다중 속성 기반 다단계 클러스터링을 이용한 이웃 선정 방법

김택헌°, 양성봉
연세대학교 컴퓨터과학과
e-mail: {kimthun, yang}@cs.yonsei.ac.kr

Neighbor Selection Methods Using Multi-Attribute Based Multi-Level Clustering

Taek-Hun Kim°, Sung-Bong Yang
Dept. of Computer Science, Yonsei University

요 약

추천시스템은 일반적으로 협동적 필터링이라는 정보 필터링 기술을 사용한다. 협동적 필터링은 유사한 성향을 갖는 다른 고객들이 상품에 대해서 매긴 평가에 기반하기 때문에 고객에게 가장 적합한 유사 이웃들을 적절히 선정해 내는 것이 추천시스템의 예측의 질 향상을 위해서 필요하다.

본 논문에서는 다중 속성 정보를 기반으로 한 다단계 클러스터링을 통한 이웃선정 방법을 제안한다. 이 방법은 대규모 데이터 셋에서 탐색 공간을 줄이기 위해 클러스터링을 수행하여 적절한 이웃 고객들의 집합을 검색하여 추출한다. 이 때, 다중 속성 정보에 따라 단계적으로 클러스터링을 수행함으로써 보다 정제된 고객 집합을 구성할 수 있도록 한다. 본 논문에서는 고객 선호도와 위치 정보 및 아이템의 선호도와 위치 정보를 대표적인 속성 정보로 사용함으로써 모바일 환경에서 보다 정확한 추천이 이루어질 수 있도록 한다.

1. 서 론

온라인 및 오프라인 상거래 시장에는 셀 수 없을 만큼 많은 종류의 상품들이 취급되고 있기 때문에 다양한 선호도를 갖는 고객들이 이들 중에서 그들이 원하는 것을 찾기 위해 많은 탐색 비용을 들이고 있다. 따라서 고객들에게 그들이 원하는 상품을 찾기 위해 들이는 탐색 비용을 줄일 수 있도록 해 주기 위해서는 더 좋은 가치를 갖는 양질의 정보를 제공하는 개인화된 추천 시스템의 개발이 필요하다.

추천 시스템은 고객의 선호도를 추출하고 분석하여 고객에게 적합한 상품을 정확하게 예측하여 추천해 줄 수 있어야 하며, 이를 위해 일반적으로 협동적 필터링(Collaborative Filtering)이라고 하는 정보 필터링 기술을 사용한다. 협동적 필터링은 상품에 대한 고객들의 선호도 상관 관계에 따른 고객들간 선호도의 유사도를 구하고 이를 예측식에 이용하여 상품에 대한 추천 여부를 결정한다. 협동적 필터링이 유사 선호도를 갖는 이웃 고객들의 평가에 근거하기 때문에 고객에게 가장 적합한 유사 이웃들을 적절히 선정해 내는 것은 추천 시스템에서 예측의 질 향상을 위해 필요하다.

이웃선정 방법은 모든 고객을 이웃으로 하여 추천을 위한 어떤 고객에게 의미 없는 고객들마저도 그 고객의 이웃 고객으로 삼는 전통적인 협동적 필터링의 단점을 보완하여 전체 고객 집단에서 의미 없는 고객들은 걸러내고 의미 있는 고객들을 찾아 이들을 이웃으로 선정함으로써 예측의 정확도를 높일 수 있게 하는 방법이다. 이웃선정 방법 중에서도 특히 클러스터링을 이용한 방법은 대규모의 데이터 셋으로부터 보다 빠른 예측과 이에 따른 추천이 이루어질 수 있도록 하는 방법으로서 큰 의미를 갖고 있다[1][3][4].

본 논문에서는 다중 속성 정보를 기반으로 한 다단

계 클러스터링을 이용한 이웃선정 방법을 제안한다. 이 방법은 대규모 데이터 셋에서 탐색 공간을 줄이기 위해 클러스터링을 수행하여 적절한 이웃 고객들의 집합을 추출한다. 이 때, 다중 속성 정보에 따라 단계적으로 클러스터링을 수행함으로서 보다 정제된 고객 집합을 구성할 수 있도록 한다.

본 논문에서는 고객 선호도와 고객의 현재 위치 정보를 대표적인 속성 정보로 사용함으로써 모바일 환경에서 보다 정확한 추천이 이루어질 수 있도록 한다. 속성을 기반으로 한 다단계 클러스터링에 의해 구성된 이웃 고객들은 그 자체로써 최종적인 이웃 고객 집합으로 선정할 수 있고 또한 이들에 대해서 다시 threshold 값에 의해 2차적으로 고객 집합을 추가 정제하여 이들을 최종 이웃으로 구성하는 것도 가능하다.

논문에서는 또한 고객에 대한 속성 정보뿐만 아니라 아이템에 대한 속성을 함께 고려한다. 다시 말해서 아이템의 선호도 속성과 아이템이 위치한 현재 위치 정보를 아이템에 대한 대표 속성 정보로 사용한다. 고객에 대한 속성 정보에 더하여 아이템에 대한 속성 정보, 특히 위치 정보 속성을 추가적으로 사용함으로써 모바일 환경에서 보다 정교한 이웃 고객을 선정할 수 있도록 하여 추천 시스템의 예측 품질을 보다 높일 수 있도록 한다.

본 논문의 구성은 다음과 같다. 2장에서 협동적 필터링에 대해서 설명하고, 3장에서 클러스터링 기반 이웃 선정 방법에 대해서 설명한다. 4장에서는 다중 속성 정보 기반 다단계 클러스터링을 통한 이웃 선정 방법을 나타내고, 5장에서 결론을 맺는다.

2. 협동적 필터링

협동적 필터링은 각 아이템에 대한 고객의 선호도로부터 고객의 프로파일을 생성함으로써 아이템을 추천한다. 협동적 필터링에서 선호도는 일반적으로 고객에 의해 평가된 수치 값으로 표현된다. 테스트 고객에게 어떤 새로운 아이템에 대한 선호도를 예측하는 것은 아이템에 대한 다른 고객(이웃)들의 평가에 기반 한다.

협동적 필터링에서 식(1)은 고객의 선호도를 예측하기 위해 사용된다. 여기에서 $w_{a,k}$는 피어슨 상관계수를 나타낸 것으로 식(2)에서 주어진 것처럼 유사 가중치를 말한다[2][3].

$$p_{a,i} = \bar{r_a} + \frac{\sum_k (w_{a,k} \times (r_{k,i} - \bar{r_k}))}{\sum_k |w_{a,k}|} \quad (1)$$

$$w_{a,k} = \frac{\sum_j (r_{a,j} - \bar{r_a})(r_{k,j} - \bar{r_k})}{\sqrt{\sum_j (r_{a,j} - \bar{r_a})^2 \sum_j (r_{k,j} - \bar{r_k})^2}} \quad (2)$$

위 식에서 $p_{a,i}$는 아이템 i에 대한 고객 a의 선호도를 나타낸다. r_a와 r_k는 고객 a의 평가와 고객 k의 평가에 대한 평균을 각각 나타낸다. $r_{k,i}$와 $r_{k,j}$는 아이템 i와 j에 대한 고객 k의 평가를 각각 나타내고, $r_{a,j}$는 아이템 j에 대한 고객 a의 평가를 나타낸다.

만약 고객 a와 k가 한 아이템에 대해서 유사한 평가를 가진다면, $w_{a,k}>0$이다. $|w_{a,k}|$는 두 고객이 이미 평가한 아이템에 대해서 고객 a가 고객 k에 얼마나 동의하는지를 나타낸다고 할 수 있다. 만약 두 고객이 한 아이템에 대해서 반대되는 평가를 내렸다면, $w_{a,k}<0$ 이고 $|w_{a,k}|$는 그들이 같은 아이템에 대해서 얼마나 동의하지 않는지를 나타낸다고 할 수 있다. 그러므로 만약 그들이 상관성이 없다면, $w_{a,k}=0$ 이다. $w_{a,k}$는 -1에서 1 사이의 값을 가진다.

협동적 필터링은 고객과 비슷한 선호도를 갖는 다른 고객들의 평가를 기반으로 하기 때문에 추천 시스템에 적합하다. 그러나 비록 협동적 필터링이 추천 시스템을 위한 좋은 선택이라고 여겨질 수 있지만, 예측의 질을 향상시키기 위한 여지가 여전히 많이 남아 있다. 이를 위해서 협동적 필터링은 유용한 이웃선정 방법이 필요하다.

3. 클러스터링 기반 이웃 선정 방법

클러스터링 방법은 서로 유사한 선호도를 갖는 고객들로 구성된 k개의 클러스터를 만든다[1][3][5]. 많은 클러스터링 방법 중에서도 k-means 클러스터링은 대표적인 클러스터링 방법으로 수치 데이터를 클러스터링하는데 매우 유용하다. k-means 클러스터링은 먼저 임의로 k명의 고객들을 k개의 클러스터에 대한 초기 중심점으로 선택한다. 그런 후에 모든 고객은 클러스터의 중심과 고객 사이의 거리가 최소가 되는 하나의 클러스터에 할당한다. 거리는 Euclidean distance를 사용하여 계산

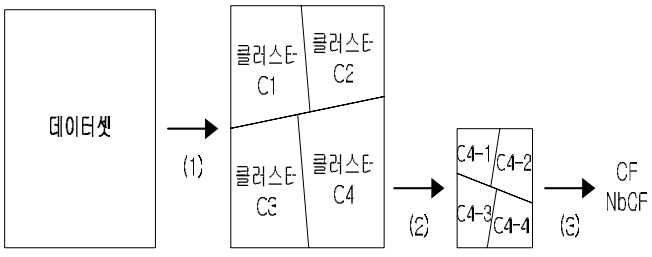
하게 되는데 이것은 고객과 각 중심점 사이의 각 속성의 차에 대한 제곱의 합에 대한 제곱근으로 구할 수 있다. 이 때, 거리 측정을 위해서는 Euclidean 거리 대신에 식(2)에 보이는 것처럼 피어슨 유사 상관 계수를 사용하여 구할 수 있다.

그 후에 각 클러스터에 대해서 클러스터에 현재 속한 고객들을 기반으로 클러스터의 평균을 다시 계산한다. 이 평균은 클러스터의 새로운 중심으로 고려된다. 새로운 중심을 찾은 후에 고객이 속해야 하는 클러스터를 찾기 위해 각 고객에 대한 거리를 계산한다. 평균을 재계산하고 거리를 계산하는 것은 종료조건을 만나게 될 때까지 반복된다. 종료조건은 일반적으로 모든 새로운 중심이 이전 중심으로부터 각각 얼마나 움직였는가 하는 것이다. 만약 모든 새로운 중심이 어떤 한계 거리 내에서 움직였다면 우리는 반복을 종료한다.

클러스터링에 의해 최종적으로 K개의 클러스터와 각 클러스터에 속하는 고객 집합이 결정되면 각 클러스터에 속한 고객 집합은 모두 후보 이웃이 된다. 이후 선호도를 예측하고자 하는 고객의 상품 속성에 대한 선호도와 각 클러스터의 대표 값이 지니는 속성 선호도 사이의 거리를 계산하여 가장 최소의 값을 갖는 클러스터를 선정한다. 이렇게 결정된 클러스터에 속하는 고객 집합이 최종적인 이웃 고객으로 선정된다.

클러스터링 기반 이웃선정 방법은 클러스터링 이후 예측하고자 하는 고객의 선호도와 가장 유사한 클러스터내의 고객들만 예측을 위한 이웃으로 결정하기 때문에 대용량 데이터 셋의 경우 빠른 예측을 통한 추천이 가능하게 되는 장점이 있다.

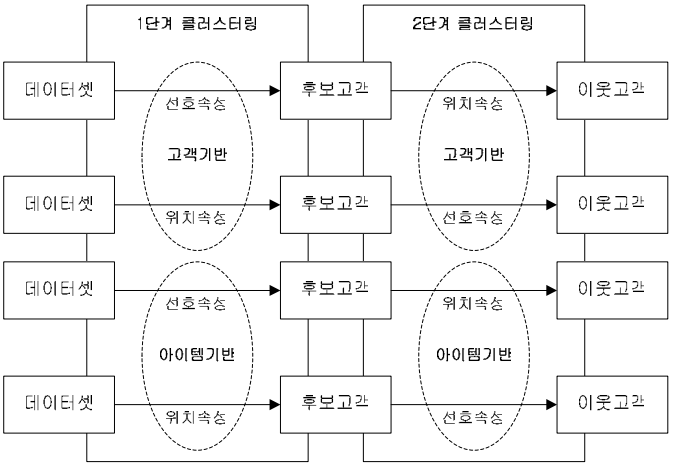
4. 다중속성기반 다단계 클러스터링을 이용한 이웃선정



(그림 1) 다단계 클러스터링

그림 1은 다단계 클러스터링을 통해 이웃을 선정하는 방법을 보여주고 있다. 클러스터링은 먼저 주어진 원시 데이터 셋 혹은 가공된 데이터 셋을 입력으로 하여 1단계 클러스터링을 수행한다(그림에서 (1)). 그림은 1단계 클러스터링 이후에 C1에서 C4까지 4개의 클러스터가 생성된 것을 보여주고 있다. 다음으로 4개의 클러스터 중에서 이웃 고객의 후보로써 하나의 클러스터를 선정한 후에(그림에서 C4) 다시 2단계 클러스터링을 수행한다(그림에서 (2)). 그림은 2단계 클러스터링 이후에 C4-1에서 C4-4까지 4개의 클러스터가 새로 생성된 것을 보여주고 있다. 이렇게 생성된 클러스터들은 다시 협동적 필터링(CF)에서 사용할 최종적인 이웃 고객을 선정하기 위해 가장 최적의 클러스터 하나를 이웃 고객으로 결정한다. 이렇게 하여 선정된 이웃 고객은 그 자체로써 협동적 필터링에서 예측을 수행하기 위한 고객 집단으로 이용할 수 있고 또한 이들 고객 집단을 threshold를 이용한 이웃 선정 방법 등으로 2차적으로 보다 정제된 이웃 고객으로 선정하는 것이 가능하다 (NbCF: Neighbor-based CF).

그림 1에서 각 단계별 클러스터링은 속성 정보를 이용하여 수행한다. 여기에서 속성은 고객의 선호도 속성과 위치 속성 그리고 아이템의 선호도 속성과 위치 속성을 사용한다. 그림 2는 이러한 속성들을 이용한 다중 속성 기반 다단계 클러스터링을 이용한 이웃 선정 방법을 보여주고 있다.



(그림 2) 다중속성기반 다단계 클러스터링

위치 속성은 모바일 환경에서 주어진 고객에 대해서 그가 속한 일정 지역 내에 있는 유사 선호 고객을 이웃 고객으로 선정하고자 할 경우에 우선적으로 클러스터링에 사용한다. 반대로 주어진 고객에 대해서 그와 유사한 선호도를 가진 다른 고객들을 우선적으로 고려하여 클러스터링을 수행한 이후에 고객이 속한 위치 속성을 이용하여 다시 클러스터링을 수행하여 이웃 고객을 선정

할 수도 있다.

이것은 아이템에 대해서도 마찬가지로 이용할 수 있다. 다시 말해서 아이템이 가진 위치 속성은 이 아이템이 일정 지역 내에서 서비스되고 있음을 나타낸다. 따라서 일정 지역 내에 있는 아이템에 대해서 이 아이템을 선호하는 고객들을 이웃 고객으로 선정하고자 할 경우에 클러스터링을 수행한다. 반대로 주어진 고객에 대해서 이 고객이 선호하는 아이템들을 우선적으로 고려하여 클러스터링을 수행한 이후에 아이템이 속한 위치 속성 정보를 이용하여 다시 클러스터링을 수행하여 이들을 선호하는 이웃 고객을 선정할 수도 있다.

그림 2에서 주어진 데이터 셋은 1단계 클러스터링을거쳐 후보 이웃군을 선정하기 위한 클러스터들을 생성한다. 여기서 데이터 셋은 원시 데이터 셋 혹은 가공된 데이터 셋을 말한다. 그 다음에 다시 2단계 클러스터링을 통해 최종 이웃 고객들을 선정하기 위한 클러스터들을 생성한다. 이 때 클러스터링 각 단계에서는 서로 다른 속성을 이용하여 클러스터링을 수행한다. 즉, 1단계 클러스터링에서 선호도 속성을 이용했다면 2단계에서는 위치 속성 정보를 이용하여 클러스터링을 수행하는 것이다. 마찬가지로 1단계 클러스터링에서 위치 속성을 이용하여 클러스터들을 생성하고 다시 최종 이웃 고객을 선정하기 위해 2단계 클러스터링을 수행한다면 이 때에는 선호 속성을 이용하여 클러스터링을 수행한다.

고객에 대한 속성 정보로 대표적인 것으로는 고객의 상품에 대한 선호도를 들 수 있으며, 모바일 환경하에서는 고객의 위치 정보 속성이 추가로 필요하다. 또한 이들 속성 정보는 아이템에 대해서도 마찬가지로 이용할 수 있다. 즉, 아이템이 가진 다양한 속성 정보뿐만 아니라 아이템이 위치한 혹은 아이템이 서비스되고 있는 지역 정보, 즉 위치 데이터 속성이 모바일 환경에서는 중요한 속성 정보로 보다 정확한 예측을 위해 필요하다.

그림에서 클러스터링 1단계에 의해 얻어진 클러스터들은 가장 적합한 클러스터를 선정하여 이웃 고객의 후보 집단으로 정하고 다시 2단계 클러스터링의 입력 데이터 셋으로 사용한다. 이어서 2단계 클러스터링의 수행 결과로 나온 결과에 대해 다시 가장 적합한 하나의 클러스터를 선정하여 이를 이웃 고객 집단으로 결정한다.

이렇게 다중 속성을 기반으로 다단계 클러스터링을 거쳐 나온 최적의 클러스터 하나는 예측을 수행하기 위한 이웃 고객들로 이들을 직접 협동적 필터링의 이웃 고객으로 사용하거나, 추가적으로 다른 이웃 선정 방법을 이용하여 보다 정제된 고객을 최종적인 이웃으로 선정하여 협동적 필터링에서 향상된 예측을 위해 사용한다.

5. 결론

추천 시스템은 고객의 선호도를 추출하고 분석함으로써 정확한 예측을 수행하는 능력을 갖는 것이 매우 필요하다. 협동적 필터링이 비록 추천 시스템에 폭넓게 사용되고 있지만, 이것의 단점을 극복하기 위한 노력들이 예측의 질을 향상시키기 위해 수행되어야 한다. 아울러 대규모 데이터 셋으로부터 추천을 수행하는데 따른 부담을 덜기 위한 빠른 추천에 대한 노력도 또한 필요하다. 특히 모바일 환경하에서는 지금까지 보다 더욱 다양하고 수 많은 데이터들이 서로 다른 위치에 존재하기 때문에 이러한 대규모 데이터 셋 문제의 해결이 더욱 필요하고 중요하게 된다.

본 논문에서는 다중 속성 정보를 기반으로 한 다단계 클러스터링을 이용한 이웃 선정 방법을 제안하였다. 이 방법은 대규모 데이터 셋에서 탐색 공간을 줄이기 위해 클러스터링을 수행하여 적절한 이웃 고객들의 집합을 추출한다. 이 때, 선호도 및 위치 정보에 대한 속성 정보에 따라 단계적으로 클러스터링을 수행 함으로서 보다 정제된 고객 집합을 구성할 수 있도록 한다. 또한 이러한 속성 정보는 고객에 대한 것뿐만 아니라 아이템에 대한 것을 함께 고려함으로써 보다 정교한 이웃 고객 선정을 통해 추천을 위한 예측의 질을 높일 수 있게 한다.

본 논문에서 제안한 방법을 통해 대규모 데이터 셋을 고객 및 아이템에 대한 선호 속성 및 위치 속성을 단계적으로 이용하여 모바일 환경에 적합한 데이터 셋으로 정제하고, 이들을 협동적 필터링에서 최종적인 이웃 고객 집합으로 이용할 수 있도록 함으로써 보다 빠르고 정확한 예측을 수행할 수 있다.

참고문헌

[1] B.M. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Recommender Systems for Large-Scale E-Commerce: Scalable Neighborhood Formation Using

Clustering," Proceedings of the Fifth International Conference on Computer and Information Technology, 2002.

[2] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J., "GroupLens: Applying Collaborative Filtering to Usenet News," The Communications of the ACM, Vol. 40., 1997.

[3] O'Connor M., and Herlocker J., "Clustering Items for Collaborative Filtering," Proceedings of the ACM SIGIR Workshop on Recommender Systems, 1999.

[4] T.H. Kim, S.B. Yang, "A Refined Neighbor Selection Algorithm for Clustering-Based Collaborative Filtering," Journal of KIPS, Vol.14-D, No.3, 2007.

[5] T.H. Kim, S.B. Yang, "Attribute-based Multi-level Clustering for Collaborative Filtering," Proceedings of the 28th KIPS Conference, 2007.

[6] J. Herlocker, J. Konstan, L. Terveen, and J. Riedle, "Evaluating Collaborative Filtering Recommender Systems," ACM Transactions on Information Systems, Vol.22, No.1, 2004.

[7] G. Xue, C. Lin, and Q.E. Yang, "Scalable Collaborative Filtering Using Cluster-based Smoothing," Proceedings of the ACM SIGIR Conference, 2005.