E-Commerce에서의 퍼지 클러스터링 알고리즘을 적용한 추천 시스템

Hae Ri. Lyou and Moon Hyun Kim

School of Electrical and Computer Engineering, SungKyunKwan University
Jangangu ChunChundong 300, suwon KyungGido, 440-746, South Korea
Tel:+82-31-290-7216, Fax:+82-31-290-7211
E-mail:sea810127@msn.com , mhkim@simsan.skku.ac.kr

성균관대학교 정보통신공학부
경기도 수원시 장안구 천천동 300번지 정보통신공학부, 440-746
Tel:+82-31-290-7216 , Fax:+82-31-290-7211
E-mail:seavillage99@hanmail.net, mhkim@simsan.skku.ac.kr

Abstract

인터넷의 발전으로 전 세계적으로 다양한 인터넷 서비스들이 점차 확대되고 있으며, 특히 수익을 내는 방법으로서의 인터넷 전자상거래는 큰 비중을 차지하고 있다. 이에 수많은 사이트, 쇼핑몰은 상품과 고객들의 수많은 데이터를 데이터베이스 모듈로 관리하고 있다. 이렇게 고객에게 맞는 상품을 추천하기 위해 효율적으로 클러스터링 하는 방법이 요구된다. 이에 본 논문에서는 여러 클러스터링 방법 중에서 퍼지 이론을 기반으로 개선된 클러스터링 알고리즘을 이용하여 상품을 추천하고자 한다. 이 방법은 클러스터의 개수가 한정되어 있는 기존의 방법에 클러스터의 유사도에 따른 유사성을 부여함으로써 더 세밀하고 정확한 클러스터링을 가능케 하여 이에 따른 개인의 성향에 맞게 개인화된 상품을 추천하는 시스템을 설계하고자 한다.

1. 서론

인터넷 환경의 급속한 발달과 함께 이를 이용한 전자상거래가 빠르게 증가하고 있다. 증가하는 전자상거래 환경에서 고객에게 필요한 제품을 신속히 제공하고, 제품 판매를 증가시킬 수 있는 새로운 전자상거래 시스템의 필요성이 점차 확대되어 가고 있다.

그러나 기존의 전자상거래 시스템은 고객이 요구하는 상품만을 제공하는 단순한 형태여서 고객의 요구 사항을 충분히 만족시키지 못하였다. 따라서 전자상거래 시스템들 간의 경쟁력이 강조되는 현재에서는 고객들에게 양질의 서비스를 제공하고, 보다 특성화된 기능의 다양한 해결 방안이 요구되고 있다.

이러한 요구에 따라 최근 전자상거래 시스템 연구에서는 추천 시스템에 대한 연구가 활발히 진행되고 있다. 하지만 지금까지의 추천 시스템은 고객의 구매 데이터가 증가하면 고객에게 추천을 제공하는데 많은 시간이 소요되는 단점을 가졌다.

따라서, 본 논문에서는 전자상거래 시스템의 효율성을 높이기 위해 다양하고 방대한 고객들의 성향을 분석 파악하여 고객에 맞는 추천 시스템을 제안한다. 또한 Fuzzy Clustering Algorithm을 적용하여 고객 정보, 행동 등의 데이터로부터 고객의 성향을 측정하고 이 성향 측정의 결과를 바탕으로 개인화 된 추천을 가능하게 할 것으로 기대

된다. 1]

2. 관련연구

2.1 ISODATA algorithm

K-means 알고리즘은 클러스터링에 영향을 끼치는 요인들에 대해 휴리스틱 지식을 활용한 알고리즘이다. K-means 알고리즘과 동일하게 클러스터의 중심 벡터들을 반복적으로 결정한다. {x^1,x^2,x^3,...,x^n}들이 알고리즘에 순차적으로 주어지며, 초기 클러스터들의 중심벡터인 $m^1(0)$, $m^2(0)$,....., m(0)을 초기에 임의로 설정한다. 여기서 설정되는 중심벡터의 개수인 T(0)은 최정적으로 구하고자 하는 클러스터의 개수 K와 반드시 동일한 필요는 없다.

2.2 FCM Clustering Method

FCM 클러스터링 방법은 하나의 클러스터에 속해져 있는 각각의 데이터 점을 소속 정도에 의해서 클러스터에 대한 데이터의 소속 정도를 일일이 열거한 데이터 분류 알고리듬이다. FCM 클러스터링은 n개의 벡터 xi , i=1 , ... , n 의 집합을 c개의 퍼지 그룹들로 분할하고, 비유사성 측정의 비용함수가 최소가 되는 것과 같은 각각의 그룹 안에서 클러스터의 중심을 찾는다. FCM 클러스터링 방법은 0과 1 사이의 소속 정도에 의해서 나타난 소속감의 정도를 가지고 주어진 데이터 점이 몇 개의 그룹에 속할 수 있다는 퍼지 분할을 사용한다. 즉 퍼지 분할을 적용하기 위해서, 소속함수 U 는 0과 1 사이의 값을 가지는 요소들을 가진다. 그러나 데이터 집합에 대한 소속감 정도의 함은 식 (2.1.1)과 같이 항상 1 이다.[2]

$$\sum_{i=1}^{c} u_{ik} = 1, \ \forall k = 1, \ldots, n$$

식 2.1.1

2.3 FCM clustering algorithm

[단계 1] 클러스터의 개수 c($2 \le c < n$)을 정하고 지수의 가중(exponential weight) m($1 < m < \infty$)을 선택한다.
초기 소속함수 $U^{(0)}$ 를 초기화한다. 알고리즘 반복 횟수를 r(r= 0 , 1 , 2 , ⋯)로 표시한다.
[단계 2] 식 (3.6)을 이용하여 퍼지 클러스터 중심 {Vi | i = 1, 2, ⋯ , c}을 계산한다.
[단계 3] 다음과 같이 새로운 소속함수를 식 (2.1.2)으로 계산한다.

$$U_{ik}^{(r+1)} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}^{r}}{d_{jk}^{r}} \right)^{2/m-1}} \quad \text{for } I_k = \emptyset$$

식 2.1.2

[단계 4] 다음 식을 계산해서 만일 $\Delta > \varepsilon$이면 r = r+1로 정하고 [단계 2]로 가서 다시 알고리듬을 반복 수행하고 그렇지 않고 $\Delta <= \varepsilon$ 이면 알고리듬을 종료한다. 여기서, ε 는 임계값이며, 식(2.1.3)과 같다.[3][6]

$$\Delta = \| U^{(r+1)} - U^{(r)} \| = MAX_{ik} | u_{ik}^{(r+1)} - u_{ik}^{(r)} |$$

식 2.1.3

2.4 ISODATA 와 Fuzzy의 비교

	advantage	disadvantage
ISODATA algorithm	➢자기조직(self-organizing)능력 ➢삭제(eliminate), 분할(divide), 병합(merge)의 유연성(flexibility)	➢복수의(multiple) 파라미터 필요. ➢클러스터의 개수(K)를 결정하는 값(value)은 유저에 의해 주어진 파라미터에 의존 - 최상의 값(best value)이 아닐 수 없다.
Fuzzy algorithm	➢ISODATA에 비해 적은 수의 파라미터 필요 ➢군집에 소속된 정도(weight)에 따라 적절한 군집구성	➢클러스터의 개수가 확정적 - 유연성(flexibility)의 제약. ➢Small cluster 의 처리문제.

Table2.4.1 ISODATA and Fuzzy 비교

3. 제안 시스템

데이터마이닝의 여러 기법 중에서 Clustering 기법은 전체 데이터의 분포 상태나 패턴등을 찾아내는데 유용하게 이용할 수 있다. Clustering이란 주어진 n 개의 점을 k 개의 그룹으로 나누는 것을 말하며, 분류와 다른 점은 각 클래스에 해당되는 정보가 제공되지 않는다는 것이다. 이 군집화 방법에는 여러 가지 알고리즘이 개발됐는데 알고리즘에 따라 다른 군집화를 만들어 낸다. 그러므로 모든 알고리즘의 특성을 잘 알고 있어야 자기 응용 분야에 맞는 것을 잘 사용할 수 있다. 따라서 본 장에서는 효율적인 전자 상거래에 사용되기 위한 개인화 기능을 지원하기 위해 Fuzzy Clustering 기법을 향상시킨 새로운 m-Fuzzy Clustering에 대해 제안한다.[3][4]

3.1 m-Fuzzy Clustering System

m-Fuzzy Clustering System은 기본적인 FCM(Fuzzy C-Means) Algorithm에 Small Cluster Filtering Part와 클러스터의 중심 거리와 유사도로 새로운 클러스터를 생성(Merge)하는 Part로 구성된다. 이 알고리즘을 적용하여 고객 정보, 행동(Action)등의 데이터로부터 고객의 성향을 측정하고, 이 성향 측정의 결과를 바탕으로 개인화 된 추천을 가능할 것으로 기대된다. Small Clusters Filter Part는 전체 데이터와 클러스터링 된 데이터의 비율과 다른 클러스터와의 유사도에 따라 측정 할 수 있으며, 여기서 검출된 Small Clusters는 이후의 Merge 과정에서 제외되어 Merge 과정의 효율성을 높인다.[5][7]

Merge Part는 가장 가까운 중심거리(클러스터의 유사도)를 바탕으로 하여 1 Pair를 선택한 후 두 클러스터간의 군집의 타당성(Clustering Validity)을 이용하여 수행된다. 이 군집의 타당성은 The Xie-Beni Validity를 이용하여 구해진다.

The Xie-Beni Validity는 군집의 조밀도(Compactness)와 분리도(Separatio) 측정의 결과이다.

이 Compactness-to-Separatio ratio v는 D가 초기 군집(Prototype) 사이의 거리가 가장 작을 때 다음 식(3.1.1)과 같이 정의된다.[8]

$$V = \{(1/k) \sum_{(h=1,k)} \sigma_h^2\} / \{D_{min}\}^2;$$

$$\sigma_h^2 = \sum_{(q=1,Q)} w_{qh} \, \| x^{(q)} - c^{(h)} \|^2 , \; k=1,..,k$$

식 3.1.1

3.2 전체 구성도

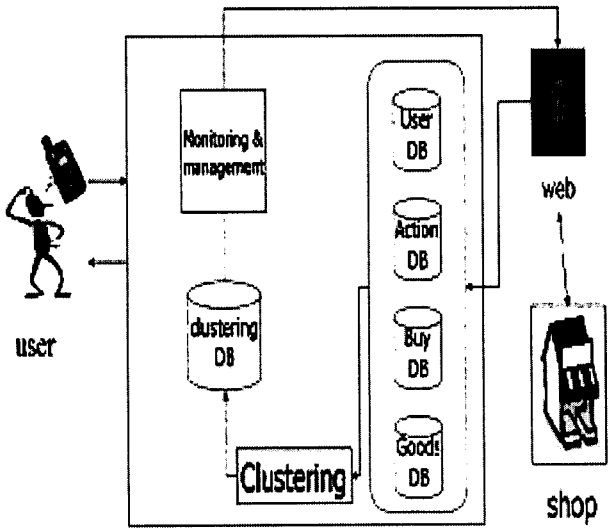


Figure3-1. m-Fuzzy clustering system의 기본 구성도

User가 유 무선 환경에서의 쇼핑몰에서 상품에 대한 선호도나 상품 구매 행위가 발생했을때, User Profile 정보는 여러 정보들을 토대로 Clustering 과정이 일어난다. 그 결과는 각각의 DB Module에 저장된 후, 새로운 User에게 클러스터링된 각각의 상품 데이터를 바탕으로 한 상품의 추천이 일어난다.

3.3 Clustering System

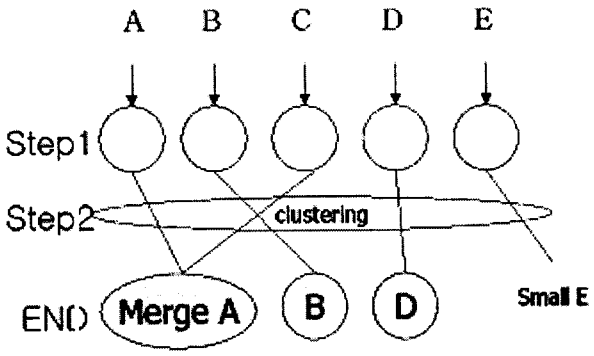


Figure3-2. Clustering Part 구성도

Step1) User, Action, Buy, Goods을 이용한 1st clustering 생성 (FCM)

Step2) Clustering의 결과에 Small Cluster Filtering 과정과 Merge 과정이 일어난다.

End) 최종 Cluster 생성

4. 구현 및 평가

본 고에서 제안하는 m-Fuzzy Clustering System은 Windows 2000 server, MSSQL server 환경에서 Visual C++6.0을 이용하여 구현하였으며, Weka 툴을 활용하였다.

데이터는 전자상거래에서의 Real Data를 활용하였다.

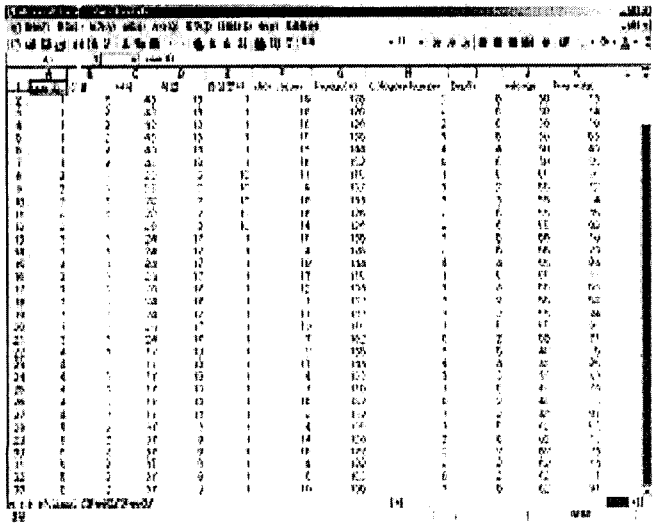


Figure4-1. 통합 테이블

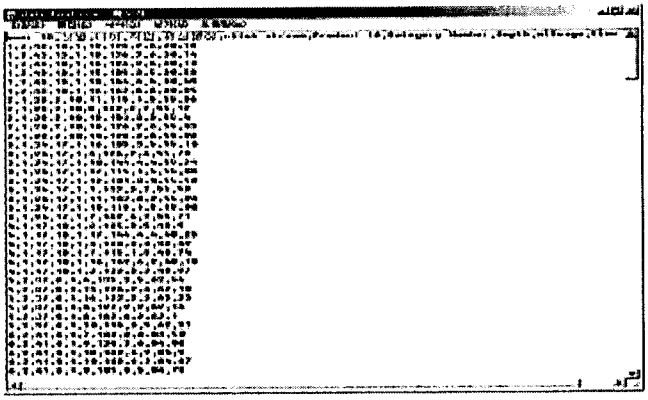


Figure4-2. Transaction File

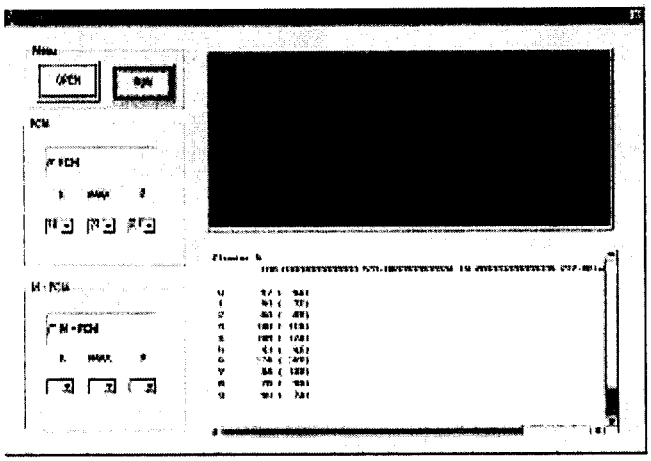


Figure4-3. m-Fuzzy Clustering Tool

4.3 Chair Data 선정 및 Simulation

Y 좌표를 나이로 고정 한 후 X 좌표를 여러 Case로 변화를 주었다. 아래 그림에서 보이는 각 Case에 대한 장바구니(Cart) 데이터와 실 구매(Buy) 데이터의 결과로서 두 데이터의 결과가 서로 유사 할 수록 어떤 상품에 대한 구매성향과 실제 구매 비율을 평가 할 수 있다.

아래의 그림을 분석한 결과 나이-직업의 Chair 데이터를 이용하는 것이 가장 신뢰도가 높다고 평가되며,나이-직업에 관한 시뮬레이션의 결과이다.

Figure4-4. 나이-클릭에 따른 구매 비율

Figure4-5. 나이-관심도에 따른 구매 비율

Figure4-6. 나이-성별에 따른 구매 비율

Figure4-7. 나이-성별에 따른 구매 비율

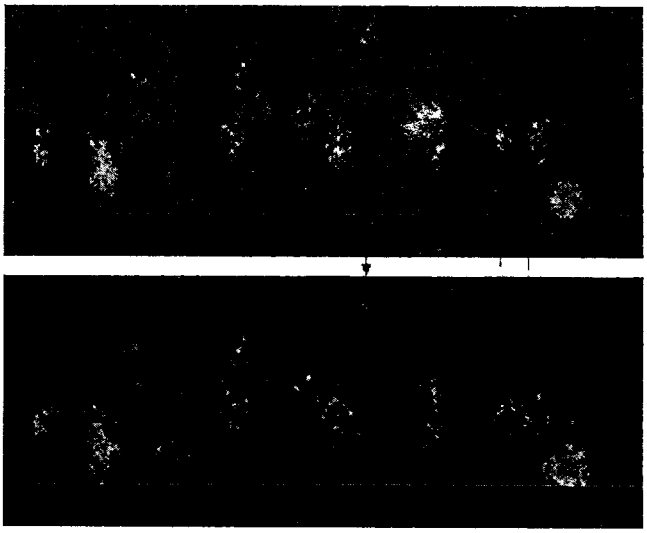
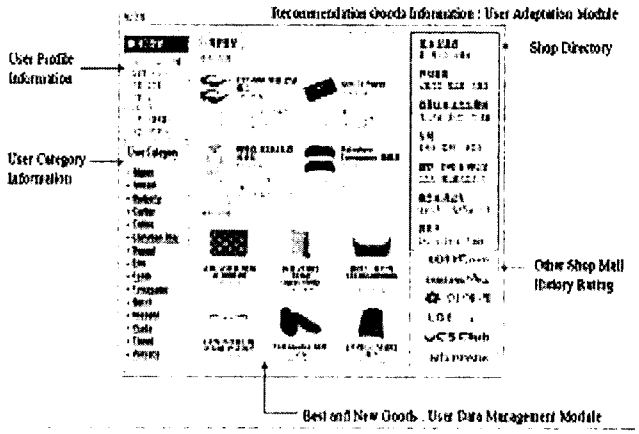


Figure4-8. 최종 결과

Figure4-8 에서 FCM 으로 수행된 각 클러스터의 결과가 Filtering and Merging 을 통해 새로운 클러스터 군집으로 표현된 것을 볼 수 있다. 이 결과를 바탕으로 보다 적은 수의 클러스터의 분석으로 유저 상품 데이터의 효율성을 높일 수 있으며, 이에 따라 아래 그림 4-9 처럼 개인의 성향에 맞추어진 개인화 추천 시스템 인터페이스를 구성하였다.



4-9. Recommendation System

5. 결론

본 고에서는 DB 에 적재된 수 많은 데이터들 중 고객정보, 행동(Action)등의 사용자 데이터로부터 개인의 성향을 측정하고, Small Cluster 의 Filtering 과정과 Cluster 의 중심 거리와 유사도로 새로운 Cluster 를 생성(Merge)하여 Recommendation System 을 제안하였다. 제안한 시스템은 여러 Cluster 기법 중 Fuzzy 기술을 변형시켜, Small Cluster 를 검출하는 과정을 통해 군집과의 유사도나 성향이 차별화 된 군집을 파악하고 이에 합병 과정(Merge Part)의 과정을 단순화였다. 또한 기존의 Fuzzy 기술의 단점인 유동적이지 못한 클러스터의 수를 자동적으로 조절하였다. 따라서 이러한 Clustering 결과를 적용하는 Personalized Recommended 부분에서 향상된 효과를 얻을 수 있다.

6. 향후 과제

본 고에서 제안하는 시스템의 향후 과제 및 발전 방향은 다음과 같다.

Small Cluster Data 에 대해 Dummy 데이터의 여부를 판단해서 삭제 또는 재 사용성의 여부를 판단하야 하는 연구가 더욱 더 필요하다. 또한 많은 Real Data 를 통해 Real World 상에서 User 에게 Recommendation / Adaptation 의 확장성을 줄 것으로 기대된다.

References

[1] Adil, G.K., Rajamani, D. and Strong, D., 1997, "Assignment allocation and simulated annealing algorithms for cell formation". IIE Transactions, 29/1, 53-67.

[2] Bezdk, J., 1981, "Pattern Recognition with Fuzzy Objective Function Algorithms", (New York: Plenum Press)

[3] Chan, H.M. and Miner, D.A., 1985, "Direct Clustering Algorithm for Group Formation in Cellular Manufacture. Journal of Manufacturing System", 1/1, 65-75

[4] S. Miyamoto and Y. Agusta, "L based fuzzy c-means for data with uncertainties," Computing Science and Statistics, Vol.29, No.2, pp.409-414, 1997

[5] W. Pedrycz, J.C. Bezdk, R.J. Hathaway and G.W. Rogers, "Two nonparametic models for fusing heterogeneous fuzzy data," IEEE Trans., on Fuzzy Syst., Vol.6, No.3, pp.411-425, 1998

[6] Chiu, S., 1994, "Fuzzy model identification based on cluster estimation". Journal of Intelligent and Fuzzy Systems, 2, 267-278

[7] M. Setnes and U. Kaymak. Extended fuzzy c-means with volume prototypes and cluster merging. In Proceeding of Sixth European Congress on Intelligent Techniques and Soft Computing, Volume2, pages 1360-1364. ELITE, Sept. 1998

[8] M. Setnes, U. Kaymak, and H. R. van Nauta Lemke. "Fuzzy target selection in direct marking." In Proceeding of CIFEr'98, New York, Mar. 1998.