데이터마이닝 기법을 이용한 구매정보 분석방법

이문표*, 조경수, 김응모
*성균관대학교 정보통신공학부
e-mail:thepleig@naver.com

The Analysis of Purchase Information Using Data Mining Methods

Mun Phyo Lee*, Kyung Soo Cho, Ung-mo Kim
*School of Information and Communication Engineering,
Sungkyunkwan University

요 약

컴퓨터와 통신의 급속한 성장은 방대한 양의 정보를 서로 공유하는 정보화 시대를 출현 시켰고 이러한 다양하고 많은 정보들로부터 유용한 정보를 얻어 내는 필요성이 대두되었다. 그리고 기업들은 효율적인 경영을 위해 전통적인 자원들을 효과적으로 운영하고 관리해야할 뿐만 아니라 고객만족을 위해 제품, 서비스의 질을 향상시켜야 하는 문제에 봉착했다. 효과적인 고객관리 전략을 수행하는 능력이 중요한 시점에서 데이터베이스 마케팅의 진보된 형태로서 데이터 마이닝 기술의 활용이 요구되고 있다. 본 논문에서는 마케팅에 이용되는 구매정보를 효율적으로 조회하는데 적용할 수 있는 방법을 제안한다. 이 방법을 통해서 판매자는 마케팅에 적용하여 서비스 향상을 꾀할 수 있을 것이다.

1. 서론

기업들의 기존 마케팅은 상품위주의 불특정 다수를 향한 마케팅이 주를 이루었다. 당시에는 필수품이 부족했기에 기능이 최소화되고, 규격화된 상품을 대량으로 생산하여 광고를 거치면 상품이 팔렸다. 하지만 시간이 지나면서 필수품에 대한 수요와 공급이 균형을 이루고 품질의 차이도 거의 없어지게 되자 고객들은 개인적인 취향과 기호에 맞는 상품을 찾기 시작 했고, 기업들도 고객의 요구가 다양하게 변한다는 사실을 깨닫고 새로운 시장의 틀에 맞는 상품을 생산하기 시작했다. 이러한 시대적 배경이 기업들의 특정세부시장을 목표로 한 마케팅을 발전시켰고, 집중적인 마케팅을 전개함으로써 효과를 극대화 시킬 필요성에 의해CRM이란 개념이 등장하게 되었다.

CRM(Customer Relationship Management)은 선별한 고객에게서 수익을 창출하고, 고객 관리를 가능하게 하는 솔루션으로써 고객과 관련된 기업의 자료를 분석·통합하여 고객의 특성에 기초한 마케팅 활동을 계획하고, 지원 및 평가하는 과정이 라고 할 수 있다. CRM은 크게 세가지로 분류될 수 있다. 먼저 분석CRM은 데이터웨어하우스, 데이터마이닝, OLAP 를 이용하여 마케팅 의사결정을 지원하는 마케팅 의사결정 시스템을 의미하고, 운영CRM는 ERP와 연관되어 있으며, 주로 영업과 서비스를 위한 시스템이다. 마지막으로 e-CRM은 인터넷을 기반으로 한 포탈서비스의 급성장과 기업의 온라인화가 가속화 되면서, 인터넷에 대응하는 CRM이다. 분석CRM에 활용되는 데이터마이닝 기법으로는 군집화, 분류, 연관규칙, 순차패턴, 의사결정트리, 인공신경망 등이 있다. [1]

본 논문은 다음과 같이 구성되어 있다. 2장에서는 관련 연구에 대해 기술하고, 3장에서는 일반화된 순차패턴마이닝을 구매정보 분석에 적용하는 방법을 제시한 후, 마지막으로 4장에서 결론을 맺고자 한다.

2. 관련연구

2.1 연관규칙

연관규칙 기법은 상품 혹은 서비스간의 관계를 살펴보고 이로부터 유용한 규칙을 찾아내고자 할 때 이용될 수 있다. 연관성측정은 구축된 데이터베이스를 기반으로 어떤 특정 문제에 대해 아직 일어나지 않은 사건을 찾아내는 작업이다. 동시에 구매될 가능성이 큰 상품들을 찾아냄으로써 시장바구니분석에서 다루는 문제들에 적용 가능하다. "아이템셋 A → 아이템셋 B"로 표시된 경우, "상품 A가 구매되어진 경우는 상품 B도 구매된다." 라고 해석된다. 연관 규칙기법을 이용할 수 있는 데이터는 판매시점에서 기록되어진 거래와 품목에 관한 정보를 담고 있어야 한다. 데이터의 형태는 비감독데이터이며 특별히 각 고객들이 누구인지에 대한 구분이나 고객들에 관한 성별, 나이 등의 인구 통계학적인 자료를 비롯한 기타 정보들을 필요로 하지는 않는다. 연관규칙은 연관성분석의 결과 이해가 쉽고, 거래내용에 대한 데이터를 변환 없이 그 자체로 이용할 수 있으므로 사용이 편리하지만 분석을 위해 필요한 계산이 많고, 분석하는 품목수가 증가할수록 분석에 필요한 계산이 기하급수적으로 늘어나는 단점이 있다. [2]

2.2 순차패턴마이닝

순차패턴마이닝은 순서대로 일어난 데이터를 분석해 빈도수가 높은 순차 패턴을 찾아내는 기술로써 동시에 구매될 가능성이 큰 상품군을 찾아내는 연관성측정에 시간이라는 개념이 포함되어 순차적인 구매 가능성이 큰 상품군을 찾아내는 것이다. 순차적 패턴 발견에서 연관 규칙 X→Y 은 “상품 X가 구매되면 일정시간의 경과한 다음에는 상품 Y가 구매된다.”라고 해석한다. 예를 들어, “컴퓨터를 구입한 사람들 중 20%는 다음달에 복합기나 프린터를 구입할 것이다”와 같은 규칙을 찾아낼 수 있다. 연관규칙과 순차패턴마이닝의 차이 점은 연관성 규칙은 X→Y, Y→X가 성립하지만, 순차패턴마이닝에서는 X→Y 관계만 성립된다. 연관규칙은 X, Y 중 어느 것이 먼저 일어나도 관계없지만 순차패턴마이닝의 경우는 시간상 X가 선행되어야 한다. 이처럼 순차적 패턴발견은 구매의 순서가 고려되어 상품간의 연관성이 측정되고 이의 정도에 따라 유용한 연관 규칙을 찾는 기법이다. 그러므로 연관성 측정에서의 데이터 형태에서 각각의 고객으로부터 발생한 구매의 시점에 대한 정보가 포함되어져야 한다. [3]

2.3 일반화된 순차패턴마이닝 (GSP 알고리즘)

GSP 알고리즘은 연관 규칙의 Apriori알고리즘에 기반을 두고 있다. GSP 알고리즘은 순차패턴마이닝에 시간제한, 슬라이딩 윈도우, 분류개념을 도입하여 순차패턴 문제를 일반화화한 것이다. GSP 알고리즘의 핵심은 후보 시퀀스 생성이 지지도 계산에 대한 데이터베이스의 완전한 패스의 결과라는 것이다. GSP는 몇 번의 DB스캔을 거쳐 DB를 변환한다. 첫 번째 스캐닝에서, 1-시퀀스들이 구해진다. 빈도가 높은 항목으로부터, 후보 2-시퀀스들의 집합은 형성된다. 빈도가 높은 2-시퀀스들은 후보 3-시퀀스들을 생성시키는 데 이용되고, 이 과정이 빈도가 더 높은 시퀀스가 발견되지 않을 때까지 반복된다. 알고리즘은 크게 후보 생성과정과 지지도 계산 과정으로 나뉘어 진다.

후보 생성과정은 Join Phase와 Prune Phase로 구분된다. Join Phase에서는 빈도가 높은 시퀀스의 집합을 만들고, 다음의 후보는 자신과 다른 시퀀스의 조합으로 후보를 생성한다. Prune Phase에서는 최소 지지도를 기반으로 Join Phase의 결과를 필터링하며 빠른 계산을 위해 후보 시퀀스는 해시 트리 내에 저장되어 처리된다.

지지도 계산에서 모든 후보 시퀀스의 부분 시퀀스가 생성되며. 각 부분 시퀀스는 해시 트리에서 만들어진다. 해시 트리에 있는 후보가 부분 시퀀스와 일치할 때, 지지도가 증가한다. [4]

2.4 Apriori 알고리즘

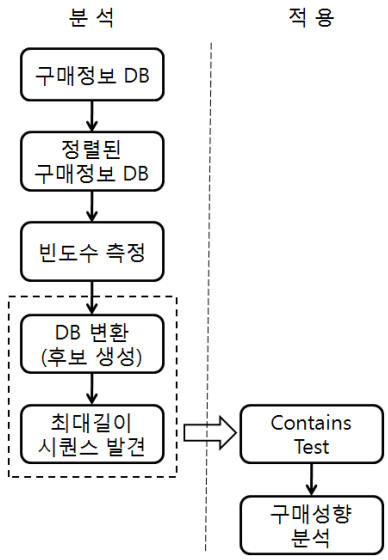
Apriori 알고리즘은 최소지지도를 만족하는 항복집합의 빈발목록을 찾는 과정을 구현한 알고리즘으로 두 단계로 구성된다. 첫 번째 단계에서는 최소 지지도 설정 값에 따라 빈도수가 높은 항목의 집합들을 찾아내고 그 다음 단계에서는 이들 집합들로부터 신뢰도 설정 값을 모두 뽑아낸다. Apriori 알고리즘의 핵심 원리는 빈발 항목집합의 모든 부분집합의 지지도는 빈발 항복집합의 지지도보다 크다는 것이다. 항목들을 먼저 구하고 그 항목들로부터 길이를 늘려가면서 빈도수가 높은 항목들의 집합을 구하는 방식으로 처리된다. [2]

2.5 AprioriAll 알고리즘

AprioriAll 알고리즘은 Apriori 알고리즘을 확장한 것으로 다음과 같이 진행된다. 먼저, 단계별로 후보 시퀀스를 생성해서 지지도를 산출해낸다. 그리고 최소 지지도를 만족하지 못하는 후보들을 제거한 후, DB를 변환해서 각 트랜잭션을 빈번한 후보들로 대체하여 순차패턴 빈발목록을 찾는다. [3]

3. 구매정보 분석방법 모델

본 논문에서 제안하는 추천 서비스 모델은 <그림 1>과 같다.

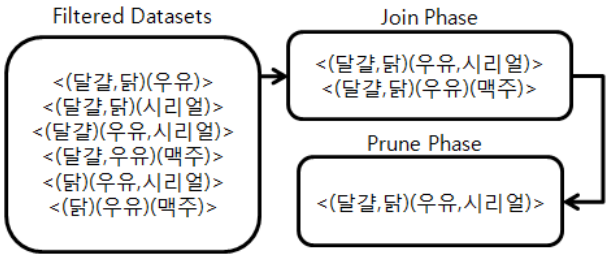


<그림 1> 구매정보 분석방법 모델

<그림 1>의 구매정보DB는 고객의 ID와 거래시간을 주키,보조키로 정렬되어 거래데이터베이스를 시퀀스 형태로 변환된다. 아이템셋의 빈도수 측정과정에서는 데이터베이스를 스캐닝해서 빈도수가 높은 아이템셋을 발견한다(빈도수가 낮은 아이템셋은 배제 시킨다). 지지도는 아이템셋을 포함하는 고객수를 말하며, 빈도수가 높은 아이템셋은 알고리즘의 효율을 위해 각각 정수로 매핑시킨다. 그리고 GSP 알고리즘에 따라 후보 시퀀스를 생성하고 높은 빈도수를 가진 시퀀스 중에서 최대길이 시퀀스를 찾는다. 이렇게 변환된 정보들은 시간제한과 슬라이딩 윈도우 개념을 이용해 아이템셋의 포함여부를 테스트하고, 그 결과물로부터 구매자의 성향을 분석한다.

3.1 일반화된 순차패턴 마이닝 적용

구매된 아이템들을 대상으로 일반화된 순차패턴 마이닝을 적용해서 후보 시퀀스를 찾는 과정으로 결과로 마이닝 전 설정된 최소 지지도 이상의 아이템셋들이 생성된다.



<그림 2> Join Phase

<그림 2>는 정렬된 구매정보가 Join단계와 Prune단계를 거쳐 4개의 시퀀스로 이뤄진 후보 시퀀스가 생성되는 과정을 보여준다. 필터링된 구매정보는 <달걀, 닭, 우유, 시리얼, 맥주>의 아이템들로 이루어진 총 6개의 시퀀스로 분류되어 있다.

Join Phase는 시퀀스S_1의 첫 번째 아이템셋의 첫 번째 아이템 하나를 버리고, 다른 시퀀스S_2의 마지막 아이템셋의 마지막 아이템 하나를 버린 결과가 같을 때, 두 시퀀스를 조인한다. <그림 2>의 Filtered Databasets에서 첫 번째 시퀀스인 <(달걀, 닭)(우유)>에서 첫 번째 아이템셋의 첫 번째 아이템인 달걀을 지우면 <(닭)(우유)>가 된다. 그리고 다섯 번째 시퀀스인 <(닭)(우유,시리얼)>에서 마지막 아이템셋의 마지막 아이템인 시리얼을 지우면 역시 <(닭)(우유)>가 된다. 그리고 이 두 시퀀스를 조인한 결과로 <(달걀, 닭)(우유, 시리얼)> 시퀀스를 얻을 수 있다. <(달걀, 닭)(우유)(맥주)> 시퀀스도 마찬가지 방식으로 얻을 수 있다.

Prune Phase에서는 패턴의 인접하는 부분열이 최소 지지값을 만족하지 못한 패턴을 제거한다. 시퀀스의 아이템셋에서 첫 번째 아이템셋의 한 아이템을 제거하거나 마지막 아이템셋의 한 아이템을 제거하는 경우가 인접하는 부분열로 조인하기 전의 시퀀스라고 할 수 있다. 그리고 어떤 시퀀스의 최소 2개의 아이템을 가진 아이템셋에서 한 아이템을 제거하는 경우역시 인접하는 부분열이다.

<표 1> Prune Phase

아이템 제거	결과 시퀀스
달걀	<(닭)(우유, 시리얼)>
닭	<(달걀)(우유, 시리얼)>
우유	<(달걀, 닭)(시리얼)>
시리얼	<달걀, 닭)(우유)>

<표 1>은 Join Phase의 첫 번째 시퀀스인 <(달걀, 닭)(우유, 시리얼)>의 인접하는 부분열을 보여준다. <표 1>의 결과 시퀀스는 모두 Filterd Datasets에 등장하므로 지지도를 만족하고 Purune Phase에서 남게 된다. Join Phase의 두 번째 시퀀스인 <(달걀,닭)(우유)(맥주)>의 경우는 닭 아이템을 제거할 경우 <(달걀)(우유)(맥주)>라는 인접하는 부분열을 얻을 수 있다. 그런데 이 시퀀스는 Filtered Datasets에 등장하지 않으므로 지지도를 만족하지 못하므로 Pruning된다. 그 결과, 최종적으로 얻게 되는 후보 순차패턴은 <(달걀,닭)(우유,시리얼)>이 된다.

3.2 Contains Test

Contain test는 데이터 시퀀스가 후보 시퀀스를 포함하는지 테스트하는 과정으로 분석과정에서 생성된 후보 시퀀스를 탐색하는데 적용된다.

<표 2> Contains test

트랜잭션	아이템셋
100	달걀,닭
200	시리얼,맥주
400	우유
500	**달걀,닭**
700	**우유**
900	닭,**시리얼**
950	맥주

Contains test는 모든 요소들이 찾아질 때까지 전방 단계와 후방 단계를 반복한다. 두 아이템의 시간차가 최대갭보다 적은 경우에 연속하는 두 아이템을 찾는 것이 전방단계이다. 시간차가 최대갭보다 커지는 경우에는 후방단계로 전환된다. 후방단계에서는 전환 전 전방단계에서 찾은 마지막 아이템의 트랜잭션 시간에서 최대갭을 뺀 후의 트랜잭션부터 다시 전방단계로 전환되어 검색을 시작한다. (표1)은 최소갭을 50, 그리고 최대갭을 400으로 설정 후,<(달걀,닭)(우유)(시리얼)> 패턴을 찾는 것을 예로 들고 있다. 먼저 (달걀, 닭) 아이템셋은 트랜잭션 100에서 최초로 등장한다. 그다음 (우유) 아이템셋은 트랜잭션 400에서 찾아진다. 마지막으로 (시리얼) 아이템셋을 찾는데 트랜잭션 900에서 찾을 수 있다. 여기까진 최소 갭인 50을 만족하지만, 최대 갭은 (달걀,닭)의 트랜잭션과 (시리얼) 아이템셋의 차이가 800이므로 최대 갭인 400을 만족시키지 못한다. 이 패턴은 조건을 만족 시키지 못했으므로 다음 (달걀, 닭) 아이템셋이 등장하는 트랜잭션 500부터 다시 Contains Test를 한다. (우유)는 트랜잭션 700에 등장하고, (시리얼)은 트랜잭션 900에 등장한다. 여기서 최소 갭은 200이고, 최대 갭은 400이므로 설정된 수치를 만족하는 순차패턴을 찾게된다.

4. 결론 및 향후 연구 과제

이 논문에서 제안한 구매정보 분석 방법은 고객의 구매패턴 후보를 도출해서 시계열 기반으로 검색 및 분석할 수 있도록 하였다. 이 방법을 통해 구매자의 구매 성향을

다각도로 파악해서 마케팅에 활용할 수 있을 것이다. 하지만 GSP알고리즘이 갖는 몇 가지 문제점은 개선이 필요하다. 먼저 GSP알고리즘의 후보 시퀀스는 데이터베이스 내에서 생성되고 수차례의 데이터베이스 스캐닝이 필요하다. 또, 긴 순차 패턴은 짧은 순차패턴을 기반으로 생성되기 때문에 길이가 긴 순차 패턴을 마이닝 시, 성능하락이 발생할 수 있다는 점에서 개선이 필요하다.

감사의 글

이 논문은 2009 년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. 2009-0075771)

참고문헌

[1] 마이클 A. 베리고든 S. 린오프, 경영을 위한 데이터 마이닝 : 마케팅과 CRM 활용을 중심으로, 2009

[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rule," Proc. Int'l Conf. Very Large Data Bases, pp. 487-499, Sept. 1994

[3] R. Agrawal and R. Srikant. Mining Sequential Patterns. In Proc. of the 11th Int'l Conference on Data Engineering, Taipei, Taiwan, March 1995.4

[4] Ramakrishnan Srikant; Rakesh AgrawalMining, "Sequential Patterns: Generalizations and Performance Improvements" EDBT'96, 1996