

연관규칙을 활용한 상품 구매 패턴분석에 관한 연구

정용규* · 박정권** · 이정찬*** · 최은영****

목 차

요약	3.1 실험 데이터
1. 서론	3.2 연구실험 및 분석
2. 관련 연구	4. 실험결과
2.1 연관규칙과 분류기법	5. 결론
2.2 소비자 구매행위	참고문헌
3. 실험	Abstract

요약

기업의 데이터베이스 규모는 계속 증가되고 있는 추세이다. 이런 대량의 데이터베이스에서 유용한 정보를 얻는 데이터마이닝은 다양한 예측기법이나 차이점의 분석을 통하여 매출의 증대나, 비용 절감 등의 효과를 줄 수 있다. 분석방법에는 정보의 다양한 분류기법 또한 다양한 각도나 영역의 관점을 통하여 분석할 수 있다. 연관규칙과 패턴의 분석은 대량의 데이터베이스에서 유용한 정보를 효과적으로 분석할 수 있다. 본 논문에서는 쇼핑물의 상품을 데이터마이닝 분석 기법중의 하나인 연관규칙을 활용하여 분석하였다. 이를 통하여 기존상품 분류 및 분석에 따른 고객의 구매패턴을 분석하고, 전략적 마케팅 수립을 통해 접근할 수 있는 데이터마이닝 분석에 관한 사례를 연구하였다.

표제어: 데이터마이닝, 분류기, 연관규칙, 고객관계마케팅, 패턴분석

접수일(2012년 2월 19일), 수정일(2012년 2월 30일), 게재확정일(2012년 3월 20일)

* 을지대학교 의료IT마케팅학과 교수, ygjung@eulji.ac.kr

** 제원인터넷네트웍, successpark@jewon1986.com

*** 한국정보보호진흥원, 교신저자, jclee@nia.or.kr

**** 을지대학교 의료산업학부 의료전산학전공, ohjjong4@nate.com

1. 서론

기존의 데이터베이스 관리 시스템이 취급하고 이에 저장되는 데이터의 양이 증가함에 따라서 저장된 자료로부터 원하는 정보를 효과적으로 추출하거나 데이터에 내재된 규칙들을 찾는 일이 대단히 어렵게 되고 있다. 또한 기업의 ERP 시스템의 정착과 더불어 ERP로부터의 방대한 데이터에 대한 분석의 필요성이 증대되고 있다. 이에 데이터웨어하우징 기술만으로는 향후 고객의 욕구를 충족시키기 어려울 뿐만 아니라 기업들 간의 경쟁심화로 인해 데이터마이닝의 중요성에 대한 인식이 확대되고 활용은 극대화되고 있다.

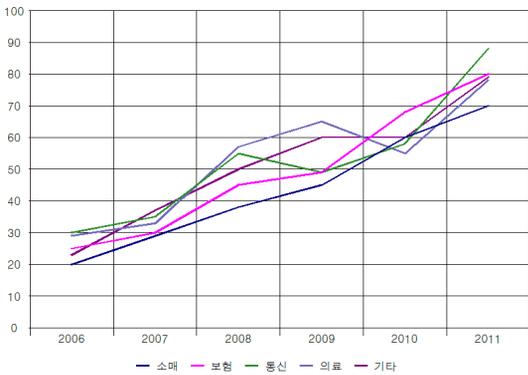


그림 1. 데이터마이닝 응용분야 추세
Fig. 1. Trend of the Data Mining Applications

또한 데이터마이닝은 정보기술의 발달과 더불어 발전되어 왔다. 특히 데이터베이스 기술의 발달과 데이터웨어하우징 기법, 인공지능의 전문가 시스템과 기계학습 등이 데이터마이닝을 발달시키는데 주요한 요인을 제공하였다. 기업경영에서 마케팅 정보에 대한 요구와 기대가 달라진 것도 데이터마이닝을 확산시키는데 중요한 요인이라고 볼 수 있다 현재 광범위한 영역에서 데이터마이닝이 활용되고 있으며, 또한 다양한 부분에 구축되어 있는 자료들을 통합해서 보다 상위수준의 전략적 지식을 도출하기 위해서도

많이 이용되고 있다.

본 논문에서는 쇼핑몰 사이트에서 판매되는 제품 간의 연관관계를 알아보기 위해 데이터마이닝 기법 중 하나인 연관규칙분석을 실시하였다. 분석을 위해서는 데이터마이닝 소프트웨어들 중 하나인 EC-Miner를 사용하였다. 연관관계를 관찰하고자 하는 제품의 수가 증가할수록, 분석을 위해 필요한 계산의 수는 기하급수적으로 증가하므로 본 논문에서는 구매된 제품들 사이의 연관성을 분석함으로써 동시에 구매되는 관계를 알아내어 교차 판매(Cross Selling)를 활용하기 위해 연관성분석을 실시하였다.

2. 관련 연구

2.1 연관규칙과 분류기법

연관규칙은 보통 $A \Rightarrow B$ 로 나타낸다. $A \Rightarrow B$ 는 어떤 고객이 A라는 물건을 사면 B라는 물건도 산다는 것을 의미한다. 여기서, A는 반드시 하나의 물건을 가리키지는 않으며 복수의 물건도 가능하지만, B는 대부분의 경우 하나의 물건이다. 이러한 연관성 규칙을 마케팅 활동에 활용하기 위해서는, 그 규칙에 관련된 상품들이 전체 고객 중 상당 %를 차지하는 고객들의 거래 내역에서 발견되어야 하며(지지도), 동시에 A 상품을 산 고객 중에서 상당 %의 고객이 B상품을 샀어야 한다(신뢰도).

이렇게 찾아낸 연관 규칙에 대해 다시 향상도 값을 계산하여 그 값이 1보다 큰 연관 규칙만을 활용하게 된다. 다음은 지지도, 신뢰도 그리고 향상도에 대한 수식이다.

$$\text{Support}(A \Rightarrow B) = \Pr(A \cap B) \quad (1)$$

(두 항목이 같이 거래된 수 / 전체 거래수)

$$\text{Confidence}(A \Rightarrow B) = \frac{\Pr(A \cap B)}{\Pr(A)} \quad (2)$$

(두 항목이 같이 거래된 수/항목 A의 거래수)

$$\text{Lift}(A \Rightarrow B) = \frac{\text{Pr}(A \cap B)}{\text{Pr}(A) \cdot \text{Pr}(B)} \quad (3)$$

(항목 A의 거래 중 항목 B가 포함된 거래의 비율/
전체거래 중 항목 B가 거래된 비율)

2.2 소비자 구매행위

소비자 구매행동(consumer purchasing behavior)은 교환과정에 참여한 사람의 구매와 소비에 관련된 모든 활동을 총칭하는 것으로, 소비자 행동이란 소비자가 자신의 욕구를 충족시킬 것으로 기대하는 제품이나 서비스 또는 아이디어를 탐색, 구매, 사용, 평가 처리함에 있어서 보여주는 행동이다. 소비자들은 제품이나 서비스 또는 상표에 신념과 가치, 태도, 습관, 기호 등 자신들의 구매와 소비에 영향을 미치는 것들과 구매 및 소매 방법을 학습을 통해 습득하며 이러한 의미에서 소비자 행동은 학습된 행동이라 할 수 있다. 따라서 소비자 행동은 어떠한 제품이나 서비스의 선택과정, 일련의 의사결정과정에 포함되는 모든 요소들을 포괄하여 파악 하여야 한다. 결과적으로 대형마트에 방문하는 고객세분화를 통하여 소비자의 욕구(needs)를 만족 시키도록 소비자의 소비형태 및 의식 변화에 마케팅 4요소인 상품(product), 가격(price), 장소(place), 촉진(promotion)등의 다양한 마케팅 수단을 활용하여 차별화해야 한다.

이러한 측면에서 소비자 행동을 이해하고 연구할 필요가 있으며 포괄적 의미의 개성(personality)이론을 통해 분석 소비자의 잠재적, 본능적 행동을 통해 소비자의 구매행동을 연관하여 파악할 수 있다.

2.2.1 특성이론(Trait-Factor Theory)

알포트(F.Allport)와 카텔(J.Catell)은 개성을 구성하는 각 요소들의 크기에 비교할 수 있다는 계량적 접근방법의 하나라고 볼 수 있으며, 같은 상황에서 개인의 반응과 판단 및 행동은 일관적이며 차별적이라는 사회현상에 기초를 두고 있다.

2.2.2 정신분석 이론(Psychoanalytic Theory)

정신분석학의 시조인 프로이트(S.Freud)는 인간의 행동을 지배하는 근본적인 동기는 무의식적 요소라고 하였다.

첫째, 인간의 행동은 우연히 일어날 것이 아니며 반드시 원인이 있다.

둘째, 인간의 행동은 본능적, 잠재의식적 특성을 내포하고 있다.

셋째, 인간의 사고, 정서 및 행동에는 무의식적 결정인자가 포함되어 있다.

2.2.3 신경신 분석 이론(Neo-Freudian Theory)

신 프로이트 이론은 융(C.Jung), 애들러(Alfred Adler)에 의해 프로이트의 정신분석이론보다 폭넓은 개념으로 사회적 영향이 내부적 요인보다 더 많은 영향을 소비자에게 줄 수 있다고 한다.

2.2.4 자기 개념 이론(Self-Concept Theory)

자기 이미지를 중심으로 인간의 개성을 설명하여 현실속의 정신과 이상적인 자신에 대한 생각을 지니고 자아와 초자아의 개념을 공유하게 되는 행동이론으로 자기개념 결정기준은 즐거움, 의존성, 실용성, 성공 등을 들 수가 있다.

따라서 소비자 구매행동은 위에서 언급한바와 같이 외적인 요소와 환경적인 내적인 요소의 세분화된 경험에 의해서 학습이 되어지는 것이다.

3. 실험

3.1 실험 데이터

실험 데이터로 기존의 상품연관성분석에 쓰였던 데이터를 쇼핑몰 내의 상품들로 구성하여 65,535개의 Record가 있고 Attributes는 6개로 고객, 의류, 신발, 모자, 액세서리, 기타잡화가 있다. 실험 방법으로는 EC-Miner를 활용하여 Aprior 방법으로 의류, 신발,

모자, 액세서리, 기타잡화 사이의 연관성 규칙을 도출하여 상품의 매출을 늘리기 위해 쓰였다.

표 1. 실험 데이터

Tab. 1. The Experimental Data

	1	2	3	4	5
	의류	신발	모자	액세서리	기타잡화
1	1	1	1	0	0
2	1	1	1	1	1
3	0	1	1	0	0
4	0	0	0	1	1
5	0	0	1	0	0
6	0	1	1	0	1
7	1	0	1	0	1
8	0	1	0	1	0
9	1	0	0	1	0
10	0	0	1	0	0
11	1	1	1	0	1
12	1	0	1	0	1
13	1	1	1	1	1
14	1	1	1	0	0
15	0	0	1	1	1
16	0	1	0	1	1
17	0	1	1	1	0
18	0	0	1	1	1
19	0	1	1	0	0
20	0	0	0	1	1

3.2 연구실험 및 분석

실험과정은 EC-Miner를 활용하여 쇼핑몰 사이트 내의 의류, 신발, 모자, 액세서리, 기타잡화 제품 구매 시 '1', 비구매 시 '0'으로 표시하였으며, Apriori 방법으로 많은 량의 데이터를 가능한 시간 내에 처리를 한다. 연관성분석에는 '고객번호'는 필요가 없으므로 'Filter' 노드를 통해 걸러낸다. 연관규칙발견 시행에서 최소지지도 2개, 아이템셋크기 3개, 최소신뢰도 50%의 31개의 Set이 생성되었다.

표 2. 속성별 연관성 분석

Tab. 2. Association Analysis by Attribute.

순번	Set	아이템수	지지도
1	의류, 신발, 모자, 액세서리, 기타잡화	5	4370
2	신발, 모자, 액세서리, 기타잡화	4	6554
3	의류, 신발, 모자, 액세서리	4	4370
4	의류, 모자, 액세서리, 기타잡화	4	4370
5	의류, 신발, 모자, 기타잡화	4	6555
6	의류, 신발, 액세서리, 기타잡화	4	4370
7	신발, 모자, 액세서리	3	13106
8	모자, 액세서리, 기타잡화	3	13107
9	신발, 모자, 기타잡화	3	10924
10	신발, 액세서리, 기타잡화	3	8738
11	의류, 모자, 액세서리	3	4370
12	의류, 신발, 모자	3	13109
13	의류, 신발, 액세서리	3	4370
14	의류, 모자, 기타잡화	3	10925
15	의류, 액세서리, 기타잡화	3	4370
16	의류, 신발, 기타잡화	3	6555
17	모자, 액세서리	2	21843
18	신발, 모자	2	28399
19	신발, 액세서리	2	21843
20	모자, 기타잡화	2	21847
21	액세서리, 기타잡화	2	21844
22	신발, 기타잡화	2	13108
23	의류, 모자	2	17479
24	의류, 액세서리	2	6555
25	의류, 신발	2	13109
26	의류, 기타잡화	2	10925
27	기타잡화	1	30584
28	액세서리	1	41502
29	모자	1	45876
30	신발	1	37136
31	의류	1	19664

아래와 같이 생성된 24개의 연관규칙에서 “[의류, 신발] → [모자] 규칙을 보면 신뢰도 100%, 향상도 1.428로서, 의류와 신발을 구매한 고객이 모자를 함께 구입할 비율이 전체고객의 모자 구매 선택비율에 비해 1.428배임을 알 수 있다.

표 3. 생성된 규칙정보

Tab. 3. Generated Rules Information

규칙	연관규칙	신뢰도 (%)	향상도	지지도 (개)
규칙 1	[모자, 액세서리] → [신발]	60.000	1.0588	13106
규칙 2	[모자, 액세서리] → [기타잡화]	60.005	1.2857	13107
규칙 3	[신발, 기타잡화] → [모자]	83.338	1.1905	10924
규칙 4	[신발, 기타잡화] → [액세서리]	66.661	1.0526	8738
규칙 5	[의류, 모자] → [신발]	74.998	1.3235	13109
규칙 6	[의류, 신발] → [모자]	100.00	1.4285	13109
규칙 7	[의류] → [신발, 모자]	66.664	1.5383	13109
규칙 8	[의류, 액세서리] → [신발]	66.666	1.1764	4370
규칙 9	[모자, 기타잡화] → [의류]	50.006	1.6665	10925
규칙 10	[의류, 모자] → [기타잡화]	62.503	1.3393	10925
규칙 11	[의류, 기타잡화] → [모자]	100.00	1.4285	10925
규칙 12	[의류] → [모자, 기타잡화]	55.558	1.6665	10925
규칙 13	[의류, 액세서리] → [기타잡화]	66.666	1.4285	4370
규칙 14	[신발, 기타잡화] → [의류]	50.007	1.6666	6555
규칙 15	[의류, 신발] → [기타잡화]	50.003	1.0714	6555
규칙 16	[의류, 기타잡화] → [신발]	60.000	1.0588	6555
규칙 17	[모자] → [신발]	61.903	1.0924	28399
규칙 18	[신발] → [모자]	76.472	1.0924	28399
규칙 19	[기타잡화] → [모자]	71.432	1.0204	21847
규칙 20	[기타잡화] → [액세서리]	71.422	1.1278	21844
규칙 21	[액세서리] → [기타잡화]	52.633	1.1278	21844
규칙 22	[의류] → [모자]	88.888	1.2697	17479
규칙 23	[의류] → [신발]	66.664	1.1764	13109
규칙 24	[의류] → [기타잡화]	55.558	1.1904	10925

4. 실험결과

본 논문의 표 3의 연관규칙분석 결과를 바탕으로 모든 제품들이 향상도 값이 모두 '1' 이상으로 의미 있는 관계를 가지고 있음을 나타내고 있다. 그러므

로 쇼핑몰 사이트 디자인 방법 수정방안에서도 메인 화면에서의 추천 상품들도 관리자의 주관 하에 디스플레이 하는 것보다 데이터마이닝 기법인 연관분석을 통해 발견된 규칙을 활용해 판매에 많은 영향을 미치는 제품을 발견하여 메인화면에 소비자들의 시야에 잘 띄는 위치에 배치할 수 있을 것이다.

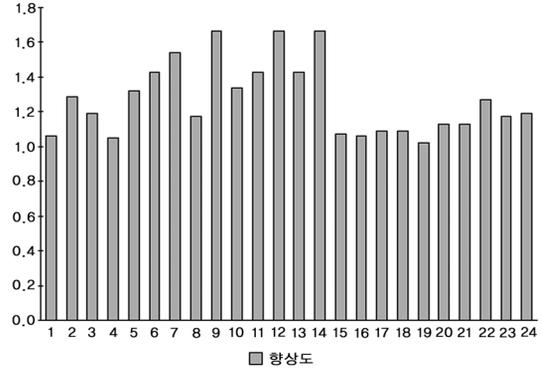


그림 2. 향상도에 따른 연관규칙들의 수
Fig. 2. Number of Association Rules According to Improvements

또 “[의류, 신발] → [모자]”, “[의류, 기타잡화] → [모자]” 규칙을 보면 신뢰도 100%로서 의류, 신발, 모자, 기타잡화의 제품들 간의 관련성을 더욱 자세히 분석하여 이벤트나 교차판매 등을 실시하여 프로모션을 하여 고객들을 선점할 수 있을 것이다.

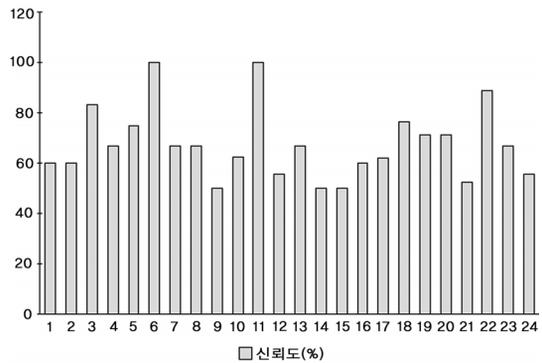


그림 3. 신뢰도에 따른 연관규칙들의 수
Fig. 3. Number of Association Rules According to Levels of Reliability

5. 결론

최근에는 고객의 요구반영뿐만 아니라 쇼핑물 운영에 있어 고객분석과 판매상품의 분석이 이루어지지 않는다면 머지않아 구매자들은 좀 더 나은 서비스와 요구를 충족시켜주는 타 업체로 이동하게 될 것이다. 신규고객을 확보하는 것 보다 기존고객을 유지시키는 것이 비용 적 측면에서 더 저렴하다는 것은 이미 알려진 사실이다. 고객유지를 위해 많은 대형 쇼핑물 업체들은 이미 CRM 프로세스를 마련하여 고객확보에 주력하고 있는 반면 중·소 업체들은 아직 기반조차 마련하지 못한 경우가 많이 있다.

본 논문에서는 쇼핑물의 구매 데이터들을 데이터마이닝 기업의 하나인 연관성규칙방법으로 분석하여 거래되는 상품들 간의 관련성을 발견하고 결과를 통해 쇼핑물의 효율적 운영을 위한 방법들을 제시하고자 하였다.

분석결과를 통해서 사이트 내의 상품매치, 고객 충성도를 높일 수 있는 방안, 제품 카테고리의 재분류 필요성 등에 관하여 논의하였다.

본 논문에서 제시된 운영방안은 쇼핑물 사이트에서 큰 부담이 없이 적용 가능하여 체계적인 고객관계 관리에 일조를 할 수 있다는데 의의가 있다고 하겠다.

참고 문헌

[국내 문헌]

- [1] 김주현, 전민수, 정용규 (2011), “사례 기반 학습을 이용한 주식데이터 분류”, 서비스연구, 1(1), 71-79.
- [2] 이수은, 정용규(2011), “다차원 연관 분석을 이용한 인터넷 이용자의 특징 분석”, 서비스연구, 1(1), 61-69.

[국외 문헌]

- [3] Agrawal, R., Imielinski, T., and Swami, A. (1993), “Mining Association Rules between Sets of Items in Large Databases,” Proceedings of ACM SIGMOD, 207-216.
- [4] Agrawal, R. and Srikant, R. (1994), “Fast Algorithm for Mining Association Rules,” Proceedings of the 20th International Conference on Very Large DataBases, 478-499.
- [5] Agrawal, R. and Srikant, R. (1995), “Mining Sequential Patterns,” Proceedings of the 11th International Conference on Data Engineering, 3-14.
- [6] Park, J. S., Chen, M. S. and Yu, P. S. (1995), “An Effective Hash-Based Algorithm for Mining Association Rules,” Proceedings of ACM SIGMOD, 175-186.



정 용 규 (Yong Gyu Jung)

서울대학교, 연세대학교, 경기대학교에서 각각 학사, 석사, 박사학위를 취득하였고, 현재 을지대학교 의료IT마케팅학과 교수로 재직 중이다. ISO/TC154, UN/Cefact의 한국대표위원으로 활동하고 있으며, 의료정보, 전자무역, 해상물류, 금융전산에 Semantic Web, Process Modelling, ebXML 등의 표준기술의 적용에 관심이 많다.



박 정 권 (Jeong Kwon Park)

환경대학교 식품생물공학과를 졸업하고 현재 삼풍비엔에프(주)에 근무하고 있으며, 비즈니스 인텔리전스를 연구하고 있다. 관심분야로는 유통 분야에서 의사결정지원을 위한 다양한 결정요인을 마이닝기법을 통해 실험하고 연구하고 있다.



이 정 찬 (Jeong Chan Lee)

안양과학대학 전자계산학과를 졸업하고 현재 한국정보화진흥원 창의 인재부에서 근무하고 있으며, 관심분야로는 대형 네트워크망과 다수의 서버환경에서 정보보안을 연구하고 있다. 특히 트래픽 필터링을 위해 행위기반과 연관규칙을 이용한 패턴분석을 연구하고 적용을 실험하고 있다.



최 은 영 (Eun Young Choi)

을지대학교 의료산업학부를 졸업하고 현재 회사에 근무하고 있으며, 의료정보시스템에 관심이 많으며, 특히 의료분야에서 의사결정지원을 위한 다양한 결정요인을 마이닝기법을 통해 실험하고 연구하고 있다.

An Study on the Product Purchase Patterns using Association Rule

Yong Gyu Jung* · Jeong Kwon Park** · Jeong Chan Lee*** · Eun Young Choi****

ABSTRACT

It is growing in size of database in companies. This caused to develop data mining techniques to predictive information from the large database. Costs and other effects can give variety of sales exploding through the analysis of the differences. Analysis of the various classification techniques, various angle can be analyzed point of view of the area information. The analysis of rules and patterns associated with a large amount of useful information from the database can be analyzed effectively. Goods store were analyzed using association rules, one of the data mining analysis techniques. Through this type of existing products according to analyze customer buying patterns, data mining has been studied to establish strategic marketing analysis.

Keywords: Data Mining, Classification, Association Rule, CRM Marketing, Pattern Recognition

* Department of Medical IT and Marketing, Eulji University, Professor, ygjung@eulji.ac.kr

** Jewon International, successpark@jewon1986.com

*** NIA, National Information Society Agency, Corresponding Author, jlee@nia.or.kr

**** Department of Medical Computer Science, Eulji University, ohjjong4@nate.com