






IT와 비즈니스혁신

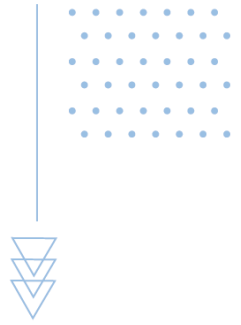
W12. 마이닝 기법 Ⅲ: 연관성 분석





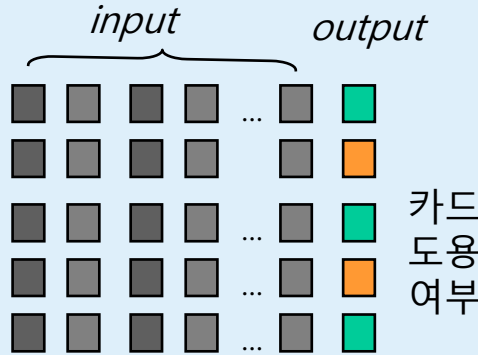
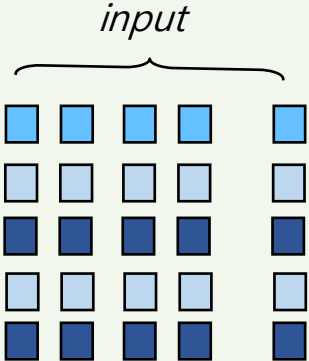
Contents

- I. 연관성 분석 개요
 - II. 연관성 분석 원리
 - III. 활용 사례
 - IV. 정리
- 
- 
- 



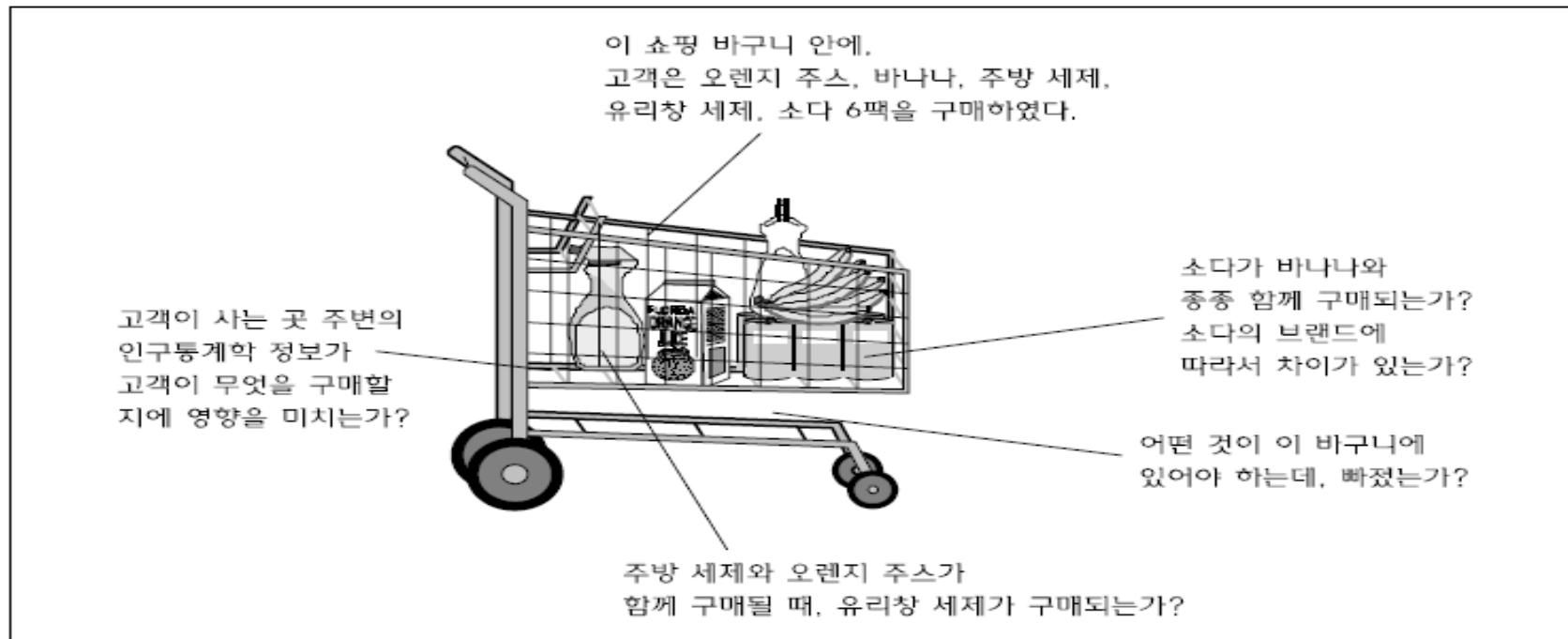
1. 연관성 분석 개요

데이터 마이닝은 크게 **출력 변수**의 존재 여부에 따라 **지도학습과 자율학습**으로 나눌 수 있음

	지도학습 (supervised learning)	자율학습 (unsupervised learning)
의미	<ul style="list-style-type: none"> 입력 데이터와 정답(Label)을 제공 받아 이를 통해 입력(독립)과 출력 (Label, 종속,타겟) 으로 매칭할 수 있는 규칙 생성 <p>예. 카드번호, 성별, 나이, 거래 내역 등 →  카드도용 여부</p>	<ul style="list-style-type: none"> 외부에서 정답(Label)이 주어지지 않음 입력 데이터에서 패턴을 찾아내는 작업 <p>예. 군집화: 주어진 데이터를 3 개의 그룹으로 나눔 </p>
특징	출력 변수가 존재함	출력 변수가 존재하지 않음
분석 기법	의사결정나무, 회귀분석, 인공신경망, 판별분석 등	군집분석, 연관성 분석 등

■ 데이터 안에 존재하는 항목들 간의 **연관규칙**을 발견하는 과정

- 1993년 IBM Almaden 연구소에서 처음 시도된 데이터 분석 방법론
- 동시에 구매될 가능성이 큰 상품들을 찾아냄으로써 시장 바구니 안의 구매 물품간의 관계를 찾는다는 의미에서 장바구니 분석 (Market Basket Analysis) 이라고도 함



* Source: Berry et al., 김종우, 경영을 위한 데이터 마이닝, 2018.

일련의 거래나 사건들의 연관성에 대한 규칙

일반적으로 알아낼 수 없는 규칙 → 실제 적용으로 좋은 결과 기대
상식적으로 널리 알려진 관련성 → 이의 발견은 큰 의미가 없다.

{Item A} → {Item B}

Item A: if절, 조건절, 선행(antecedent), 전제(premise)

Item B: then절, 주절, 후행(consequent), 결론(conclusion)

예. 빵과 우유를 구매하면 주스도 함께 구매한다.

신발을 구매한 고객은 양말도 동시에 구입한다.

연관 규칙의 유형

- **실행 가능한 규칙(Actionable Rules)** : 유용한 정보를 가지고 실제로 행동을 취할 수 있는 규칙
 - 목요일 저녁에 식료품 가게를 찾는 고객은 아기 기저귀와 맥주를 함께 구입하는 경향이 있다.
- **사소한 규칙(Trivial Rules)** : 해당 분야에 익숙한 사람이라면 누구나 이미 알고 있는 규칙
 - 한 회사의 전자제품을 구매하던 고객은 전자제품을 살 때 같은 회사의 제품을 사는 경향이 있다.
- **설명 불가능한 규칙(Inexplicable Rules)** : 설명할 수 없고, 실제로 행동을 취할 수도 없는 규칙
 - 새로 연 건축 자재점에서는 싱크대가 많이 팔린다.

■ 비즈니스 의사 결정 지원

- 마케팅 : 상품의 패키징화, 효율적인 매장 진열, 교차판매 전략, 기획 상품의 결정 등에 응용
- 활용 예시
 - 구매 상품의 연관성을 분석하여 상품 A와 함께 구매할 가능성이 높은 상품을 추천, 쿠폰이나 카탈로그 제공
 - 상품 진열시 구매 가능성이 높은 상품을 함께 배치, 구매 활성화 및 고객의 동선 축소로 만족도 향상
 - 모바일에서 주로 이용되는 유료 서비스들 간의 연관규칙을 파악하여 화면 설계 및 번들 상품 개발을 통해 판매 촉진
 - 신용카드, 대출 등의 은행 서비스 내역을 기반으로 특정 서비스 구매 가능성이 높은 고객 예측
 - 환자의 의무기록에서 여러 치료가 같이 이루어진 경우 합병증 발생의 징후 예측

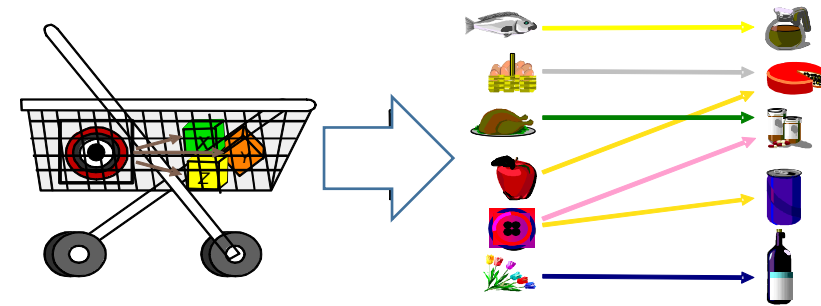
■ 연관성 분석을 활용한 대표적인 마케팅 방법

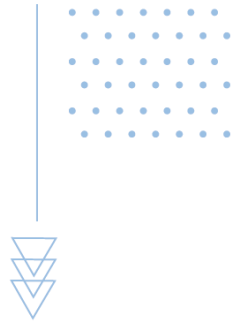
● 교차 판매 (Cross Sell)

- 어떤 상품을 구입한 고객에게 다른 상품도 판매하는 전략
- 예. PC 구입시에 프린터, 스피커 등 다른 제품까지 구매하게 되는 경우

●상향 판매 (Up Sell)

- 어떤 상품을 구입한 고객에게 보다 고급의 상품을 판매하는 전략
- 예. PC 구입시 더 좋은 사양의 PC를 고르게 되는 경우

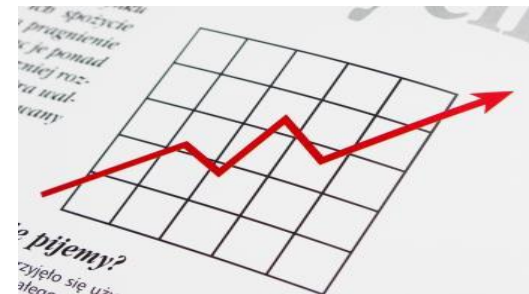




II. 연관성 분석 원리

■ 연관 규칙의 생성

- 유용한 연관 규칙을 찾기 위해서는 규칙에 의해 실제로 얼마나 거래가 되는지 빈도수를 측정
- 연관 규칙 생성을 위한 데이터
 - 고객 정보, 주문정보, 물품 정보
- 빈도수를 기반으로 연관 정도를 정량화 하기 위한 지표
 - 지지도 (Support): 연관규칙의 중요성에 대한 척도
 - 신뢰도 (Confidence): 연관규칙의 신뢰성에 대한 척도
 - 향상도 (Lift, Improvement): 연관규칙이 임의 추측보다 얼마나 더 예측력을 갖는지 평가하는 척도



■ 연관성 분석을 위한 거래 데이터

- 거래 유무를 나타내는 이진 범주형으로 변환

주문번호	구매상품 리스트
K07001	오렌지 주스, 소다
K07002	우유, 오렌지 주스, 유리창 세제
K07003	오렌지 주스, 주방 세제
K07004	오렌지 주스, 주방 세제, 소다
K07005	유리창 세제, 소다

POS 거래데이터 (원 데이터)



주문번호	오렌지 주스	유리창 세제	우유	소다	주방 세제
K07001	1	0	0	1	0
K07002	1	1	1	0	0
K07003	1	0	0	0	1
K07004	1	0	0	1	1
K07005	0	1	0	1	0

분석을 위한 데이터

- 간단한 구매 패턴 파악

- 오렌지 주스와 소다는 다른 두 물품들에 비해 함께 잘 팔리는 경향이 있다.
- 주방 세제는 유리창 세제나 우유와는 같이 팔리는 경우가 없다.
- 우유는 소다 혹은 주방세제와 같이 팔리는 경우가 없다.

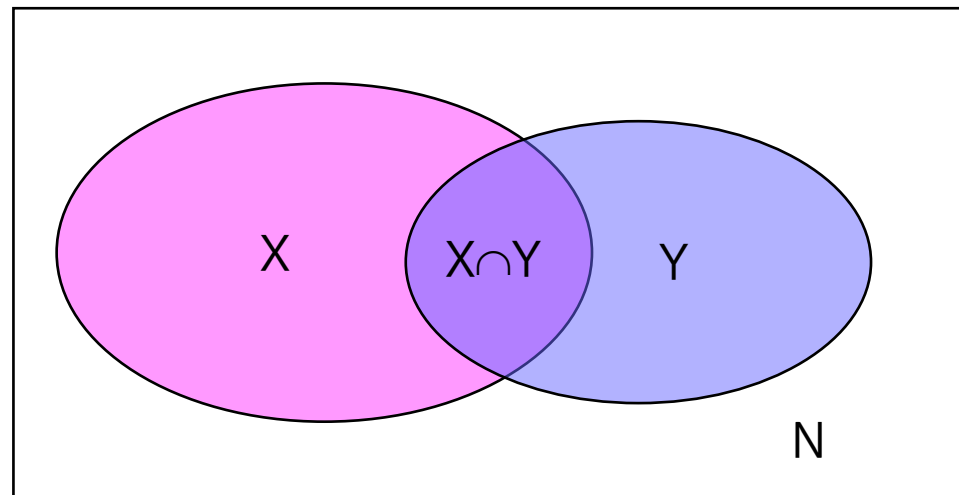
	오렌지 주스	유리창 세제	우유	소다	주방 세제
오렌지 주스	4	1	1	2	1
유리창 세제	1	2	1	1	0
우유	1	1	1	0	0
소다	2	1	0	3	1
주방 세제	1	0	0	1	2

상품의 동시발생 테이블

지지도(Support)

- 연관규칙의 중요성에 대한 척도
- 전체 거래에서 X와 Y를 모두 포함하고 있는 거래의 비율, 대칭적
- X와 Y를 동시에 포함하는 거래수/전체거래수
- $\text{Support}(X \rightarrow Y) = \Pr(X \cap Y)$
- $\text{Support}(\text{소다} \rightarrow \text{오렌지 주스})$
 $= P(\text{소다} \cap \text{오렌지 주스}) = 2/5 = 40\%$

주문번호	구매상품 리스트
K07001	오렌지 주스, 소다
K07002	우유, 오렌지 주스, 유리창 세제
K07003	오렌지 주스, 주방 세제
K07004	오렌지 주스, 주방 세제, 소다
K07005	유리창 세제, 소다



신뢰도(Confidence)

- 연관규칙의 신뢰성에 대한 척도, X를 구매한 경우 이 중에서 얼마나 Y구매로 이어지는지를 의미
- X가 포함된 거래 내에서 Y를 포함하고 있는 거래의 비율, 비대칭적

주문번호	구매상품 리스트
K07001	오렌지 주스, 소다
K07002	우유, 오렌지 주스, 유리창 세제
K07003	오렌지 주스, 주방 세제
K07004	오렌지 주스, 주방 세제, 소다
K07005	유리창 세제, 소다

- $\text{Confidence}(X \rightarrow Y) = P(X \cap Y) / P(X) = P(Y|X)$
- $\text{Confidence}(\text{소다} \rightarrow \text{오렌지 주스}) = P(\text{소다} \cap \text{오렌지 주스}) / P(\text{소다}) = (2/5) / (3/5) = 2/3$
 - 소다를 구매할 때 오렌지주스도 같이 구매한다.
 - 소다를 구매한 3개의 거래 중 2개의 거래가 오렌지주스를 포함 (67%)
- $\text{Confidence}(\text{오렌지 주스} \rightarrow \text{소다}) = P(\text{오렌지 주스} \cap \text{소다}) / P(\text{오렌지 주스}) = (2/5) / (4/5) = 1/2$
 - 오렌지 주스를 구매한 4개의 거래 중 2개의 거래가 소다를 포함 (50%)

■ 향상도(Lift)

- 연관규칙이 임의 추측보다 얼마나 더 예측력을 갖는지 평가하는 척도
- 임의로 Y를 구매한 경우에 비해 연관 규칙 (X→Y)에 의해 Y를 구매하는 경우의 비율
- 향상도 값이 1보다 크면 예측력이 있다고 간주, $P(Y|X) > P(Y)$
- $Lift(X \rightarrow Y) = P(Y|X)/P(Y) = P(X \cap Y)/(P(X) \times P(Y))$
- $Lift(\text{오렌지 주스} \rightarrow \text{소다}) = \text{Confidence}(\text{오렌지 주스} \rightarrow \text{소다}) / P(\text{소다})$
 $= P(\text{오렌지 주스} \cap \text{소다}) / (P(\text{오렌지 주스}) \times P(\text{소다}))$
 $= (1/2) / (3/5) = 5/6$

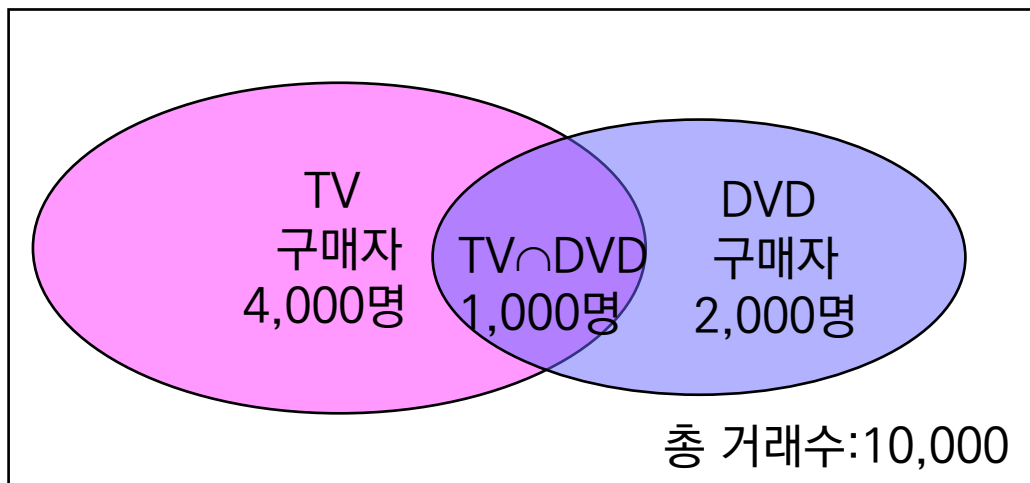
주문번호	구매상품 리스트
K07001	오렌지 주스, 소다
K07002	우유, 오렌지 주스, 유리창 세제
K07003	오렌지 주스, 주방 세제
K07004	오렌지 주스, 주방 세제, 소다
K07005	유리창 세제, 소다

향상도 > 1 → 양의 연관관계

향상도 = 1 → 두 품목은 독립

향상도 < 1 → 음의 연관관계

■ 연관 규칙: TV를 사면 DVD를 함께 산다.



연관규칙 해석

- 지지도: TV와 DVD를 모두 구매하는 비율이 10%임
- 신뢰도: TV를 구매한 고객 중 DVD를 구매한 고객의 비율이 25%임
- 향상도: DVD를 포함하는 거래가 언제 일어날 지를 예측하는데, TV를 구매한 조건이 있는 경우가 조건이 없는 경우보다 1.25배 우수함.

- 지지도 = $1,000 / 10,000 = 10\%$
- 신뢰도 = $1,000 / 4,000 = 25\%$
- 향상도 =
$$\frac{(1,000 / 10,000)}{(4,000 / 10,000) \times (2,000 / 10,000)} = 1.25$$

- 향상도 값이 1보다 크므로 “TV를 사면 DVD를 함께 산다” 는 의미 있는 연관 규칙으로 생각할 수 있음

1. 품목 선택

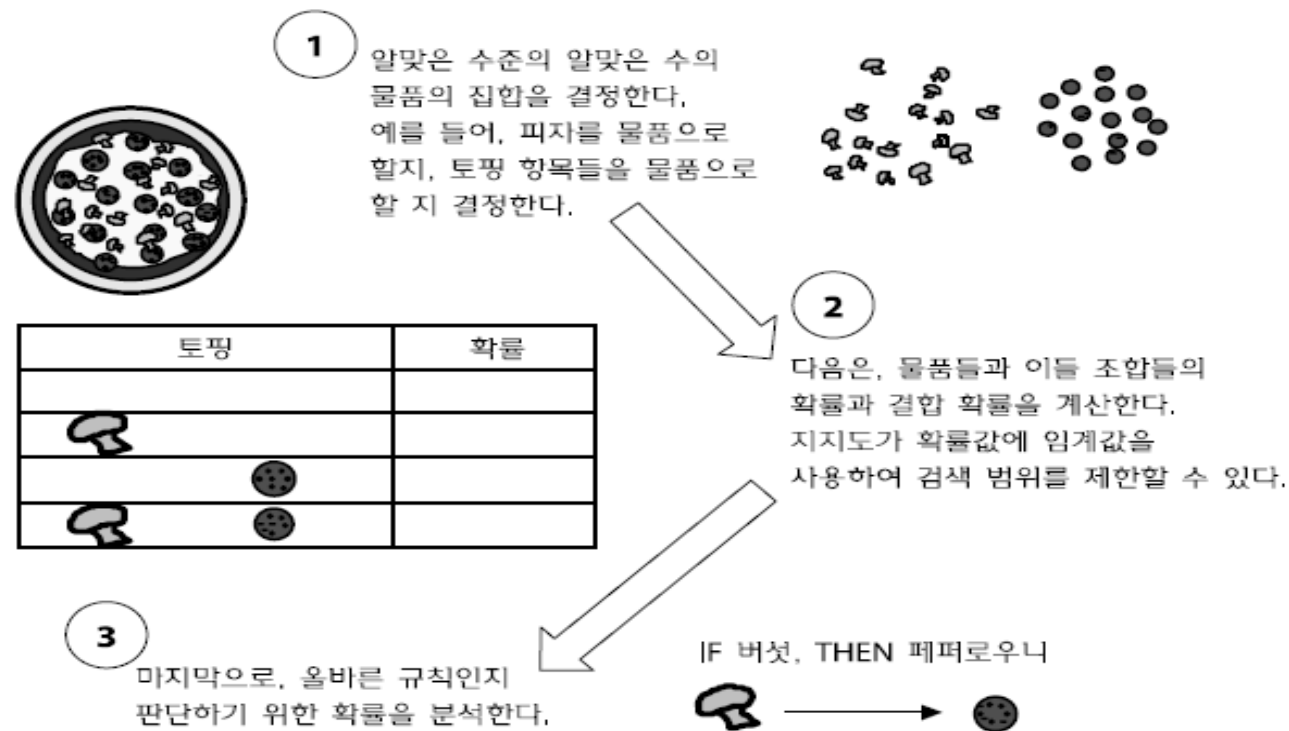
- 적절한 상세 수준의 품목 선택
- 품목 조합의 수에 따라 규칙이 복잡해질 수 있음
- 관심 있는 물품이 시간에 따라 달라질 수 있음

2. 연관 규칙 발견

- 최소 지지도와 신뢰도 기준을 만족하는 규칙 도출
- 자주 발생하는 항목집합(빈발항목집합)들로 후보자 목록 생성
- 빈발항목 집합을 찾는 효과적인 알고리즘으로는 선형적(Apriori) 알고리즘과 빈발패턴(Frequent Pattern) 성장 알고리즘이 가장 대표적

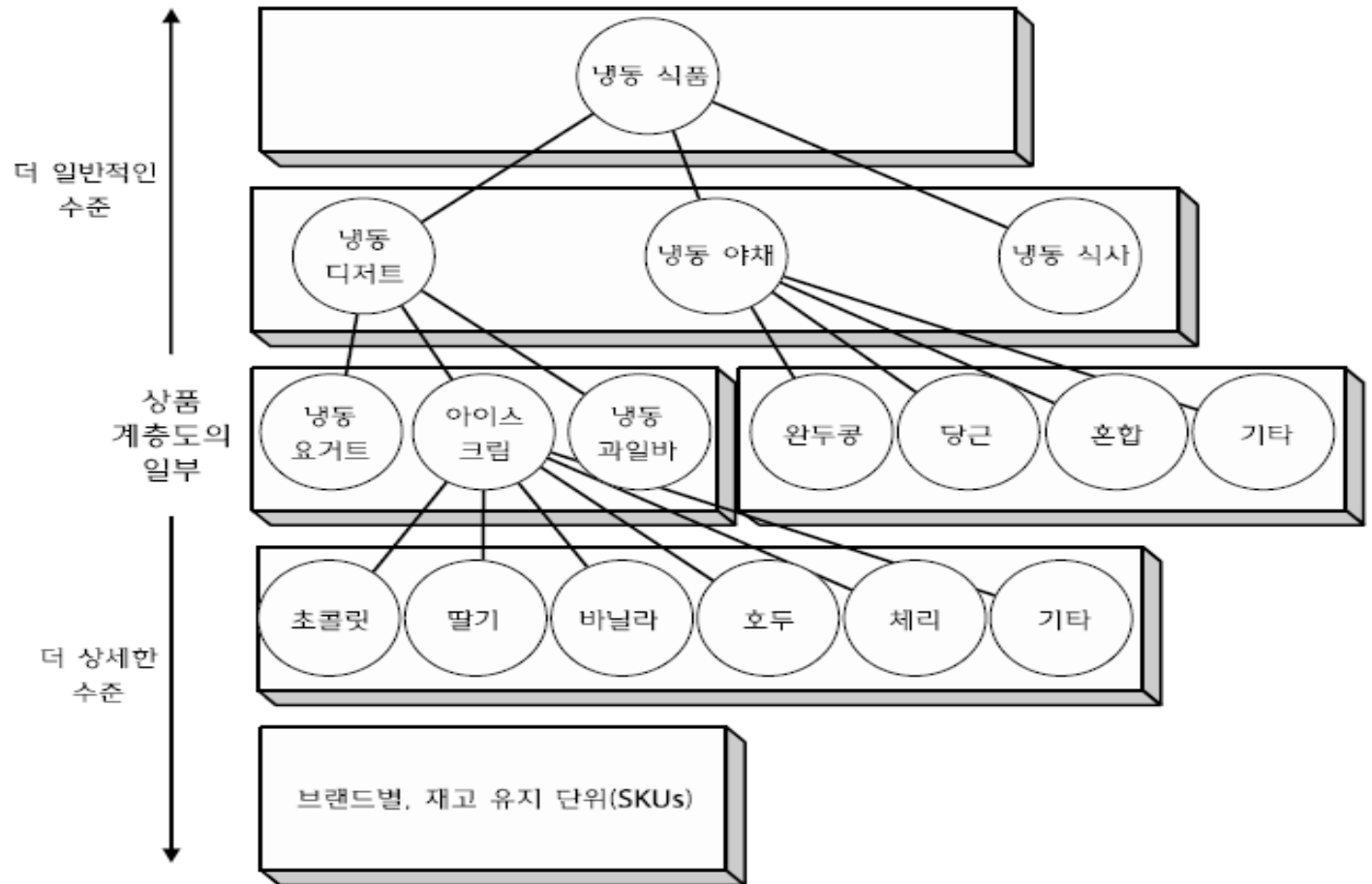
3. 연관 규칙 평가

- 연관 규칙이 의미가 있는지 확인 (향상도 값이 1보다 크다면 의미가 있음)



❏ 품목 선택 – 품목 계층도

- 가장 일반적인 것에서 시작하여 점차 상세한 것으로 이동
- 보다 구체적인 결과일수록 결과들이 행동으로 옮겨질 가능성이 높음
- 상품 계층을 올라가면 올라갈수록 물품의 수는 줄어들음
- 일반화된 물품들은 충분한 지지도를 가진 규칙을 찾는 데 도움이 됨
- 어떤 물품들이 일반화되었다는 것은 모든 물품들이 같은 수준으로 올라가야 한다는 것을 뜻하지는 않음
- 물품의 종류, 행동 가능한 결과를 만드는 일에 대한 기여도, 데이터 내의 빈도 등에 의존

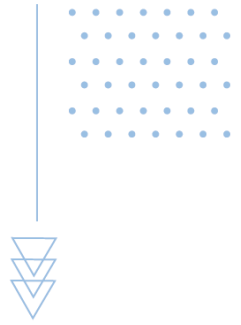


시차를 두고 일어나는 사건들의 연관성 분석

- 순차적으로 일어나는 사건들의 연관성 분석
 - 예1. 새로운 집을 가진 주인들은 가구를 사기 전에 샤워 커튼부터 산다.
 - 예2. 새로운 잔디 깎는 기계를 구입한 고객들은 그 후 6주 내에 새로운 정원 호스를 구입할 가능성이 매우 높다.
 - 예3. 고객이 은행에 들어가 계좌 정리를 요구한다면, 그 사람이 모든 계좌들을 해지할 가능성이 상당히 높다.

시계열 분석을 위한 거래 데이터의 특징

- 식별 정보(Identification information): 계좌번호, 고객 ID 등
- 거래의 시각(timestamp) 또는 순서(sequencing information) 정보
 - 예. 고객의 거래 시간을 고려하여 날짜별로 거래의 순열로 표현하고 몇 % 이상 공통으로 나타나는 순열을 찾음
{사과, 우유}, {사과, 신문}, {음료수, 과자}



III. 연관성 분석 활용 사례

미국 블루밍데일 백화점 사례

- 화장품을 구매한 여성 고객은 핸드백을 구매한다는 것을 파악
- 핸드백 매장과 화장품 매장을 함께 진열해 놓음으로써 매출 증가

국내 C 통신사 사례

- 모바일 결제 데이터를 이용하여 모바일 유료 서비스 사용 실태를 조사
- 모바일 콘텐츠들 간의 연관 규칙 도출
 - 드라마 영상 정보를 구입한 고객 중 33%는 드라마에 출연한 배우 사진도 구매함
 - 출연 배우 사진을 구입한 고객 중 17%는 음원도 구입
 - 최소 지지도 만족, 향상도 2 이상
 - 드라마 영상 정보와 주연 배우들과의 관련 콘텐츠를 엮어서 패키지 상품 판매하고 패키지 상품을 구입한 고객에게 드라마 OST 할인 쿠폰 발송 (교차 판매 전략)



국내 트리플 여행 플랫폼 사례

- 트리플은 AI와 빅데이터에 기반한 여행 준비 추천 시스템 ‘조이’와 사용자 커뮤니티 서비스 ‘라운지’를 도입(2020.4)
- ‘조이’는 사용자가 여행을 준비하며 해야 할 일을 개인 여행 스타일에 따라 단계별로 추천해주는 AI 시스템
 - 사용자가 자신의 여행 계획과 스타일을 알려주면 AI가 기존 여행자들의 빅데이터를 분석해 개별 여행 스타일에 맞는 장소, 호텔, 일정, 투어·액티비티 등 여행 정보와 상품을 추천



■ Stitch Fix의 비즈니스 모델 (2011~)

- “소비자는 자신에게 어울리는 단 한 벌의 청바지를 찾고 싶어 하지, 수많은 선택권을 원하지 않는다.” (Katrina Lake, Stitch Fix CEO)
- 고객이 가입시 작성하는 정보 (신체조건, 취향, 선호 브랜드, 지출 예산 등)를 바탕으로 알고리즘이 추천 목록을 만들고, 이 중에서 전문 스타일리스트가 선택한 5가지 품목 배송
- 마음에 드는 옷이나 액세서리 구매, 그렇지 않은 품목들은 다시 반송 봉투에 넣어 돌려 보내는 방식 (모두 구매할 경우 25% 할인)
- 배송된 품목에 대한 피드백을 받아 이를 다시 추천 품목을 찾는데 이용



“We differentiate ourselves through personalization.”

Stitch Fix – 데이터 분석

■ 데이터 분석 알고리즘과 전문 스타일리스트

- 전문 스타일리스트 3700명을 고용하고 있지만 Stitch Fix의 원천 기술은 패션이 아닌 분석 알고리즘에 있음
- 데이터 과학은 Stitch Fix의 기업문화 그 자체임 – 전통적인 조직구조에 데이터 과학을 추가했다기보다는 데이터과학을 중심으로 사업을 시작, 회사의 알고리즘을 고객의 필요에 맞추어 구축
- 2012년 8월, 넷플릭스의 데이터 과학·엔지니어 부회장 에릭 콜슨 영입 (최소 알고리즘 책임자)
- 100여 명의 데이터 과학자 고용 – 수학, 신경과학, 통계학, 천체물리학과 같은 정량적인 분야 박사 학위자
- 회사에 투자할 수 있는 1달러가 주어지고 마케팅과 제품, 데이터 과학 중에서 선택해 투자할 수 있다면 항상 데이터과학을 선택
- 고객이 구매하려고 하는 블라우스와 어울리는 값비싼 벨트를 사라고 권하는 업셀링 방법을 사용하지 않으며, 이전에 구매한 특정 브랜드도 강요하지 않음
- 데이터와 인간의 전문적인 판단력과 결합해 개성 있고 개인적인 상품 추천

Stitch Fix – Mix & Match (연관성 분석 활용)

FIX 1



The client's style profile guided both the algorithm's choice of this shirt and the stylist's choice of pale pink. ✓



The stylist approved the algorithm's choice of this all-season top, even though it's out of the stated price range, because the client likes florals. ✓



These slip-on sneakers have a high match rate among clients looking for a casual shoe. The stylist thought the floral pattern would add originality. ✓



The client asked for skinny jeans. The stylist selected green from among the algorithm's denim recommendations. ✗



Because the client's style profile said she loves textures, the stylist chose this studded blouse. ✗

✗ RETURNED ✓ BOUGHT

FIX 2



The client was looking for a versatile top. The algorithm identified this cashmere sweater because it has been extremely successful with women of her age and physical dimensions. ✓



The client did not like the fit of the green jeans, so the algorithm found a pair that fit better, and the stylist chose blue denim. ✗



The client loved the lightweight floral top in the previous box, so the stylist found this more vibrant variation, which the algorithm suggested would fit well. ✓



The client also loved the pink shirt in the previous box, so the stylist found a different take within the same color palette. ✓



The client wanted a new bag, and the algorithm found this one trending among women of her age. The stylist picked light green to pop against the red palette of the tops in the box. ✗

FIX 3

Because the client kept the cashmere sweater from the previous Fix, the stylist thought this piece, a little bolder, was worth taking a risk on. ✓



The algorithm chose this popular coat for its versatility and affordability. ✓

Stitch Fix now knows the client's preferred color and fit for jeans, so the stylist felt confident in exceeding her price range with this pair. ✓

The algorithm recommended this blouse because the client responded warmly to the color palette in the previous Fix. ✓

The stylist knows that the client is single and dating, so she chose these playful heels to dress up the skinny jeans. ✓

Stitch Fix – 성공 요인

■ 데이터 과학자와 CEO의 소통

- 대부분 데이터 과학자들은 기술 책임자나 재무 쪽에 보고를 하지만, Stitch Fix에서는 CEO에게 직접 보고함
- 마케팅과 엔지니어링과 같은 다른 부서들도 데이터과학 팀과 긴밀하게 협력
- 알고리즘에 의한 추천이 얼마나 뛰어난가에 따라 매출이 결정됨 – 기업의 가치관과 전략방향을 나타냄

■ 데이터 과학이 혁신을 이루어 냄

- 상품 추천 뿐 아니라 고객이 원하는 새로운 상품 개발 (수요가 있지만 재고가 없는 제품들)
- 옷의 유형에 따라 30~100개의 치수를 측정하고 고객이 느끼는 불편 정도, 가슴둘레와 셔츠 폭의 최적 비율 등 보다 세부적인 맞춤이 가능하게 함

■ 사람이 중요함을 잊지 않음

- 쇼핑은 본질적으로 개인적이고 인간적인 행위이므로, 알고리즘이 이해할 수 없는 다양한 상황과 감성은 스타일리스트들이 고려함 (이벤트, 취직 등에 맞추어 창의적인 추천, 이를 통해 브랜드 충성도를 이끌어 낼 수 있음)
- 인간과 알고리즘의 결합이 최고의 스타일리스트나 최고의 알고리즘만 이용할 때 보다 뛰어남

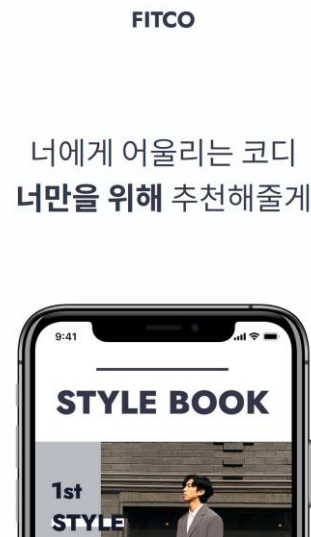
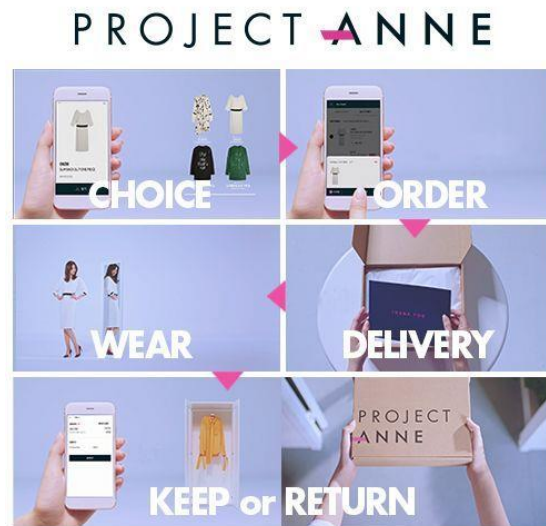
Stitch Fix – 유사 서비스

● 구매와 대여 서비스

- Stitch Fix 의 비즈니스 모델은 기본적으로 제품 구매를 기반
- Rent the Runway의 모델은 대여를 기반

● 국내 유사 서비스

- SK 플래닛의 의류 공유 서비스 프로젝트 앤 (2016.9~2018.5)
- 안 입는 옷을 빌려주고 필요한 옷을 빌려 입는 ‘클로젯셰어’
- 25~34세 남성 대상 스타일 추천 서비스 핏코 (2018.7~)



Market Summary > 스티치 픽스

3.70 USD

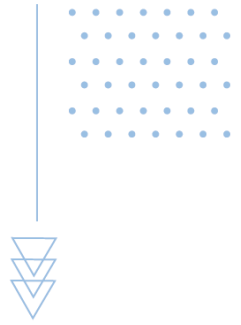
-11.45 (-75.58%) ↓ all time

Closed: Nov 18, 6:37 PM EST • Disclaimer

After hours 3.60 -0.10 (2.70%)

1D | 5D | 1M | 6M | YTD | 1Y | 5Y | Max





IV. 정리



■ 연관규칙의 평가 기준

- 지지도, 신뢰도, 향상도

■ 연관성 분석의 장단점

- 조건식 (if-then)으로 표현되는 연관 규칙이 이해하기 쉽고 실제 적용이 용이함
- 사전에 분석 방향이나 분석 목적이 특별히 없는 경우 유용함
- 데이터의 구조가 복잡하지 않으며 분석을 위한 계산이 비교적 간단함
- 데이터 형식이 제한적임: 거래 유무를 나타내는 이진 범주형 변수
- 품목수가 증가하면 분석에 필요한 계산은 기하급수적으로 늘어남
- 너무 세분화된 품목을 가지고 연관성 규칙을 찾으면 의미 없는 분석이 될 수도 있음
- 제품 판매 이익이나 판매수량과 같이 비즈니스적으로 중요한 요소를 고려할 수 없음