비구매 고객이 섞여있는 정도를 의미한다. 해당 고객군집에서 구매나 비구매 어느 한가지가 많이 나타날수록 엔트로피는 낮아지며 구매 고객수와 비구매 고객수가 동일할 때 가장 큰 엔트로피를 값을 갖는다. 본 연구에서 제안하는 프로파일 기반 추천에서는 전체 고객군에서의 특정 상품의 구매에 대한 엔트로피를 계산하고, 선택된 특정 고객군에서의 엔트로피를 계산한다.

$$P_{profile}(a,i) = Entropy_{profile} - Entropy_{total} \qquad (식4)$$

이 식에서 $P_{profile}(a,i)$는 프로파일에 의한 고객 a의 상품 i에 대한 예측 선호도이고, $Entropy_{profile}$는 적합 고객군의 상품 i의 구매에 대한 엔트로피이며, $Entropy_{total}$는 전체 고객군에서의 상품 i에 대한 엔트로피이다.

3.3 협력적 추천의 적용 방법

본 논문에서는 특정 상품에 대한 타켓 마케팅을 목표로 하기 때문에 개인화 추천에서와 동일한 유사도 계산식과 예측식을 사용하지만 이 식을 적용하는 방식에 약간의 차이가 있다. 타겟 마케팅은 특정한 상품에 대해 마케팅에 적합한 고객들을 찾는 과정이기 때문에 마케팅 하고자 하는 상품을 고정하고 모든 고객에 대하여 적용하여 예측값을 계산하였을 때 높은 예측값을 갖는 고객을 우선순위로 하여 찾을 수 있다.

3.4 프로파일 기반 추천과 협력적 추천의 결합

본 연구에서는 추천의 정확도를 높이기 위해 프로파일 기반 추천과 협력적 추천의 결과를 결합하였다. 결합하고자 하는 협력적 추천은 0에서 1사이의 값을 가지며, 프로파일 기반 추천은 -1에서 1사이의 값을 갖는다. (식5)는 협력적 추천에 의해 얻어진 예측값을 고객의 프로파일을 고려하여 보정해 주는 역할을 한다.

$$P(a,i) = P_{cf}(a,i) + P_{profile}(a,i) \times (1 - P_{cf}(a,i)) \qquad \text{(식5)}$$

제 4 장 실험 및 분석

4.1 실험 데이터

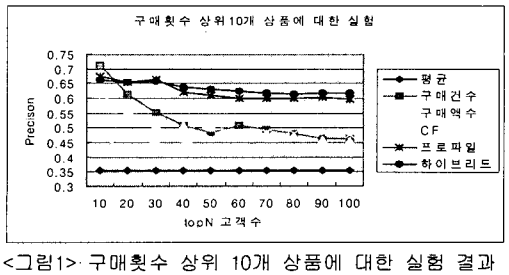
본 연구에 대한 실험 데이터로 백화점의 데이터를 사용하였다. 백화점 거래 데이터는 고객ID, 구매일, 구매지점, 거래파트, 브랜드명, 상품명, 거래액 등의 정보로 구성되어 있다. 또한 백화점 데이터는 거래 내역 이외에 고객들의 프로파일 정보를 제공하고 있다. 고객 데이터는 고객ID, 성별, 나이, 대분류 주소, 중분류 주소, 소분류 주소, 직업, 거주형태, 주택소유여부, 결혼여부, 회원등록일, 회원등록형태, 등록점 등의 정보로 구성되어 있다.

4.2 실험 방법

본 연구에서는 제안한 고객 추천 방법을 평가하기 위하여 정확도(Precision)를 사용하였다. 정확도는 $topN$에 의해 추천된 고객 중에서 실제 구매한 고객의 비율을 의미한다. 추출하는 고객 수, $topN$은 10에서 100까지 10명의 단위로 실험을 하였다.

4.3 실험 결과

제안하는 방법의 효율성 검증을 위하여 구매횟수가 높은 상품 상위 10개에 대하여 정확도(Precision)를 측정하고 다른 방법들과 비교하였다. 실험 결과는 <그림1>과 같다.



<그림1> 구매횟수 상위 10개 상품에 대한 실험 결과

각 상품의 실험에서 상품에 따라 CF(협력적 추천)와 Profile(프로파일 기반 추천)의 결과의 정확도에 차이가 있었으나 대체로 결합하였을 때의 결과인 Combine이 각각의 결과보다 향상된 정확도를 나타내었다. 이 백화점 데이터에서는 CF와 Monetary(구매액수), Frequency(구매건수 추천)의 결과가 50 이상으로 충분히 커질 때 유사한 성능을 보이고 Profile(프로파일 기반 추천)이 이들보다 높은 정확도를 나타낸다.

제 5 장 결론 및 향후과제

본 논문에서는 상품별 구매 패턴을 이용하는 프로파일 기반 추천 방법을 제안하였고, 이 방법과 기존의 협력적 추천 방법을 결합하여 추천 정확도를 향상 시키는 추천 방법을 제안하였다. 프로파일 기반 추천에서는 상품별로 구매에 영향을 많이 주는 인구 통계학적 특성을 추출하여 상품별 구매 패턴 모델을 생성하였다. 상품별 모델 생성 시 카이스퀘어 검정 방법을 이용하므로 신뢰도 범위 안에서 상품 구매에 영향을 미치는 인구 통계학적 속성을 추출하였다. 생성된 상품의 모델에 고객의 프로파일 정보를 적용하여 선호도를 예측하여 추천하였을 때 좋은 성능을 나타내었다.

기존의 협력적 추천과 프로파일 기반 추천 방법을 결합하였을 때 한가지 방법만을 사용할 때보다 높은 정확도를 나타내었다. 결합 방법은 한 가지의 추천 방법만을 적용할 때 나타날 수 있는 오류들을 보정하여 전체적인 추천 정확도를 높일 수 있었다. 또한 이러한 추천 방법은 기존의 개인화 추천에서 적절히 지원하지 못하던 특정 상품을 마케팅하기 위해 고객군을 추출하는 타겟 마케팅을 지원한다.

6. 참고문헌

[1] Sarwar, B., Karypis, G., Konstan, J., Riedel, J., "Analysis of Recommendation Algorithms for E-Commerce" GroupLens Research Group and Army HPC Research Center Department of Computer Science and Engineering University of Minnesota Minneapolis, MN 55455.

[2] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedel, J., "GroupLens: An Open Architecture for Collaborative Filtering of Netnews", In proceedings of CSCW Chapel hill, NC, 1994.

[4] Sarwar, B., Karypis, G,, Konstan, J., Riedl, J., "Item-based Collaborative Filtering Recommender Algorithms", Accepted for publication at the WWW10 Conference, 2001.

[5] Pazzani, J., "A Framework for Collaborative, Content-Based and Demographic Filtering", Artificial Intelligence Review 13(5-6): pages 393-408, 1999.

[7] Balabanovic, M., Shoham, Y., "Fab: Content-Based, Collaborative Recommendation", 1997.

[9] Sarwar, B., Karypis, G., Konstan, J., Riedel, J., "Item-Based Collaborative Filtering Recommendation Algorithms", GoupLens Research Group/Army HPC Research Center, 2000.

[12] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M., "Combining Content-Based and Collaborative Filters in an Online Newspaper.", In Proceedings of the ACM SIGIR'99 Workshop on Recommender Systems: Algorithms and Evaluation. University of California, Berkeley, 1999.

[13] 송문섭, 허문열, "수리통계학", 박영사, 2002.