orange 활용 데이터 분석 및 머신 러닝

12차시

장바구니 분석을 활용한 연관분석의 이해

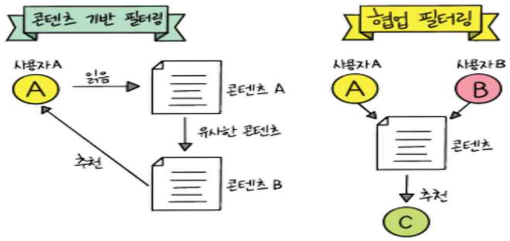
연관규칙 분석
(Association Analysis)

연관규칙 분석이란?

- 연관 분석 또는 연관 규칙(association rule) 학습은 대형 데이터베이스에서 변수 간의 흥미로운 관계를 발견하기 위한 규칙-기반 기계 학습 방법
- 대량의 트랜잭션 정보(예: 고객의 쇼핑 이력)로부터 개별 데이터(변수) 사이에서 항목 간의 관련성-연관규칙(x면 y가 발생)을 찾는 것.
- 슈퍼마켓의 구매내역에서 특정 물건의 판매 발생 빈도를 기반으로 'A물건을 구매하는 사람들은 B물건을 구매하는 경향이 있다.'라는 규칙을 찾을 수 있다. 다른 말로 장바구니 분석(Market Basket Analysis)이라고 한다.

대량의 거래 데이터
- 데이터 탐색
- 데이터 가공

함께 구매되는 상품 찾기
- 각 거래 단위로 함께 구매되는 조합이 높은 것 찾기
- 측정 지표와 알고리즘 필요



콘텐츠 기반 필터링과 협업 필터링

고객별 맞춤 추천 서비스
- 영화 추천
- 쇼핑상품 추천

출처 : 경영자를 위한 디지털 전략 가이드, 스마투스 비즈니스 리뷰(http://www.sbr.ai)

연관규칙 분석이란?

- 콘텐츠 기반 추천(Contents based Recommendation)의 기본 방법론
- 시간과 메모리를 절약하면서 효과적으로 조합을 찾아내는 알고리즘이 중요
- 고객별 맞춤 추천을 할 수 있게 된다.
- 아이템A를 구입한 후에 B를 구매: 서열 분석(Sequence Analysis)
- 연관성 규칙의 예
  - 목요일 식료품 가게를 찾는 고객은 아기 기저귀와 맥주를 함께 구입하는 경향이 있다.
  - 한 회사의 전자제품을 구매하던 고객은 전자제품을 살 때 같은 회사의 제품을 사는 경향이 있다.
  - 새로 연 건축 자재점에서는 변기덮개가 많이 팔린다.
    - 첫 번째 규칙은 유용한 규칙으로 이를 이용하여 식료품 가게의 매출을 증가시킬 수 있다.
    - 두 번째 규칙은 자명한 규칙으로, 대부분의 사람들이 이미 알고 있다. 기존의 정보를 재 확인 하는 의미가 있다
    - 세 번째 규칙은 설명이 불가능한 규칙이며, 좀더 세밀한 조사가 필요하다.

https://blog.naver.com/dear_inwoo/110129191704
출처 : 경영자를 위한 디지털 전략 가이드, 스마투스 비즈니스 리뷰(http://www.sbr.ai)

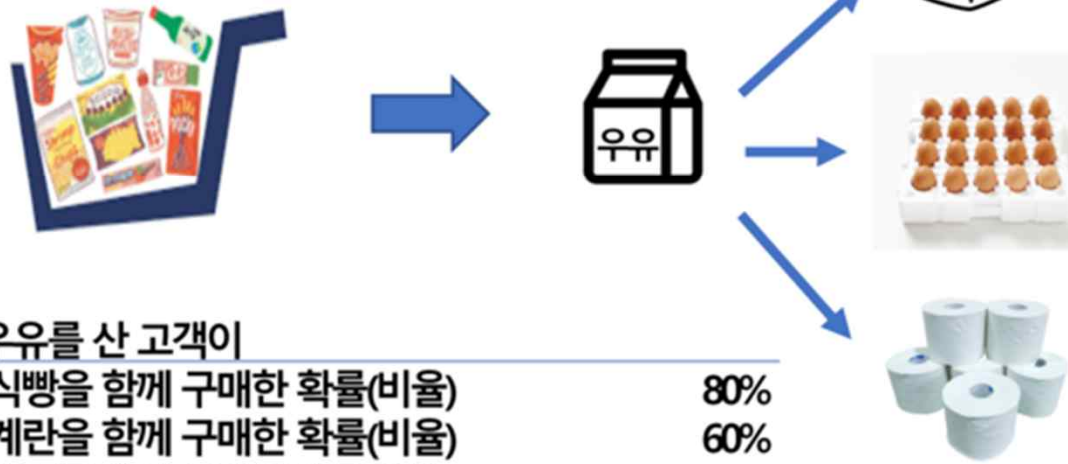
연관 규칙 분석은 왜 비지도 학습인가?

- 지도학습에서 학습한 바와 같이 어떤 문제에 대한 '정답'에 해당하는 사전정보가 없는 상태(비지도 상태, 선생님이 답을 알려 주지 않는 상황)에서 유용한 정보나 패턴을 탐색적으로 발견하기 때문이다.
- 비지도 학습은 목적변수(혹은 반응변수, 종속변수, 목표변수, 출력값)에 대한 정보 없이 학습이 이루어지며, 예측(회귀/분류)의 문제보다는 주로 현상의 기술(Description)이나 특징 도출, 패턴 도출 등의 문제에 활용된다.

출처 : 경영자를 위한 디지털 전략 가이드, 스마투스 비즈니스 리뷰(http://www.sbr.ai)

장바구니 분석

연관규칙분석은 일종의 아래와 같은 조건부 확률을 구하는 것



우유를 산 고객이

식빵을 함께 구매한 확률(비율)	80%
계란을 함께 구매한 확률(비율)	60%
휴지를 함께 구매한 확률(비율)	45%

연관 규칙 중 무엇이 좋은 규칙인가?

◦ dataset=[
['식빵','우유'],
['생수','우유','계란','고등어'],
['우유','사과', ' 휴지']
]

◦ 규칙
- 우유를 산 사람은 식빵을 산다.
- 우유를 산 사람은 계란을 산다.
- 우유를 산 사람은 휴지를 산다.

.....

◦ 어떤 규칙이 좋은 규칙인지 어떻게 판단할 수 있을까?

좋은 규칙을 판단하는 세가지 지표 - 지지도, 신뢰도, 향상도

◦ 지지도(*Support*) $= \frac{\text{A와 B가 동시에 포함된 거래 수}}{\text{전체 거래 수}} = \frac{\text{A} \cap \text{B}}{\text{전체 거래 수}} = P(A \cap B)$

← 특정 아이템이 데이터에서 발생하는 빈도. 규칙의 **유용성**의 척도

장을 본 목록을 확인했을 때 우유와 식빵이 꼭 함께 있을 확률

◦ 신뢰도(*Confidence*) $= \frac{\text{A와 B가 동시에 포함된 거래 수}}{\text{A가 포함된 거래 수}} = \frac{P(A \cap B)}{P(A)} = \frac{\text{지지도}}{P(A)}$

← 두 아이템 간 연관규칙의 **확실성**의 척도 C(A→B)

우유를 구매했을 때 식빵이 장바구니로 함께 들어갈 확률

출처 : 경영자를 위한 디지털 전략 가이드, 스마투스 비즈니스 리뷰(http://www.sbr.ai)

좋은 규칙을 판단하는 세가지 지표 - 지지도, 신뢰도, 향상도

◦ 향상도(*Lift*) $= \frac{\text{A와 B가 동시에 포함된 거래 수}}{\text{A가 포함된 거래 수 X B가 포함된 거래 수}} = \frac{P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)} = \frac{\text{신뢰도}}{P(B)}$

← A가 주어지지 않은 상태에서 B의 확률에 대하여 A가 주어졌을 때 B의 확률 증가비율

두 아이템의 연관규칙이 우연인지 아닌지를 나타내는 척도 lift(A→B)

식빵을 구매한 횟수 대비 우유와 식빵을 함께 구매했을 횟수

◦ 향상도(lift) 값이 1이면 서로 독립적인 관계이며 1보다 크면 두 품목이 서로 양의 상관관계, 1보다 작으면 두 품목이 서로 음의 상관관계이다. A와 B가 독립이면 분모, 분자가 같기 때문에 1이 나온다.

◦ 신뢰도가 높으면 좋지만, 그게 최선의 연관성 규칙이라는 뜻은 아님

◦ 신뢰도 지지도가 우연히 높게 나올 수 있으므로 향상도를 반드시 확인해야 함.

출처 : 경영자를 위한 디지털 전략 가이드, 스마투스 비즈니스 리뷰(http://www.sbr.ai)

향상도(lift) 값이 따른 관계와 의미

향상도(lift)	의미
1	서로 독립적인 관계이다. A와 B는 연관성이 없다.
<1	A와 B는 음의 상관관계이다.
>1	A와 B는 양의 상관관계이다. 품목 B를 구매할 확률보다 품목 A를 구매한 후에 품목 B를 구매할 확률이 더 높다(A와 B의 연관성이 높다)

지지도, 신뢰도, 향상도


지지도(Support), 신뢰도(Confidence), 향상도(Lift) 예시
Customer ID
Transaction ID
Items
1131
1번
계란, 우유
2094
2번
계란, 기저귀, 맥주, 사과
4122
3번
우유, 기저귀, 맥주, 콜라
4811
4번
계란, 우유, 맥주, 기저귀
8091
5번
계란, 우유, 맥주, 콜라
N = 5 (전체 transaction 개 수)
s(Y) = n(Y)/ N
= n{2번, 3번, 4번}/N = 3/5 = 0.6
연관규칙 {계란, 맥주} → {기저귀} 에 대해
X
Y
▪ 지지도(Support)
s(X→Y) = n(X∪Y)/N
= n{2번, 4번}/ N
= 2/5 = 0.4
▪ 신뢰도(Confidence)
c(X→Y) = n(X∪Y)/ n(X)
= n{2번, 4번}/ n{2번, 4번, 5번}
= 2/3 = 0.667
▪ 향상도(Lift)
Lift(X→Y) = c(X→Y)/ s(Y)
= 0.667 / 0.6 = 1.111
[R 분석과 프로그래밍] http://rfriend.tistory.com

◦ 마트 거래 데이터

transaction	구매물품
1	주스, 탄산음료
2	우유, 주스, 유리창세제
3	주스, 주방세제
4	주스, 주방세제, 탄산음료
5	유리창세제, 탄산음료

◦ 상품의 동시발생 Matrix : 가능한 상품구매조합은 상품의 개수가 N개 인 경우 2^N-1 이고 이 모든 조합에 대한 구매경력을 기록해야 함.

[주스->탄산음료] 의 연관규칙의 평가

◦ 지지도 (support)

주스와 탄산음료를 함께 구매한 횟수/전체구매건수 = 2/5=0.4

◦ 신뢰도 (confidence)

주스와 음료를 함께 구매한 횟수 / 주스를 구매한 횟수=2/4 =0.5

참고) 탄산음료->주스의 신뢰도는 2/3=0.67 (비대칭적이다.)

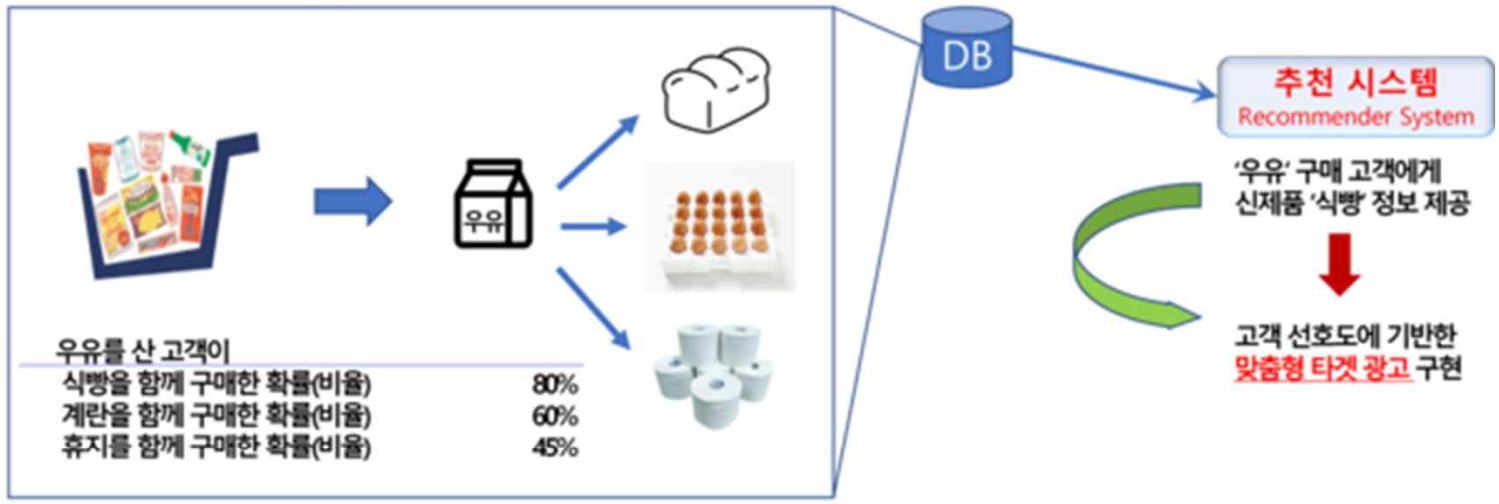
◦ 향상도 (lift)

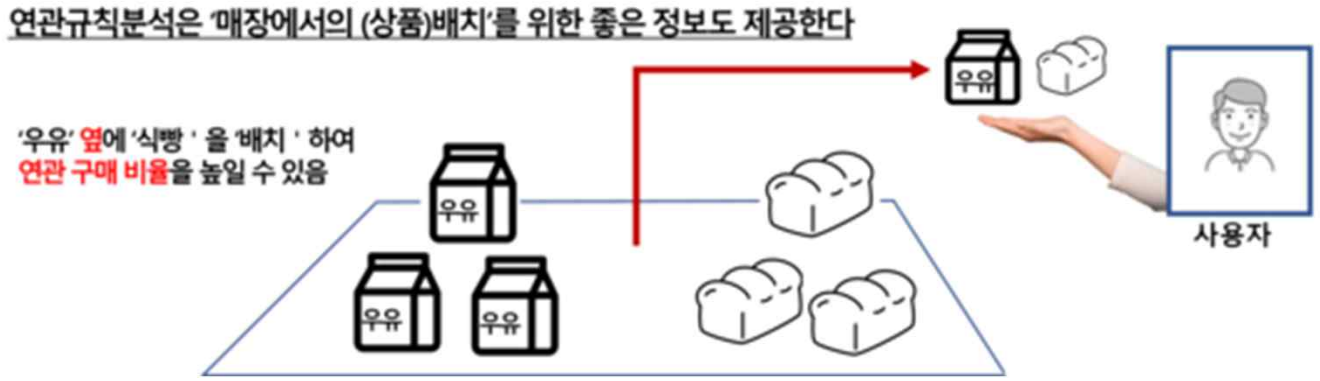
주스를 샀을 때 탄산음료를 같이 산 구매비율/탄산음료 구매비율 =0.5/0.6=0.83

향상도는 적어도 1보다는 커야 양의 연관성이 있다고 봄

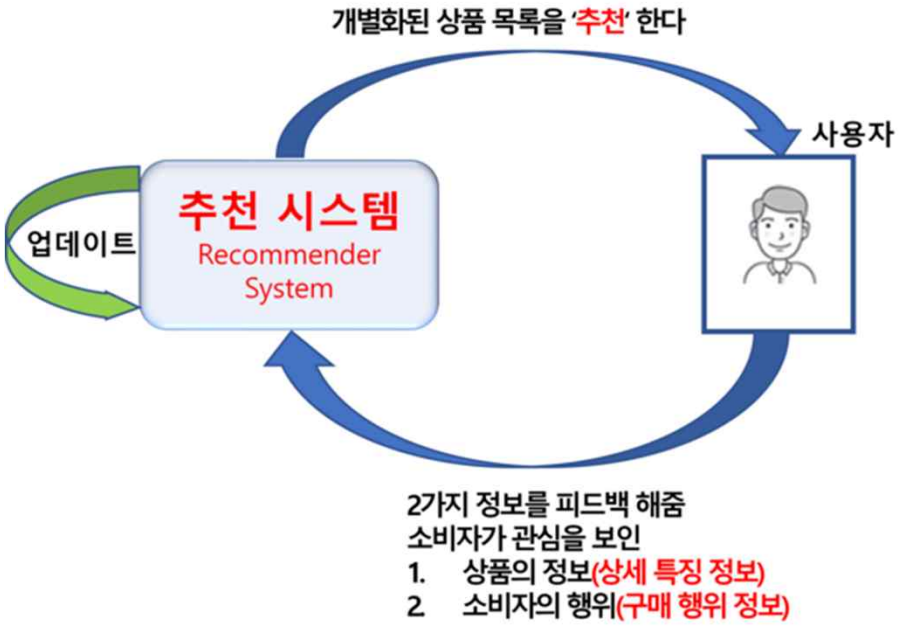
장바구니 분석

고객의 상품 구매 행위에 담긴 '정보'를 추출하여 연관도가 높은 상품의 배치 및 추천에 사용



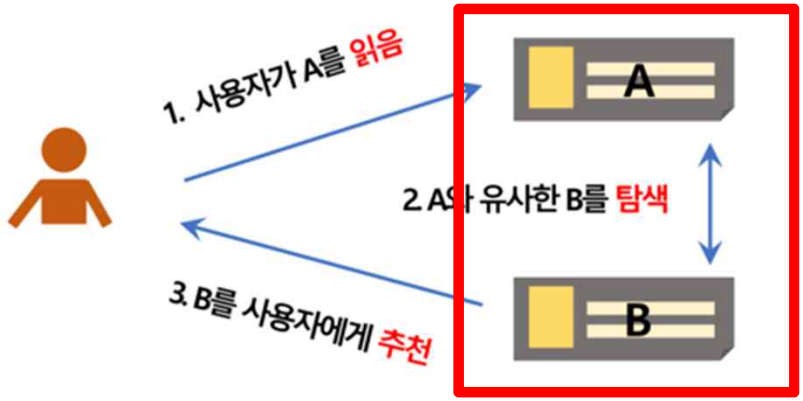


추천 시스템



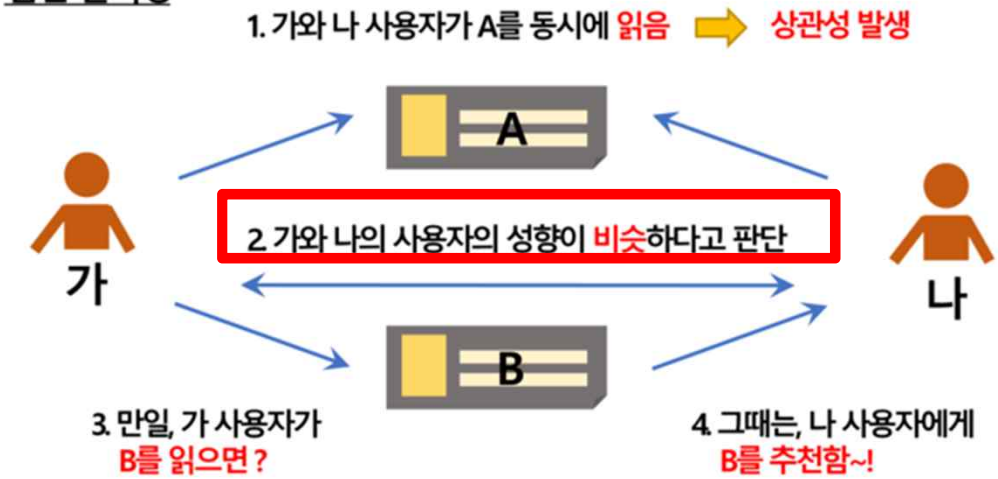
콘텐츠 기반 필터링

컨텐츠 간의 유사도
-사용자가 선호하는 영화와 유사한 영화를 추천



협업 필터링

사용자의 구매패턴의 유사도



A Priori 알고리즘

a priori

미국·영국[ˌeɪ praɪˈɔːraɪ] 영국식

선험적인, 연역적인 (→**a** posteriori)

◦ 후보항목집합을 구성한 후 사전지식(priori knowledge)을 이용하여 빈발 패턴 아이템(또는 항목) 집합을 생성하는 방법

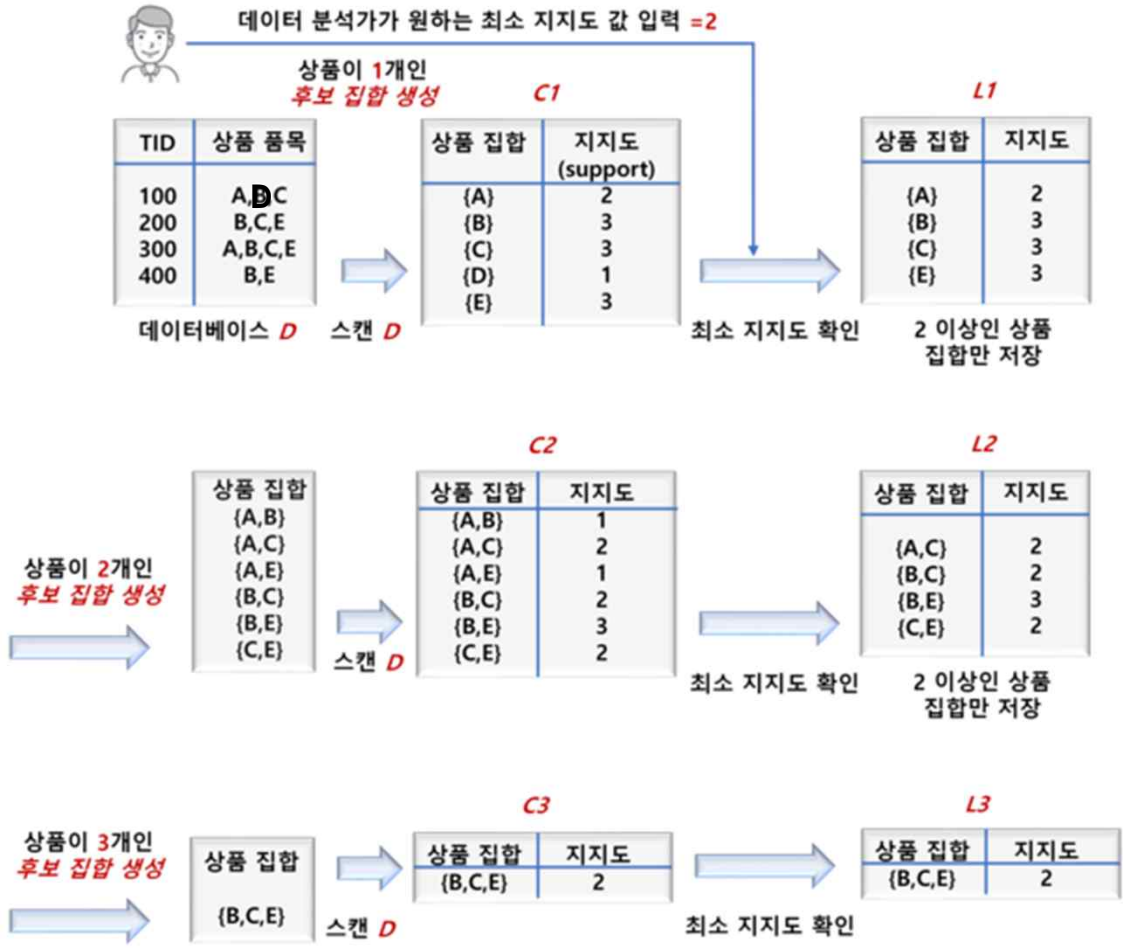
(Item set A) => (Item set B) (if A then B : 만일 A 가 일어나면 B 가 일어난다.)

◦ 빈발 패턴을 찾는 작업은 데이터에서 연관관계, 상관관계, 관심대상 관계를 분석하는 데 있어, 중요한 역할을 한다. 그리고 데이터분류, 군집화 등 다른 데이터 마이닝 작업에 도움이 된다

A Priori 알고리즘

◦ 빈발항목집합을 생성하기 위한 순서

첫째, 1-빈발항목집합을 찾는다. 이 집합을 L1로 나타내면, L1은 2-빈발항목집합인 L2를 찾는데 사용되며 L2는 3-빈발항목집합 L3을 찾는데 이용되는 식으로 계속되어 더 이상의 k-번째 빈발항목집합이 없을 때까지 진행된다.



A Priori 알고리즘

◦ Apriori 알고리즘의 장점

이해하기 쉬운 알고리즘으로써, 알고리즘내의 가지치기 연산은 대규모 데이터베이스의 규모가 큰 항목 집합에서도 쉽게 구현할 수 있다.

◦ Apriori 알고리즘의 단점

항목 집합이 매우 크고 최소 지원이 매우 낮게 유지되는 경우 계산량이 많이 필요하다. 즉 시간이 많이 소모된다.

또한, 실제 많은 콘텐츠 서비스 관련 응용분야에서는 낮은 빈도를 가짐에도 불구하고 빈발항목으로 구성해야 할 경우가 많이 발생한다.

이런 경우 Apriori 알고리즘에서는 최소 지지도를 낮게 설정하여 문제를 해결할 수 있지만, 이럴 경우 후보항목집합들이 많이 늘어나게 되어 탐색 시간에 대한 효율성이 떨어진다.

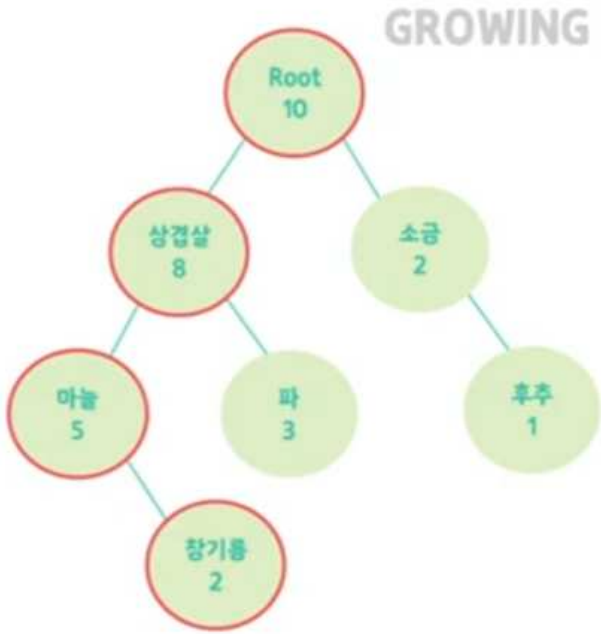
FP-Growth 알고리즘 : 오렌지에서 활용

◦ 빈출 패턴 성장(FP-Growth: Frequent Pattern Growth algorithm) 알고리즘 : 기존의 Apriori 알고리즘의 단점인 데이터 집합에 있는 각각의 아이템 항목들을 조회하면서 빈발항목 집합의 조건에 포함되는지를 계속 판단해줘야 하는 문제를 해결하기 위해 Tree 구조를 활용하여 검색시간을 줄임. FP-Tree(Frequent Pattern Tree)

FP-Growth 알고리즘 순서

FP-Growth 알고리즘

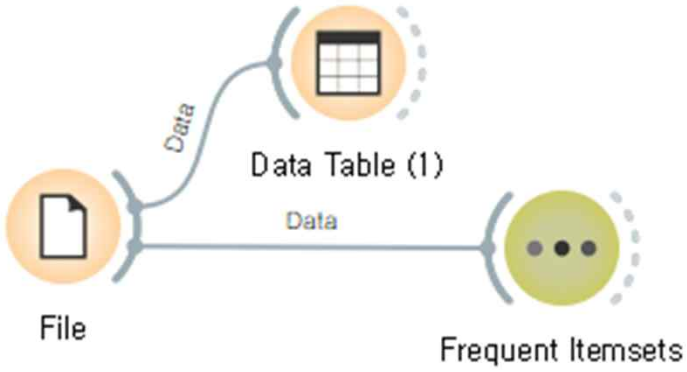
- Frequent Pattern Tree 생성
- 거래 빈도가 높은 순으로 상품 나열
  (빈도가 같으면 알파벳 순)
  Root 노드를 만들고 빈도가 높은 순으로 노드 추가
- 어떤 아이템이 들어왔을 때, 트리를 통해 추천 아이템 전달
- 분할 정복 방식을 통해 Apriori 알고리즘보다 더 빠르게
  빈발항목집합을 추출



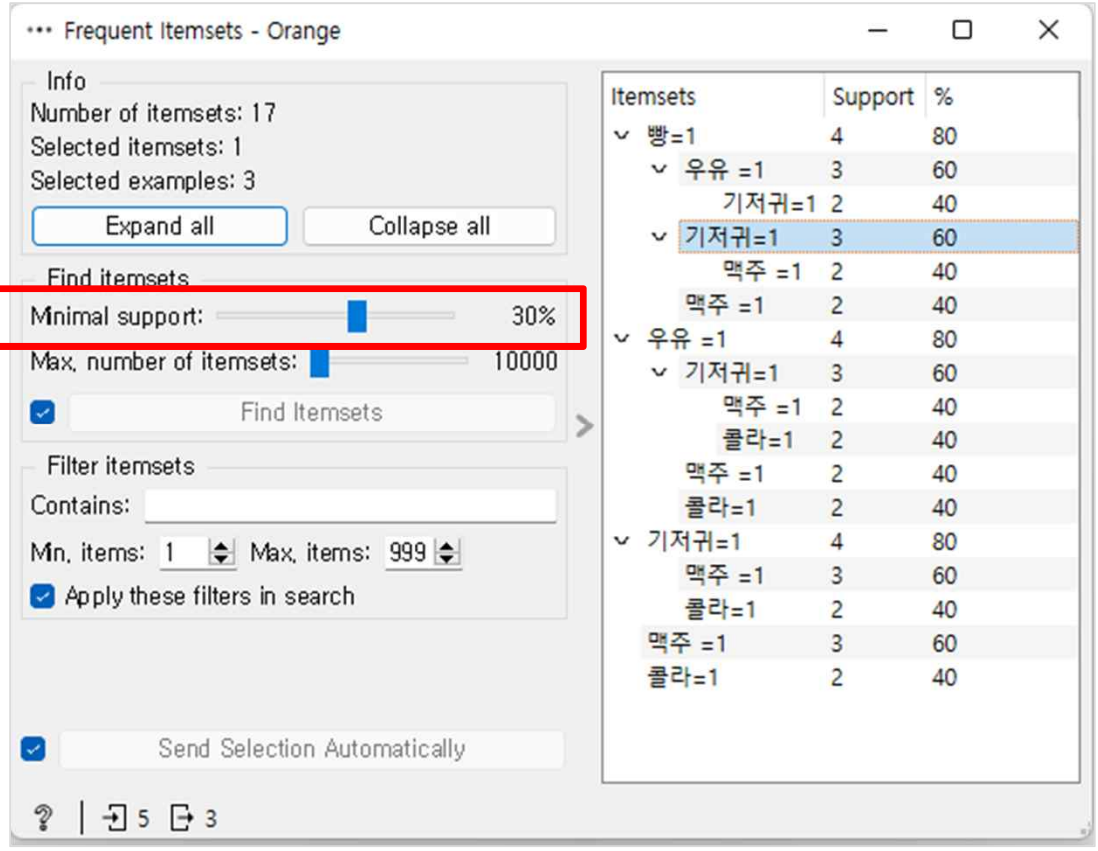
항목	거래수
상겹살, 마늘	3
상겹살, 파	3
상겹살, 마늘, 참기름	2
소금	1
소금, 후추	1

빈발 항목 집합 (Frequent Item Sets)

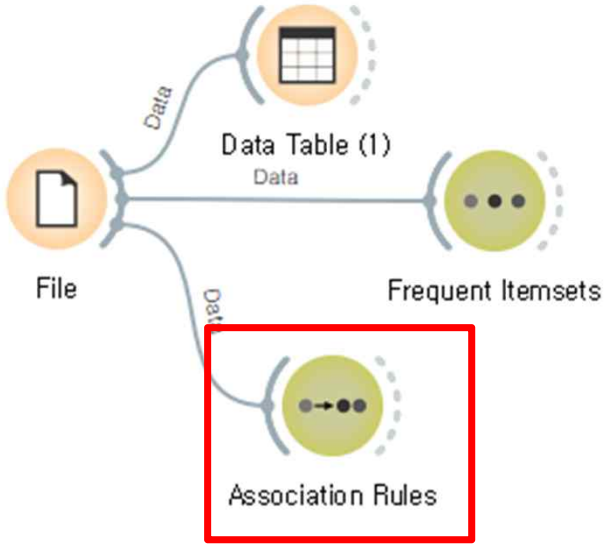
	A	B	C	D	E	F
1	빵	우유	기저귀	맥주	콜라	계란
2	1	1				
3	1		1	1		1
4		1	1	1	1	
5	1	1	1	1		
6	1	1	1		1	



자주 같이 사는 상품들의 목록을 확인



연관규칙 (Association Rules)



A를 샀을 때 B를 살 확률

Association Rules - Orange

Info
Rules: 38 (shown 38)

Find association rules
Min. supp.: 1 %
Min. conf.: 90 %
Max. rules: 10k
Induce only classification rules
Restrict search by below filters
Find Rules

Filter by Antecedent
Contains:
Items, min: 1 max: 999

Filter by Consequent
Contains:
Items, min: 1 max: 999

Send selection

5 - | 38

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.600	1.000	0.600	1.333	1.250	0.120	맥주 =1	→	기저귀=1
0.400	1.000	0.400	2.000	1.250	0.080	빵=1, 맥주 =1	→	기저귀=1
0.400	1.000	0.400	2.000	1.250	0.080	우유 =1, 맥주 =1	→	기저귀=1
0.400	1.000	0.400	2.000	1.250	0.080	콜라=1	→	우유 =1
0.400	1.000	0.400	2.000	1.250	0.080	콜라=1	→	기저귀=1
0.400	1.000	0.400	2.000	1.250	0.080	기저귀=1, 콜라=1	→	우유 =1
0.400	1.000	0.400	2.000	1.250	0.080	우유 =1, 콜라=1	→	기저귀=1
0.400	1.000	0.400	1.500	1.667	0.160	콜라=1	→	우유 =1, 기저귀=1
0.200	1.000	0.200	4.000	1.250	0.040	빵=1, 우유 =1, 맥주 =1	→	기저귀=1
0.200	1.000	0.200	4.000	1.250	0.040	빵=1, 콜라=1	→	우유 =1
0.200	1.000	0.200	4.000	1.250	0.040	빵=1, 콜라=1	→	기저귀=1
0.200	1.000	0.200	4.000	1.250	0.040	빵=1, 기저귀=1, 콜라=1	→	우유 =1
0.200	1.000	0.200	4.000	1.250	0.040	빵=1, 우유 =1, 콜라=1	→	기저귀=1
0.200	1.000	0.200	3.000	1.667	0.080	빵=1, 콜라=1	→	우유 =1, 기저귀=1
0.200	1.000	0.200	4.000	1.250	0.040	맥주 =1, 콜라=1	→	우유 =1
0.200	1.000	0.200	4.000	1.250	0.040	맥주 =1, 콜라=1	→	기저귀=1
0.200	1.000	0.200	4.000	1.250	0.040	기저귀=1, 맥주 =1, 콜라=1	→	우유 =1
0.200	1.000	0.200	4.000	1.250	0.040	우유 =1, 맥주 =1, 콜라=1	→	기저귀=1
0.200	1.000	0.200	3.000	1.667	0.080	맥주 =1, 콜라=1	→	우유 =1, 기저귀=1
0.200	1.000	0.200	4.000	1.250	0.040	계란=1	→	빵=1
0.200	1.000	0.200	4.000	1.250	0.040	계란=1	→	기저귀=1

실제 활용되는 영화추천 알고리즘

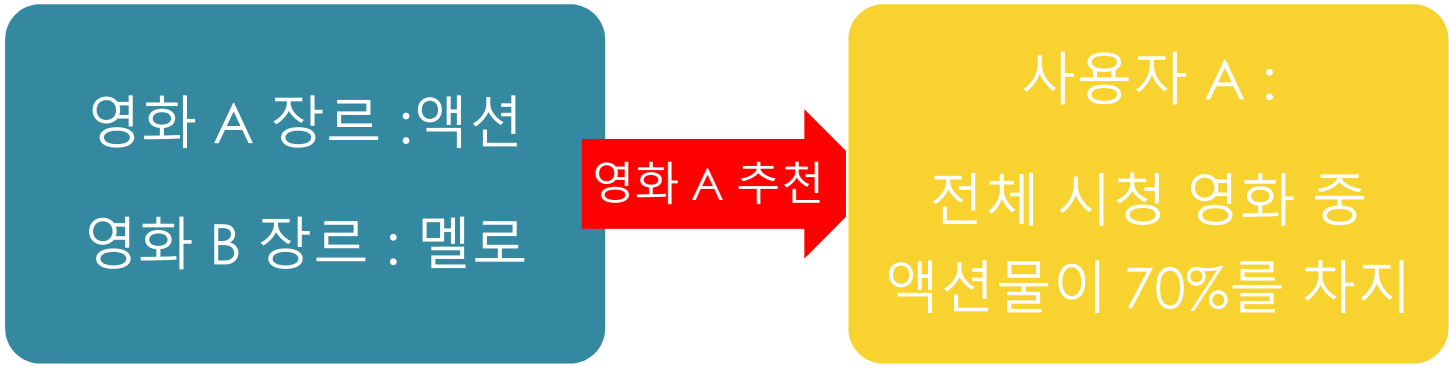
콘텐츠 기반 필터링

◦ 장점 :

- 다른 사용자의 데이터가 필요하지 않다.
- 추천할 수 있는 아이템의 범위가 넓다. 즉 새로운 영화나 인기 없는 영화도 추천이 가능하다.
- 추천하는 이유를 제시할 수 있다. 추천대상의 작품이 과거 사용자가 시청했던 시청목록의 특정 작품과 비슷합니다. 등…

◦ 단점 :

- 새로운 사용자를 위한 추천이 어렵다.
- 선호하는 특성을 가진 항목을 반복 추천한다.



협업 필터링

1. 사용자기반 협업 필터링 (User-based CF) : 나와 비슷한 성향을 가진 사람들이 사용한 아이템을 추천해주는 방식

   예) 사용자 A : 치킨, 피자, 콜라 구매

   사용자 B : 치킨, 콜라 구매

   사용자 C : 떡볶이, 우동, 콜라 구매

* 페이스북, 링크드인 등 대다수의 SNS 친구추천 서비스가 채택하는 방법

협업 필터링

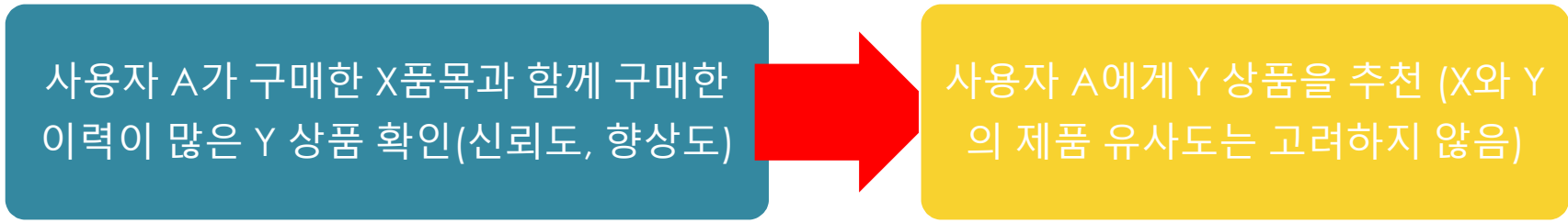
2. 아이템기반 협업 필터링 (Item-based CF) : 내가 구매하려는 물품과 함께 구매된 경우가 많은 아이템을 추천 (연관 규칙 활용)

   예) 공책과 볼펜을 함께 구매하는 소비자가 많다면 공책을 구매한 사용자에게 볼펜을 추천한다 . 이때 두 상품의 특징의 유사도는 고려하거나 파악하지 않는다.

◦ 장점 : 많은 사용자에게서 얻은 기호정보로 새로운 아이템을 추천한다.
직관적으로 이해하기 쉽고 합리적으로 보인다.

◦ 단점 :

- 콜드 스타트 (새로운 아이템이나 사용자가 추가되면 충분한 사용기록이 확보될 때 까지는 적절한 추천이 어렵다.)
- 롱테일 : 인기편향성의 문제라고도 하며, 사용자가 소수의 아이템만 선호하여 대다수의 비인기 아이템들은 추천을 위한 충분한 정보가 쌓이지 못한다.
- 계산효율저하 : 사용자 수가 많은 경우 계산 시간이 오래 걸린다.

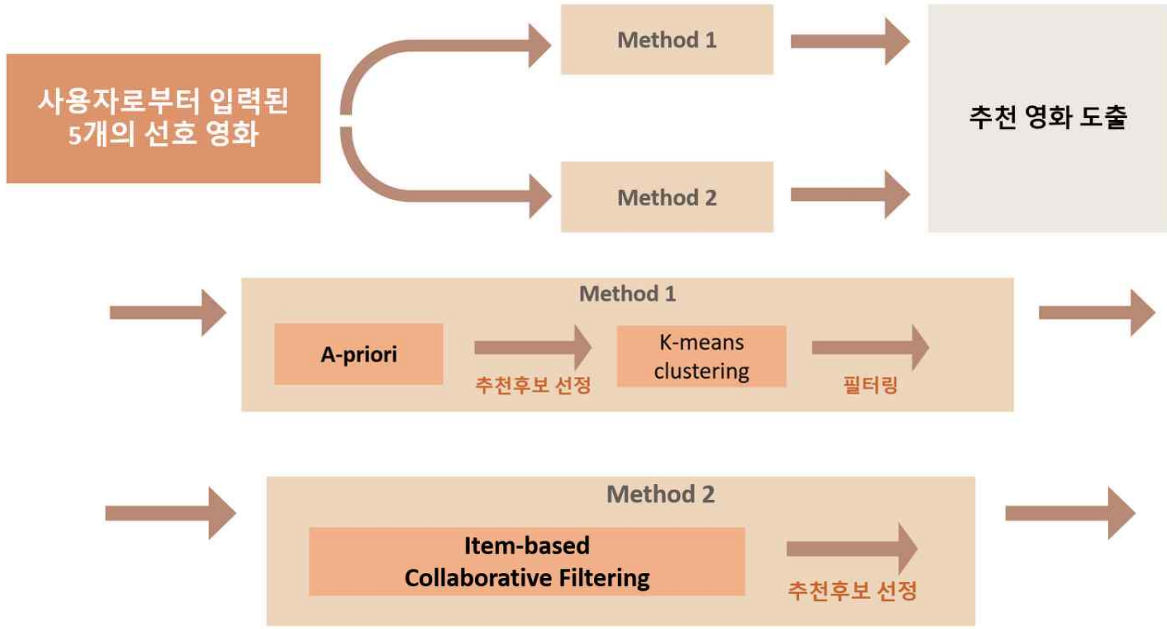


협업 필터링의 문제 보완

- 넷플릭스 : 신규 가입자가 좋아하는 콘텐츠 3개를 고르도록 권장, 선택하지 않은 경우는 인기 콘텐츠 위주로 추천
- 왓차 : 신규 이용자가 첫 이용 시 최소 10개의 콘텐츠에 별점을 부여함으로 유저의 데이터를 파악해 추천

추천시스템의 예

- 방법 1. 사용자가 선택한 영화와 자주 같이 시청된 영화들 중 비슷한 종류의 영화들을 추천 : A priori 알고리즘과 K-means 알고리즘 활용
- 방법 2 : 다른 사용자의 평가를 바탕으로 선호할 만한 영화를 추천 : Item based collaborative filtering



https://ddongwon.tistory.com/115ㅍ

다음 시간에는 시간의 흐름에 따른
데이터의 변동 예측에 관한
시계열분석에 대해 알아보겠습니다.