한국지능정보시스템학회논문지 제9권 제1호 2003년 6월(pp.179~193)

항목 유사도를 고려한 트랜잭션 클러스터링

이상욱
한양대학교 산업공학과 석사
(dreamboy2832@hanmail.net)

김재련
한양대학교 산업공학과 교수
(jyk@hanyang.ac.kr)

군집화(clustering)는 주어진 객체들 중에서 유사한 것들을 몇몇의 집단으로 그룹화 하여 각 집단의 성격을 파악하는데, 실제적으로 각 객체가 유사한지 그렇지 않은지를 측정할 수 있는 도구가 필요하다. 기존의 군집화에서 객체간에 유사하다는 의미는 각 군집(cluster)안에 있는 객체들이 같은 속성 값이 많으면 많을수록 객체간에 유사성이 높아 유사도가 높은 객체끼리 군집을 이루게 된다는 것을 의미했다. 그 중에서도 범주형 속성을 갖는 군집화는 같은 속성 값이면 1, 서로 다르면 0으로 표현하여 유사성을 측정하는 방법이다. 제안된 알고리듬은 속성 값을 0과 1로만 표현하는 것에 대한 문제점을 제시하고 서로 다른 속성이라도 속성간에 친밀한 관계가 있다는 개념을 도입하여 어느 정도 유사한 지를 보여준다. 같은 객체간에 같은 값을 갖는 속성이 하나로 없더라도 구해진 유사도에 의해 유사한 개체끼리는 하나의 군집이 될 수 있는 알고리듬을 만든 후 그 군집에 속해 있는 고객들의 니즈와 구매 선호도에 따라 적절한 타겟 마케팅(Target Marketing)을 할 수 있다.

논문접수일 : 2002년 12월 게재확정일 : 2003년 4월 교신저자 : 이상욱

1. 서론

1.1 연구배경

군집화(clustering)기법은 통계학(statistics), 패턴 인식(pattern recognition)등의 분야에서 광범위하게 연구되어 왔다. 현재는 데이터 마이닝 분야에서 이 기법을 도입하여 응용하려는 연구가 활발히 진행되고 있다. 데이터 마이닝은 일반적으로 대용량의 데이터 베이스로부터 의사 결정에 유용한 숨어있는 패턴을 식별하는 과정이라고 할 수 있다. 현재 널리 쓰이고 있는 데이터 마이닝 기법들은 연관(association), 군집화(clustering), 분류(classification)등이 있다.

연관규칙(association rule)은 최소 지지도(minimum support)를 만족하는 항목들간의 연관성을 찾아내어 유용한 규칙들을 발견하는 기법이다. 가령, "붉은 색(x1)의 sports car(x2)를 타고 다니고 애완견(x3)이 있는 여성들의 90%는 Chanel No.5(x4)를 사용한다."에서 "x1,x2,x3 ⇒ x4"인 연관 규칙을 생각할 수 있다.

분류(classification)란 데이터 베이스에 있는 레코드들을 여러 개의 클래스들로 분류했을 때 각 클래스의 특징을 찾아내는 작업이다. 예를 들어, 기존 고객들의 신용 위험도를 분류하면 새로운 고객의 신용 경력, 현재의 부채, 담보물 및 소득에 의해 고객의 신용 위험도가 추정될 수 있을 것이다. 활발하게 연구되고 있는 분류 방법으로

는 의사 결정나무(decision tree), 신경 회로망(neural network)등이 있다.

군집화는 속성들의 값에 의거하여 유사한 속성 값을 가지는 객체들끼리 그룹핑(grouping)하는 작업이다. 현재의 군집화 기법들은 크게 분할(partition)방법과 계층적(hierarchical)인 방법의 두 가지로 나눌 수 있다. 분할 방법은 어떤 클러스터 내부의 객체들이 다른 클러스터에 있는 객체들보다 유사하도록 객체들을 분할한다. 계층적 클러스터링은 통합(agglomerative)방법과 분리(divisive)방법으로 나눌 수 있는 데, 전자는 처음에 각 객체를 하나의 작은 클러스터로 형성하는 것으로부터 시작해서 유사한 객체들을 병합하여 하나의 클러스터를 생성하는 단계를 거쳐 최종적으로 한 클러스터에 모든 객체들이 포함될 때까지 과정을 진행한다. 후자는 전자의 방법과 반대로 과정을 진행한다(Han and Kamber, 2001).

군집화는 주어진 객체들 중에서 주어진 객체 중에서 유사한 것들을 몇몇의 집단으로 그룹화하여 각 집단의 성격을 파악하는데 실제적으로 각 객체가 유사한지 그렇지 않은지를 측정할 수 있는 도구가 필요하다. 유사성은 동일한 관찰치에 대한 정의가 모호하므로, 군집분석에서는 일반적으로 비유사성을 기준으로 하고, 그 척도로서 거리(distance)를 주로 사용한다. 일반적으로 범주형 변수로 측정되는 두 개체 사이의 거리는 두 개체가 서로 다른 범주에 속한 회수를 이용하는데, 다음의 예를 들어보자. 성, 학력, 출신지역으로 관찰된 세 관찰치가 있다고 가정하자.

A= (남자, 고졸, 경기)
B= (여자, 고졸, 전남)
C= (남자, 대졸, 경기)

이 경우, 관찰치 A,B 사이의 거리는 2(불일치수), A,C 는 1, B,C 는 3이 된다.

이처럼 기존의 군집화에서 유사하다는 의미는 각 군집 안에 있는 개체들이 같은 속성 값이 많으면 많을수록 좋은 군집을 이루었다는 것을 의미했다. 기존의 군집화는 반드시 같은 항목(item)이 포함되어 있는 것끼리만 군집(cluster)을 이루고 좋은 군집화(clustering)로 인식되어졌다. 즉 같은 속성이면 1, 서로 다르면 0으로 표현한다(Guha and Rastogi, 1999).

1.2 연구목적

군집화는 주어진 관찰치 중에서 유사한 것들을 몇몇의 집단으로 그룹화 하여, 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해를 돕고자 하는 분석방법이다. 대용량 데이터에서 개개의 관찰치를 요약하는 것보다는 전체를 유사한 관찰치 들의 군집으로 구분하여 복잡한 전체보다는 그를 잘 대표하는 군집들을 관찰함으로써 전체 데이터에 대한 의미 있는 정보를 얻어낼 수 있을 것이다(Han and Kamber, 2001).

장바구니 데이타를 군집하는 주된 목적은 같은 물건을 사는 고객들끼리 군집화 하여 각 군집들에 속한 고객들의 니즈와 구매성향을 분석하여 각 군집에 알맞은 마케팅(marketing)을 하는 것이다. 그러나 요즈음 수많은 제품들이 쏟아져 나오고 고객들이 구매하는 물건들도 다양해지기 때문에 동일한 제품을 사는 고객들만 군집화 하기에는 어려움이 많고 문제점이 있다. 장바구니 데이타를 군집화 하는 목적은 그 군집에 속한 고객들을 분석하는 것인데 반드시 동일한 제품이 아니더라도 유사한 제품을 구입하는 고객들을 하나의 그룹으로 묶음으로써 보다 더 실질적인 마케

팅 전략을 가질 수 있게된다.

기존의 방법으로는 유사한 물건들을 산 고객들을 군집화 하는 것은 불가능하다. 왜냐하면 동일한 물건들을 산 고객들만 의미 있는 정보로 받아들이고 유사하더라도 동일하지 않기 때문에 같은 군집에 속할 가능성은 없어지게 된다. 하지만 제안된 알고리듬은 제품간에 유사한 관계가 있다는 개념을 도입하여 친밀한 관계를 가지는 제품끼리는 하나의 군집(cluster)이 될 수 있다는 것을 보인다. 그렇기에 기존의 방법과는 다른 결과가 나오며 그 군집화의 결과가 좋다 나쁘다 의 기준이 틀려지게 된다.

이 논문의 구성은 다음과 같다. 2장에서는 기존에 제안된 알고리듬에 대해 고찰을 하고, 3장에서는 본 논문이 제안하는 알고리듬에 대해 설명하고, 4장에서는 제안하는 알고리듬에 대한 수치예제를 보여 설명하고 그 효과를 보인 후, 마지막으로, 5장에서 결론을 내린다.

2. 기존연구고찰

군집화(Clustering)를 하는 데이터는 크게 수치형 값과 범주형 값으로 나뉜다. 빈발항목(Large Items)개념을 이용한 Transaction Clustering 알고리듬은 범주형 값 중에서도 장바구니 데이터를 Apriori 알고리듬에서 사용되는 빈발 항목 개념을 도입하여 빈발하는 item 끼리만 군집화를 한다.

범주형 속성들을 비교하는 방법에는 범주형 속성을 수치형 속성으로 바꾸어 좌표로 인식하여 두 속성간에 거리를 구하는 방법이 있다.

다음은 범주형 속성 중에서도 장바구니 데이터를 클러스터링 하는 알고리듬과 범주형 값을 가지는 두 속성의 유사도를 알아보는 두 가지 방법에 대해 2.1장과 2.2장에서 설명한다.

2.1 빈발항목을 이용한 장바구니 데이터 군집화

이 논문은 장바구니 데이터를 군집화 한 알고리듬으로, 같은 물건을 많이 구입한 사람끼리 같은 클러스터로 놓고 동일하지 않은 물건을 산 사람들은 다른 클러스터로 놓는다. 좋은 클러스터링 결과를 얻기 위해서는 한 클러스터에 많은 빈발항목(large item)이 있어야 하고 클러스터들 간에 겹치는 item이 적으면 적을수록 좋다(Wang and Xu, 1999).

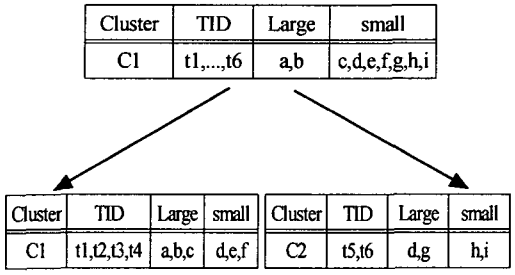
여기서 말하는 빈발항목(large item)이란 한 클러스터 안에 일정 이상 빈번하게 발생하는 item들이고, 비빈발항목(small item)이란 한 클러스터 안에 빈번하게 발생하지 않는 항목들을 말한다. <표 1>에는 군집화 하려는 예제 데이터베이스가 주어져 있고, 군집화 하는 전 과정이 <그림 1>에 나타나 있다.

<표 1> 예제 데이터베이스

TID	Item Set
t1	a,b,c
t2	a,b,c,d
t3	a,b,c,e
t4	a,b,f
t5	d,g,h
t6	d,g,i

<표 1>은 트랜잭션(TID)이 6개(t1,...,t6)이고 항목(item)이 9개(a,b,...,i)인 데이터베이스이며, 최소지지도를 60%로 정한다. 즉 항목집합이 빈

발이기 위해서는 6개의 트랜잭션 중에서 4개 이상의 트랜잭션이 그 항목집합을 포함하고 있어야 한다.



<그림 1> 알고리듬 실행예제

위의 예제를 보면 전체 트랜잭션 6개가 하나의 군집(Cluster)로 묶어 시작하여 빈발 항목 (a,b)이 포함된 트랜잭션 t1, t2, t3, t4가 하나의 군집으로, 그 외의 나머지 트랜잭션 t5, t6이 다른 하나의 군집으로 묶인다.

2.2 유클리디안 거리법을 사용한 데이터 군집화

범주형 속성을 가지는 객체를 클러스터링 할 때 해당하는 속성이 있으면 1, 없으면 0으로 표현한다. 그러면 각 객체들을 하나의 좌표 값으로 인식하여 근접한 거리에 있는 점들이 하나의 군집이 된다. (Guha and Rastogi, 1999)

```
item 종류 :     a, b, c, d, e, f
t1 = {a,b,c,e}  -> (1, 1, 1, 0, 1, 0)
t2 = {b,c,d,e}  -> (0, 1, 1, 1, 1, 0)
t3 = {a,d}      -> (1, 0, 0, 1, 0, 0)
t4 = {f}        -> (0, 0, 0, 0, 0, 1)
```

<그림 2> 속성 값을 좌표로 변환

<그림 2>는 4개의 객체에 들어있는 속성 값들을 좌표로 변환 한 것을 보여준다. 유클리디안 거리법(Euclidean distance)을 사용하여 위의 4개의 좌표 중 인접한 것끼리 군집이 이루어 질 수 있다.

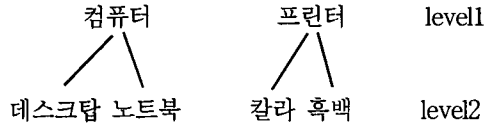
2.3 Jaccard coefficient를 사용한 군집화

Jaccard coefficient는 범주형 속성을 가지는 두 객체의 유사성을 파악하는 대표적인 척도로써 값이 클수록 유사도가 커져 한 클러스터로 이루어질 확률이 크게 된다.(Morzy et al. , 2001) 트랜잭션 T1과 T2의 유사도(similarity) sim(T1, T2)는 다음과 같이 정의한다.

$$sim(T1,T2) = \frac{| T_1 \cap T_2 |}{| T_1 \cup T_2 |}$$

예를 들어, T1 = {a, b, c}, T2 = {c, d, e}, T3 = {a, b, d} 이 세 개의 트랜잭션의 유사도를 알고싶다면 sim(T1, T2)=1/5=0.2, sim(T1, T3)=2/4=0.5, sim(T2, T3)=1/5=0.2 가되어 T1과 T3의 유사도 값이 가장 크기에 같은 군집을 이룰 가능성이 가장 높다.

2.4 Multilevel에서의 연관규칙



<그림 3> Multilevel 항목

항목(item)들이 낮은 수준(level)으로 내려갈수록 점점 많은 항목들로 인해 의미있는 연관관계

를 찾기 힘들다. 그래서 보통 강한 연관관계는 높은 수준의 항목들간에 발견된다. <그림 3>에서 처럼 어느 고객이 구입한 항목이 (노트북, 칼라프린트)라면 (컴퓨터, 프린터)로 한 수준을 올려야 빈발할 확률이 높아진다. 또한 경우에 따라서 (데스크탑, 흑백프린터)처럼 서로 다른 수준의 의미있는 빈발 항목을 찾을 수 있다. 어느 특정 수준을 선택 할 것인지는 그 수준의 빈발 항목을 고려하여 결정 할 수 있다.(J. Han and Y.Fu, 1995)

3. 제안하는 알고리듬

앞장에서 설명한 장바구니 데이터를 군집화(Clustering)하는 알고리듬과 트랜잭션간의 유사도(similarity)를 측정해 유사한 트랜잭션끼리 군집화하는 알고리듬의 단점은 반드시 같은 속성을 갖고 있어야 의미 있는 정보로 인식한다는 점이다.

본 논문에서 제안하는 알고리듬은 주어진 장바구니 데이터를 가지고 제품간에 친밀한 관계가 있다는 개념을 도입하여 친밀한 관계를 가지는 제품끼리는 하나의 군집(cluster)이 되는 것을 보인다. 따라서 3.1 절에서는 본 논문이 제안하는 알고리듬 수행절차를, 3.2 절, 3.3절, 3.4절, 3.5절에서는 수행절차를 각 단계별로 설명한다.

3.1 항목유사도를 고려한 장바구니 데이터 군집화

본 논문의 핵심 요소는 항목간에 얼마나 유사한지를 측정하여 트랜잭션(transaction)간의 유사도를 구할 때 사용한다. 해당 트랜잭션에 속해있는 항목들의 유사도가 커지면 커질수록 그만큼 트랜잭션간의 유사도는 커진다.

본 논문의 절차는 다음과 같다. 우선 전체 데이터베이스(DataBase)를 전처리과정(pre-processing)을 통해 item의 종류와 개수를 줄인다. 항목의 종류를 줄이기 위해서는 먼저 항목의 적절한 수준(level)을 정해야 하고, 항목의 개수를 줄이기 위해서는 Apriori 알고리듬에서 사용하는 빈발항목 개념을 사용하여 일정이상의 일빈발률을 만족하지 못하는 항목은 제거한다. 다음으로는 두 항목간의 유사도는 이 빈발률을 유사도 값으로 정한다. 트랜잭션들간의 군집화를 하기 위해서는 트랜잭션들이 서로 얼마나 유사한지를 알아야 하는데 전단계에서 구해진 항목간의 유사도를 기반으로 트랜잭션간의 유사도를 구할 수 있다. 유사한 트랜잭션들은 하나의 군집이 되어 다른 트랜잭션들과 다시 유사도를 구하며 군집들을 형성하게 된다.

아래의 4가지 단계는 본 논문이 제안하는 알고리듬의 절차를 나타낸 것이다.

알고리듬 단계

[단계1] Data Base 전처리과정(pre-processing)
- 항목의 종류를 level-up 하여 항목 종류를 줄임
- 최소지지도를 만족하지 못하는1빈발 항목 개수를 줄임

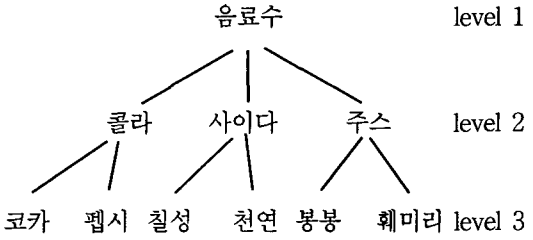
[단계2] 항목 들간 친밀도 측정
- 최소지지도를 만족하는 2빈발 확률 값을 구함

[단계3] 트랜잭션(transaction) 간의 유사도 측정
- 유사도 테이블에 측정결과 기록

[단계4] 유사도가 높은 트랜잭션간 군집형성

3.2 데이터베이스 전처리과정

일반적으로 장바구니 데이타에서 얻어지는 데이터의 양은 많고 희소하게 발생하는 항목은 중요하지 않으므로 필요에 따라서 전체 데이터베이스(data base)를 전처리과정(Pre-processing)단계를 거치므로 항목의 종류와 개수를 줄이므로 데이터의 양을 줄일 수 있다. 항목(item)의 종류를 줄이기 위해서는 항목의 수준을 level-up을 시켜 항목 종류를 일반화시키는 데 <그림 4>와 같이 적절한 수준(level)을 정해야 한다. 항목의 개수를 줄이기 위해서는 Apriori 알고리듬에서 사용하는 빈발개념을 이용하여 일정 이상의 1빈발률을 만족시키지 못하는 항목을 제거하여 의미없는 정보를 가지는 트랜잭션(transaction)과 항목은 제거를 한다. 즉, 거의 발생하지 않는 항목들은 제거 할 수 있다.



<그림 4> 제품의 종류를 줄이는 방법

3.3 항목 유사도 측정

모든 항목간의 유사도를 다 구하는 것은 많은 시간이 소요되고 모두가 다 의미 있는 정보가 아니다. 그러므로 같이 구입할 가능성이 높은지는 Apriori 알고리듬에서 사용되는 빈발항목 개념을 도입하여 기존의 트랜잭션 데이타에서 일정 이상으로 빈발하는 2빈발 항목의 확률 값을 친밀도 값으로 정한다. Apriori 알고리듬에서 사용되는 빈발 개념은 전체 거래 데이터에서 얼마나 해당 항목들의 거래가 자주 발생했는지를 보여준다. 특히 2빈발은 두 개의 항목들이 전체 거래 데이터에서 얼마나 2 항목이 같이 빈번하게 발생했는지를 보여준다. 최소 지지도 이상으로 2 항목이 동시에 거래가 이루어졌다는 의미는 다수의 고객이 이 2 항목을 같이 구입할 가능성이 높기 때문에 2 항목이 서로 연관성이 있어 유사하다는 것을 보여준다.

항목 a와 b의 유사도 sup(a, b)는 다음과 같이 정의된다.

$$sup\,(a,b) = \frac{a와b가\ 동시에\ 거래되는\ 경우수}{전체\ 거래수}$$

고객이 2개이상의 물건을 구입할 때 같이 구입할 가능성이 높은 물건들은 유사도가 높게 된다.

일정 이상의 2빈발을 만족하지 않는 항목은 항목 간의 친밀함이 전혀 없다고 정의를 내린다.

예를 들어 전체 data에서 최소지지도(minimum support)를 만족하는 2빈발 data <양복, 넥타이> = 40% 라면 양복과 넥타이의 친밀도 sup = 0.4라고 생각 할 수 있다.

3.4 트랜잭션 유사도 측정

위에서 구해진 항목간 유사도를 기반으로 트랜잭션(transaction) T1과 T2의 유사도(Similarity coefficient) SIM(T1,T2)는 Jaccard coefficient를 기반으로 다음과 같이 구해진다.

트랜잭션 T1=(a_1, a_2, ... , a_n)과 트랜잭션 T2=(b_1, b_2, ... , b_m)가 존재할 때

SIM(T1,T2)

$$= Jaccard\ \ coefficient + \frac{\sum_{a_i \in T_1,\ b_j \in T_2} sup(a_i,\ b_j)}{n \times m}$$

Jaccard coefficient는 범주형 속성을 가지는 두 객체의 유사성을 파악하는 대표적 척도로써 T1,T2간에 동일한 항목을 많이 가질수록 유사도 값이 커져 한 군집을 이룰 확률이 커진다. 하지만 T1과 T2간의 동일하지 않은 항목간의 유사도를 측정하지는 못하으로 전단계에서 구한 모든 항목간의 유사도를 고려하여 T1, T2간의 발생 가능한 모든 항목들의 유사도 값과 Jaccard coefficien값을 합하여 트랜잭션간 유사도 값을 구할 수 있다. 즉, T1,T2,간 동일한 항목은 Jaccard coefficien로 구하고 동일하지 않은 항목은 항목 유사도로 값을 구하여 두 값을 합한다. 예를 들어 전체 데이타베이스에서 구해진 친밀도 sup(a,d) = 45%, sup(b,e) = 35% , sup(a,c) = 30%를 기반으로 다음의 t1과 t2의 유사도를 알고 싶다면,

t1 : {a, b, c} t2 : {c, d, e}

① Jaccard coefficient = 1/5 = 0.2 ,

② $\frac{\sum_{a_i \in T_1,\ b_j \in T_2} sup(a_i,\ b_j)}{n \times m}$

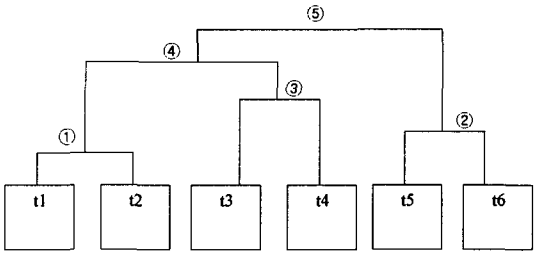
$= \frac{sup(a,d) + sup(b,e) + sup(a,c)}{9}$

= 0.12 로 t1과 t2의 유사도 SIM(t1,t2)는

0.2 + 0.12 = 0.32 가 된다.

3.5 유사가 높은 트랜잭션간 군집형성

<그림 5>와 같이 처음에는 각각의 트랜잭션(transaction)들은 하나의 군집(cluster)으로 설정한 후 이들 쌍간의 유사도를 기반으로 가장 큰 유사도를 가지는 트랜잭션들끼리 합병을 수행한다. 최종적으로는 한 군집에 모든 트랜잭션들이 포함될 때까지 과정을 수행한다.



<그림 5> 트랜잭션의 계층적 군집 형성의 예

4. 수치예제 및 효과

4.1 수치예제

이 절에서는 본 논문에서 제안하는 알고리듬을 가지고 장바구니 데이터를 군집화 하는 단계를 보여준다. <표 2>은 트랜잭션(TID)이 22개이고 품목(Item)이29개인 장바구니 데이터이다.

<표 2> 예제 장바구니 데이터

TID	Item	TID	Item
1	a1 b2 h1	12	f2 n1
2	a2 b2 n2	13	g1 m1
3	a1 b3 g2	14	k2 l2 q1
4	a3 b1 c2 g1	15	d1 k1 l1
5	a2 d2	16	c1 k1 l1
6	b1 g1 z1	17	k2 m2
7	f2 g1 l3	18	h2 l2 x1
8	f2 g1 k1	19	b1 g1
9	f3 g1 h2	20	a1 f2
10	f1 g2 b2	21	t1 y1
11	b1 f1 g2	22	o1

첫 번째 단계인 전처리 과정(Pre-processing)을 통해 품목의 종류를 level-up시켜 품목의 종류를 29개에서 14개로 줄인 결과가 <표 3>에 나타나 있다. 그리고 최소지지도 10%를 만족하지 못하는 항목과 트랜잭션을 제거한 결과가 <표 4>에 나타나 있다.

<표 3> 품목이 level-up된 장바구니 데이터

TID	Item	TID	Item
1	A B H	12	F N
2	A B N	13	G M
3	A B G	14	K L Q
4	A B C G	15	D K L
5	A D	16	C K L
6	B G Z	17	K M
7	F G L	18	H L X
8	F G K	19	B G
9	F G H	20	A F
10	F G B	21	T Y
11	B F G	22	0

<표 4> 최소지지도를 만족하는 장바구니 데이터

TID	Item	TID	Item
1	A B H	11	B F G
2	A B N	12	F N
3	A B G	13	G M
4	A B C G	14	K L
5	A D	15	D K L
6	B G	16	C K L
7	F G L	17	K M
8	F G K	18	H L
9	F G H	19	B G
10	F G B	20	A F

다음 단계로 <표 4>을 통하여 항목간 유사도를 구하기 위해서는 최소지지도를 만족하는 2빈발 항목은 다음과 같다.

sup(a, b) = 4/20 = 0.2
sup(f, g) = 5/20 = 0.25
sup(k, l) = 3/20 = 0.15
sup(b, g) = 6/20 = 0.30

위의 4개의 항목간 유사도 값이 나온다. 나온 항목간 유사도를 기반으로 20개의 트랜잭션들간의 유사도를 구하여 테이블로 나타낸 결과가 <그림 6>에 나타나 있다.

유사도를 기록한 위의 <그림 6> 테이블을 통해서 t6과 t19간의 유사도가 가장 크다는 것을 알 수가 있다. 그러므로 t6과 t19는 같은 군집(cluster)이 되어 다른 트랜잭션들과 다시 유사도를 구해 비교한 결과가 <그림 7>에 나타나 있다.

이와 같은 과정을 계속 반복하여 최종적으로는 한 군집에 모든 트랜잭션들이 포함될 때까지 과정을 수행한다. 하지만 실제 마케팅에서 사용하려면 현재의 군집의 개수가 사전에 지정된 군집 개수 보다 크고 최소 유사도 이하의 군집이 없을 때까지 앞의 과정을 반복 수행한다. <그림 8> 단계까지 보면 사전에 정해진 최소 유사도 값(0.5)이상 발견되는 유사도가 없으므로 이 단계에서 알고리듬은 마쳐진다. 결론적으로,

C1 = (t3, t4, t6, t10, t11, t19)
C2 = (t14, t15, t16)
C3 = (t7, t8, t9)
C4 = (t1, t2)

이렇게 4개의 군집(Cluster)을 얻을 수 있고 그 외의 트랜잭션들은 다른 트랜잭션들과 합병이 한번도 이루어지지 않았기에 의미가 없는 군집으로 간주하여 제거한다. 위의 4개의 군집들 중 C1에서 t3과 t11이 하나의 군집을 이룬 것을 볼 수

있는데 <그림 9>에서 처럼 t3과 t11이 하나의 군집을 이룰 수 있던 주요한 이유는 먼저 항목의 수준을 높여 품목을 단일화 시켰을 때 T3, T11로 되어 공통되는 항목이 B,G가 있고 A와B, F와 G는 서로 유사한 항목이기에 하나의 군집을 이루게 되었다.

위의 과정을 거쳐 실제로 다음과 같은 결과가 나왔다고 가정하면,

T1 = {양복, 넥타이}
T2 = {양말, 와이셔츠}

이렇게 두 개의 트랜잭션이 있을 때 기존 군집화에서는 공통되는 항목이 하나도 없기에 하나의 군집화로 나올 가능성은 거의 없고 같은 군집에 속한다는 결과가 나온다 하더라도 그 결과가 좋지 않다고 판명을 짓는다. 하지만 이 두 트랜잭션을 자세히 살펴보면 두 개다 남성을 위한 제품들이라는 것을 알 수 있다. 즉 비슷한 제품군이지만 기존의 방법으로는 비슷한 제품이다라는 것을 밝힐 수도 없었고 그럴 필요성도 느끼지 않았다. 그러나 제안된 알고리듬은 항목간에 유사도를 고려하여 유사한 관계를 가지는 항목끼리는 하나의 군집을 이룰 수 있게 된다.

4.2 효과

위의 알고리듬을 실행시켜 나온 군집들은 유사한 속성들을 가지는 군집별로 나온 것이다. 그 군집들을 대상으로 타겟마케팅(target marketing)을 하고자 할 때 마케팅 전략에서 사용되는 STP(Segment, targeting, Positioning)과정의 한 부분으로 도입할 수 있다(문준연, 1990).

첫 번째 단계로 시장세분화(Segment)는 시장의 전체 소비자들을 구매관련 니즈, 행동 또는 특성 면에서 유사한 하부 집단으로 구분하는 것이다. 시장세분화는 각기 유사한 소비자들로 구성된 하부 집단별 차이를 파악하고, 자사가 우위를 확보할 수 있도록 이러한 차이를 이용하는 것이다. 본 논문은 군집화(clustering)를 통해 바로 소비자의 니즈나 구매반응이 유사한 사람들의 집단을 파악하여 소비자에 따라 시장을 세분화했다. 특히 분석하고자 하는 분야가 한정된 대상일 때 군집화를 통한 시장세분화는 매우 중요하다.

두 번째 단계로 표적시장 선정(Targeting)으로 각 세분시장의 매력도를 평가하여 자사에 적합한 세분시장을 선정하는 것이다. 이를 위해서는 시장의 매력성과 기업의 역량 등을 감안하여 시장규모, 성장성, 자사의 경쟁적 우위성, 예상 수익들을 면밀하게 살펴보아야 한다. 목표 l장을 선정한 후 이들 집단에게 어떤 방식으로 접근할 것인지를 결정해야 한다. 세부 고객을 공략하는 전략은 크게 비 차별적 마케팅 전략, 차별적 마케팅 전략, 집중적 마케팅 전략의 세 가지로 나뉜다.

인터넷 마케팅에서 비 차별적 마케팅은 일반적으로 검색엔진 등에서 볼 수 있다. 야후, 심마니, 네이버 등과 같은 검색 사이트는 모든 소비자들이 자신들이 원하는 정보를 보다 빠르고 정확하게 검색하는 것을 목표로 삼고 있다. 하지만 이러한 검색 사이트들도 최근에는 특정 영역으로 세분화하고 이에 맞는 서비스를 제공하려 하고 있다.

차별적 마케팅이란 각 세분 시장마다 차별화된 마케팅 활동을 전개하는 것을 말한다. 즉, 서로 다른 세분 시장에 맞는 상이한 마케팅 프로그램을 수행한다. 차별적 마케팅이 인터넷 비즈니스에서 점점 활발해지는 이유는 고객의 데이터 및 행동 데이터 등을 적은 비용으로 쉽게 얻을

수 있기 때문이다. 고객의 정보 수집이 보다 용이해지면서 서로 다른 요구를 가지고 있는 고객들에게 그에 맞는 서비스를 제공하면서 수익을 증대할 수 있다. 그러나 이 전략은 자칫 잘못하면 비용의 증가라는 문제를 초래할 수 있다.

집중적 마케팅 전략은 하나의 세부시장에 기업의 역량을 집중하는 방법이다. 전체 시장이나 많은 세부시장을 공략하기에는 기업의 자원이나 역량이 부족할 때 주로 쓰이는 방법이다. 예를 들어 lefthand.com은 왼손잡이의 사람들에게 가위나 주방용품 등을 판매하고 있다. 우리 나라의 왼손잡이가 현재 200만 명이라는 점을 감안할 때, 이들에게 맞는 제품을 판매하는데 기업의 역량을 집중하는 것은 적절한 전략이라고 할 수 있다.

본 논문에서 제시한 알고리듬을 통해 군집화된 그룹들을 수익관점이나 선호하는 제품들이 어떤 것이 있는지 분석할 수 있을 것이다. 분석된 그 그룹들의 특징에 맞게 차별적 마케팅을 하여 고객들에게 맞는 적절한 서비스를 통해서 수익을 증대 할 수 있을 것이다(김재일, 2002).

세 번째로 포지셔닝(Positioning)으로 선정된 세분시장에서 자사 제품이 경쟁사들과 비교하여 뚜렷하게 구분되는 독특한 이미지를 갖도록 하는 방법을 모색하는 것이다. 이러한 독특한 이미지를 구축하는 과정이 포지셔닝이다. 제품의 이미지를 구축해야 하는데 그 제품이 소비자들의 마음 속 에서 경쟁품에 비교하여 우수한 위치를 차

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		0.54	0.58	0.46	0.28	0.33	0.03	0.03	0.23	0.06	0.06	0	0.05	0	0	0	0	0.25	0.33	0.28
2			0.58	0.46	0.28	0.33	0.03	0.03	0.03	0.26	0.26	0.25	0.05	0	0	0	0	0	0.33	0.28
3				0.81	0.28	0.74	0.23	0.23	0.23	0.56	0.56	0	0.28	0	0	0	0	0	0.74	0.28
4					0.23	0.56	0.19	0.19	0.19	0.44	0.44	0	0.24	0	0	0.16	0	0	0.58	0.23
5						0.05	0	0	0	0.03	0.28	0	0	0	0.25	0	0	0	0.05	0.33
6							0.3	0.3	0.3	0.76	0.76	0	0.41	0	0	0	0	0	1	0.05
7								0.57	0.56	0.59	0.26	0.29	0.29	0.28	0.22	0.22	0.03	0.25	0.34	0.29
8									0.56	0.59	0.26	0.29	0.29	0.28	0.27	0.27	0.25	0.03	0.34	0.29
9										0.59	0.26	0.29	0.29	0	0	0	0	0.25	0.34	0.29
10											0.59	0.29	0.34	0	0	0	0	0	0.8	0.29
11												0.04	0.3	0	0.2	0	0	0	0.76	0.04
12													0.06	0	0	0	0	0	0.06	0.33
13														0	0	0	0.33	0	0.41	0.06
14															0.71	0.71	0.37	0.37	0	0
15																0.53	0.28	0.28	0	0
16																	0.28	0.28	0	0
17																		0	0	0
18																			0	0
19																				0.11
20																				

<그림 6> 초기유사도 테이블

	6,19	1	2	3	4	5	7	8	9	10	11	12	13	14	15	16	17	18	20
6,19		0.33	0.33	0.74	0.56	0.05	0.3	0.3	0.3	0.76	0.76	0	0.41	0	0	0	0	0	0.05
1			0.54	0.58	0.46	0.28	0.03	0.03	0.23	0.06	0.06	0	0.05	0	0	0	0	0.25	0.28
2				0.58	0.46	0.28	0.03	0.03	0.03	0.26	0.26	0.25	0.05	0	0	0	0	0	0.28
3					**0.81**	0.28	0.23	0.23	0.23	0.56	0.56	0	0.28	0	0	0	0	0	0.28
4						0.23	0.56	0.19	0.19	0.19	0.44	0.44	0	0.24	0	0	0.16	0	0.23
5							0.05	0	0	0	0.03	0.28	0	0	0	0.25	0	0	0.33
7								0.57	0.56	0.59	0.26	0.29	0.29	0.28	0.22	0.22	0.03	0.25	0.29
8									0.56	0.59	0.26	0.29	0.29	0.28	0.27	0.27	0.25	0.03	0.29
9										0.59	0.26	0.29	0.29	0	0	0	0	0.25	0.29
10											0.59	0.29	0.34	0	0	0	0	0	0.29
11												0.04	0.3	0	0.2	0	0	0	0.04
12													0.06	0	0	0	0	0	0.33
13														0	0	0	0.33	0	0.06
14															0.71	0.71	0.37	0.37	0
15																0.53	0.28	0.28	0
16																	0.28	0.28	0
17																		0	0
18																			0
20																			

<그림 7> 첫 번째 반복후 유사도 테이블

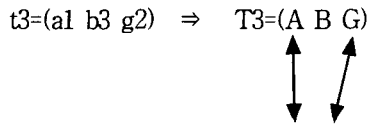
지해야 한다. 군집화된 그룹들 안의 고객들은 이미 본인이 선호하는 제품군들이 각인 되어 있다. 그 고객들에게 굳이 다른 그룹의 제품군을 홍보하기보다는 동일 그룹의 친밀도가 높은 제품을 홍보하는 것이 보다 효과적일 것이다.

예를 들어 <그림 9>에서 나온 결과에서 T3의 고객에게 F를 추천하거나 반대로 T11고객에게 A를 추천할 수 있다. 또한 A제품 군을 산 T3 고객은 실제 a1을 샀기에 A제품 군에 포함된 a2,a3 같은 제품을 추천하여 고객의 구매 심리를 일으킬 수 있게된다.

장바구니 데이터를 군집화하는 목적은 같은 물건을 사는 고객들끼리 군집화하여 각 군집에 알맞은 타겟 마케팅을 하는 것이다.

	6,19 10,11 3,4	14, 15,16	7,8 9	1,2	5	12	13	17	18	20
6,19 10,11 3,4		0.16	0.28	0.37	0.24	0.05	0.3	0	0	0.19
14,15 ,16			0.13	0	0.07	0	0	0.31	0.31	0
7,8 9				0.08	0	0.29	0.29	0.07	0.2	0.29
1,2					0.28	0.13	0.05	0	0.13	0.28
5						0.28	0	0	0	0.33
12							0.06	0	0	0.33
13								0.33	0	0.06
17									0	0
18										0
20										

<그림 8> 열번째 반복후 유사도 테이블



<그림 9> 군집을 이룬 두 트랜잭션

5. 결론 및 향후과제

트랜잭션들을 군집화 하는 주된 목적은 같은 제품을 사는 고객들끼리 군집화 하여 각 군집에 알맞은 타겟 마케팅을 하는 것이다.

이 논문에서는 항목간의 유사성을 고려한 트랜잭션들의 군집화 문제를 연구하였다. 본 문제를 풀기 위하여 항목의 유사도 척도를 제안하였고, 이 척도를 이용하여 트랜잭션 데이터를 가지는 군집화 알고리듬을 제안하였다. 제시된 알고리듬을 가지고 장바구니 데이타를 분석하게 될 때, 기존에 항목의 유사성을 고려하지 않은 방법보다는 보다 현실적으로 군집이 이루어지게 된다. 선행 연구들이 주로 상품군의 유사도를 기반으로 한 장바구니 분석을 이용하여 상품군의 구매연관관계를 분석한데 반해 본 연구에서는 트랜잭션의 유사도를 측정하고 이에 따라 트랜잭션을 군집화하는 차별화된 방법과 적용분야를 제시하였다. 또한 기존 연구에서 시장 세분화를 위해 고객의 인구 통계학적 자료를 기반으로 고객군을 군집화 하는데 반하여 본 연구에서는 트랜잭션의 유사도를 측정하여 시장 세분화를 하려는 점에서 기존 연구와 차별되는 점이 있다. 결과로 나오는 각각의 군집들로 물건을 사는 고객들의 구매 선호도를 알 수 있다. 그런데 만약 두 번 이상 구매한 고객이 있고, 이 고객이 두 번의 구매에서 다른 상품을 구입한 경우에는 분석의 결과에 따라 추천하는 상품이 달라질 수 있다. 두 번 이상 구매한 고객의 구입 품목이 중복되는 경우에는 구입한 품목의 회수를 고려하여 가중치를 두어 알고리듬을 구현할 수 있으나 본 논문에서는 제시되지 않고 향후 과제로 남겨두었다. 제안된 알고리듬은 기존에 제시된 속성간에 유사도를 고려하지 않고 장바구니 데이터를 군집화 한 것보다 군집에 대한 현실적인 분석을 할 수 있으므로 효과적인 타겟 마케팅을 할 수 있다.

참고문헌

김재일, 인터넷 마케팅, 박영사, 2002

문준연, 마케팅, 청목출판사, 1990.

Han. J. and M. Kamber; Data Mining: Concepts and Techniques, Morgan kaufmann Publishers, 2001

Han J. and Y.Fu, "Discovery of multiple-level association rules from large databases." *VLDB*, 1995, 420~431

MacQueen J., "Some methods for classification and analysis of multivariate observations.", *Math Statist, Prob.*, 1967, 1:281~297.

Wang, K., C. Xu, and B. Liu. "Clustering Transactions Using Large Items." *ACM CIKM*, Nov. 1999, 483~490.

Agrawal, R. and R. Srikant. "Fast algorithms for mining association rules." *In Proc. 1994 Int. Conf. VLDB*, Santiago, Chile, September 1994, 487~499.

Guha S., R. Rastog and K. Shim, "ROCK : a robust clustering algorithm for categorical attributes," in *Proceedings of the 15th International Conference on Data Engineering*, 1999, 512~521

Guha S., R. Rastogi and K. Shim, "CURE : An Efficient Clustering Algorithm for Large Databases", *SIGMOD98*, 1998, 73~84

Morzy T., M. Wojciechowski and M. Zakrzewicz, "Scalable Hierarchical Clustering Method for Sequences of Categorical Values", *Proc. of the 5th PAKDD*, Kowloon, Hong Kong, 2001

Zhang T., R. Ramakrishnan and M. Livny; "BIRCH : An Efficient Data Clustering Method for Very Large Databases", *ACM SIGMOD96*, 1996, 103~114

Raymond T. and J. Han. "Efficient and Effective Clustering Method for Spatial Data Mining", *VLDB* 1994.

Huang Z., "Extensions to the k-means algorithm for clustering large data sets with cataegorical values." *Data Mining and Knowledge Discovery*, 1998, 2:283~304.

Abstract

Transactions Clustering based on Item Similarity

Sang Wook Lee* · Jae Yearn Kim*

Clustering is a data mining method which help discovering interesting data groups in large databases. In traditional data clustering, similarity between objects in the cluster is measured by pairwise similarity of objects. But we devise an advanced measurement called item similarity in this paper, in terms of nature of clustering transaction data and use this measurement to perform clustering. This new algorithm show the similarity by accepting the concept of relationship between different attributes. With this item similarity measurement, we develop an efficient clustering algorithm for target marketing in each group.

Key words : 군집화(clustering), 아이템 유사도(item similarity), 타겟마케팅(target marketing)

* Department of Industrial Engineering, Hanyang University