CSFSM7053-00
Business Analytics

College of Hotel & Tourism
Kyung Hee University

Lecture Note 11
연관규칙분석(Association Rule)

Dohyung Bang

Fall, 2021

Syllabus

Week	Date	Topic	Note
1	9/6(월)	R Basic - R 기초 문법 학습	
2	9/13(월)	R Basic – Data Manipulation I	과제#1
3	9/20(월) (추석)	<추석> (보충영상) R Basic - Data Manipulation II	
4	9/27(월)	Descriptive Analytics I - 데이터 요약하기/상관관계/차이검증	~~과제#2~~
5	10/4(월) (대체공휴일)	<대체공휴일> (보충영상) Descriptive Analytics II - 데이터 시각화	과제#2
6	10/11(월) (대체공휴일)	<대체공휴일> (보충영상) Supplementary Topic I - 외부 데이터 수집 (정적 컨텐츠 수집)	~~과제#4~~ 과제#3
7	10/18(월)	Predictive Analytics I – Linear regression	
8	10/25(월)	Predictive Analytics II – Logistic Regression	시험 대체 수업
9	11/1(월)	Predictive Analytics III – Clustering & Latent Class Analysis	과제#4
10	11/8(월)	Predictive Analytics IV – Tree-based Model and Bagging (Random Forest)	
11	**11/15(월)**	**Predictive Analytics V – Association Rules**	
12	11/22(월)	Prescriptive Analytics I – Linear Programming	과제#5
13	11/29(월)	Prescriptive Analytics II – Data Envelopment Analysis (DEA)	
14	12/6(월)	Prescriptive Analytics III – Integer Programming	과제#6
15	12/13(월)	Prescriptive Analytics IV – Simulation	Quiz
16	**12/20(월)**	**Final Presentation**	

Lecture 11-1

장바구니 분석
(연관규칙분석)

추천시스템이란?

ebay amazon twitter

NETFLIX LinkedIn

"The Power of Collective Intelligence"
추천은 집단 지성의 산물이다.

- Netflix 소비자가 시청하는 컨텐츠의 75%, Amazon은 판매액의 35%가 추천에 기반함
- 기업이 이윤을 추구한 이래로 "추천"은 이윤극대화를 위해 소비자의 소비를 촉구하는 마케팅의 근원적 활동 중 하나로 자리잡아 왔음. 즉, 과거부터 이미 존재해온 개념임
- "**집단지성**"이 축적가능한 형태로 쌓이면서 우리는 기존 이용자들의 행위 패턴에서 규칙을 찾아 유사한 패턴을 보이는 사람에게 "**보다 정교한 추천**"을 제공할 수 있게 됨

Source : https://www.businessinsider.com/netflixs-recommendation-engine-drives-75-of-viewership-2012-4

연관규칙분석의 Motivation

아이언맨 vs **신데렐라** ?

Target group에 어떤 영화를
추천해줘야 할까?

어벤져스, **토르** 본 사람 중
아이언맨 도 본 사람이 **99%**
신데렐라 본 사람은 **0.001%**

What is Association Rule(AR)?

연관규칙(Association rule) 분석의 Idea는 장바구니 분석(Market Basket Analysis)에서부터 출발함

Customer #1

맥주, 과자, 사과, 계란, 빵

Customer #2

계란, 빵, 시리얼, 우유

Customer #3

빵. 우유

내 고객이 무엇을 사는가? 어떤 제품을 함께 구매하는가?

고객에게 어떤 제품을 추천해줘야 하는가?

AR의 목적은 서로 다른 아이템(Item) 간 연관성(Association) 및 상관관계(Correlations)를 찾음으로써 아이템 간 연관 발생 패턴(ex. 연관구매 패턴)을 분석하는 것이 주 목적임

What is Association Rule(AR)?

연관규칙 (Association Rule)

- 특정 아이템 집합을 구매했을 때, 또 다른 아이템 집합을 구매하는 규칙을 의미함
- 마트의 예 : {빵, 계란} => {우유}

  해석 – '빵'과 '계란'을 구매한 고객은 '우유'를 동시에 구매한다.

연관규칙의 문제와 해결방안

- 제품의 수(Item)가 N개이면, 가능한 부분집합의 수는 2^N개다.

  Ex) Items = {계란, 빵, 우유} 이면, 나올 수 있는 구매의 조합은

  → {∅}, {계란}, {빵}, {우유}, {계란, 빵}, {계란,우유}, {빵, 우유}, {계란, 빵, 우유}

- 어떻게 연관 규칙을 찾을 것인가?

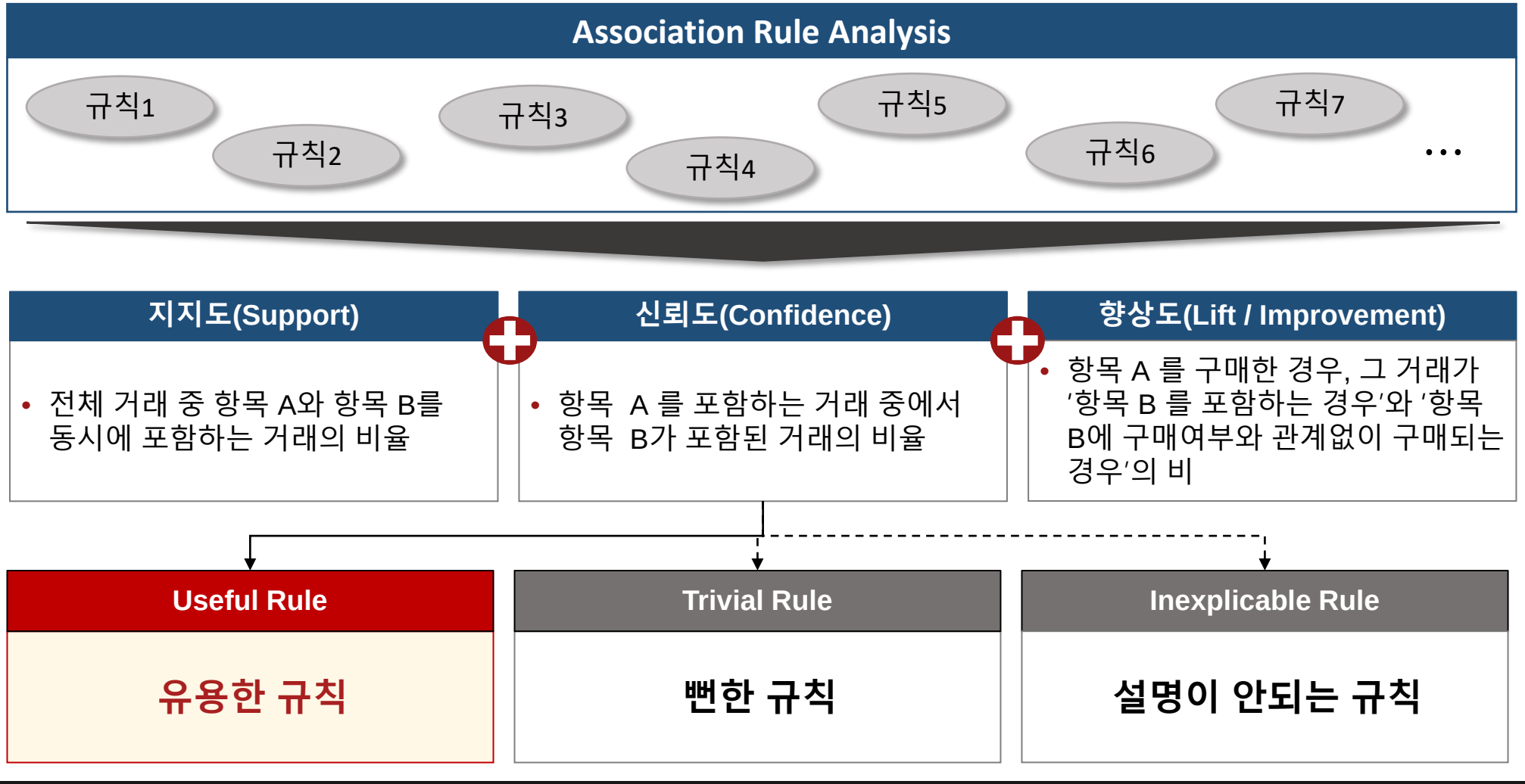
→ **A Priori Algorithm**

- 많은 규칙 중 어떻게 유용한 규칙을 발견할 것인가?
- 많은 규칙 중 원하는 규칙을 어떻게 발견할 것인가?

→ **규칙(Rule)의 평가**

How to find Useful Rule

연관규칙은 수 많은 규칙 중에서 '**유용한 규칙(Useful Rule)**'을 발견하는 것이 목적이며,
유용하지 않은 규칙들 사이에서 유용한 규칙을 발견하는 Insight가 중요함



How to find Useful Rule

'**유용한 규칙(Useful Rule)**'이란 상식적으로 설명이 되면서, 실행가능한 Rule을 의미하며, Trivial한 Rule은 상식적으로 설명은 되나, 이미 실행한 결과로 나타난 값이거나 실행이 불가한 경우, Inexplicable한 Rule은 아예 설명도 안되지만 실행하기도 어려운 경우를 의미함

	Useful Rule	Trivial Rule	Inexplicable Rule
Explainable 상식적으로 설명되는가?	○	○	×
Actionable 실행가능한가?	○	×	×
Example	{금요일, 30대, 남성, 기저귀} -> {맥주}	{책상} -> {의자} : 올해 책상/의자 번들 특판 효과	{변기커버} -> {우산} : 특정 날짜에 갑자기 비가 와서 증가한 매출은 일반화 안됨

규칙을 어떻게 평가할 것인가? #1-지지도(Support)

예시 Rule : $\{x_1, x_2, \dots, x_p\}$ 를 구매한 소비자가 Y도 구매하는가?

소비자#1 : $\{x_1, x_2, \dots, x_p, Y\}$
소비자#2 : $\{x_1, x_2, \dots, x_p, Y\}$
소비자#3 : $\{x_1, x_2, \dots, x_p\}$
소비자#4 : $\{x_1, x_2, \dots, x_p, Y\}$
소비자#5 : $\{x_1, x_2, \dots, x_p, Y\}$
소비자#6 : $\{x_1, x_2, \dots, x_p, Y\}$
소비자#7 : $\{x_1, x_2, \dots, x_p\}$
소비자#8 : $\{x_1, x_2, \dots, x_p, Y\}$
소비자#9 : $\{x_1, x_2, \dots, x_p\}$
소비자#10 : $\{x_1, x_2, \dots, x_p, Y\}$
⋮

- 전체 Item 집합 중에서 $\{x_1, x_2, \dots, x_p\}$와 Y를 동시에 구매할 확률을 Support라고 함
- Confidence와의 차이점은 Confidence는 조건부 확률이라면, Support는 전체 Item 집합 수가 분모가 됨
- 즉, $\{x_1, x_2, \dots, x_p\}$와 Y를 함께 구매한 소비자는 7명, 전체 Item 집합의 수를 30이라 하면, 해당 규칙에 대한 지지도(Support)는 $\frac{7}{30}$

$$supp(\{x_1, x_2, \dots, x_p\} \Rightarrow Y) = \frac{\{x_1, x_2, \dots, x_p\}\text{와 } Y \text{ 동시 구매 건수}}{\text{전체 구매 } Item \text{ 집합}} = \Pr(\{x_1, x_2, \dots, x_p\} \cap Y)$$

규칙을 어떻게 평가할 것인가? #2-신뢰성(Confidence)

예시 Rule : $\{x_1, x_2, \ldots, x_p\}$ 를 구매한 소비자가 Y도 구매하는가?

소비자#1 : $\{x_1, x_2, \ldots, x_p, Y\}$
소비자#2 : $\{x_1, x_2, \ldots, x_p, Y\}$
소비자#3 : $\{x_1, x_2, \ldots, x_p\}$
소비자#4 : $\{x_1, x_2, \ldots, x_p, Y\}$
소비자#5 : $\{x_1, x_2, \ldots, x_p, Y\}$
소비자#6 : $\{x_1, x_2, \ldots, x_p, Y\}$
소비자#7 : $\{x_1, x_2, \ldots, x_p\}$
소비자#8 : $\{x_1, x_2, \ldots, x_p, Y\}$
소비자#9 : $\{x_1, x_2, \ldots, x_p\}$
소비자#10 : $\{x_1, x_2, \ldots, x_p, Y\}$
⋮

- 전체 Item 집합 중에서 $\{x_1, x_2, \ldots, x_p\}$를 구매한 소비자가 Y를 구매할 연관성에 대해, $\{x_1, x_2, \ldots, x_p\}$를 구매한 소비자가 Y도 구매했을 확률을 Confidence라고 함
- 즉, $\{x_1, x_2, \ldots, x_p\}$ 구매한 소비자는 10명, Y를 함께 구매한 소비자는 7명으로 해당 규칙에 대한 신뢰도(Confidnece)는 $\frac{7}{10}$

$$conf(\{x_1, x_2, \ldots, x_p\} \Rightarrow Y) = \frac{\{x_1, x_2, \ldots, x_p\}\text{와 } Y \text{ 동시 구매 건수}}{\{x_1, x_2, \ldots, x_p\}\text{ 구매 빈도}} = \Pr(Y \mid \{x_1, x_2, \ldots, x_p\})$$

규칙을 어떻게 평가할 것인가? #3-향상도(Lift)

예시 Rule : $\{x_1, x_2, \dots, x_p\}$ 를 구매한 소비자가 Y도 구매하는가?

소비자#1 : $\{x_1, x_2, \dots, x_p, Y\}$
소비자#2 : $\{x_1, x_2, \dots, x_p, Y\}$
소비자#3 : $\{x_1, x_2, \dots, x_p\}$
소비자#4 : $\{x_1, x_2, \dots, x_p, Y\}$
소비자#5 : $\{x_1, x_2, \dots, x_p, Y\}$
소비자#6 : $\{x_1, x_2, \dots, x_p, Y\}$
소비자#7 : $\{x_1, x_2, \dots, x_p\}$
소비자#8 : $\{x_1, x_2, \dots, x_p, Y\}$
소비자#9 : $\{x_1, x_2, \dots, x_p\}$
소비자#10 : $\{x_1, x_2, \dots, x_p, Y\}$
⋮

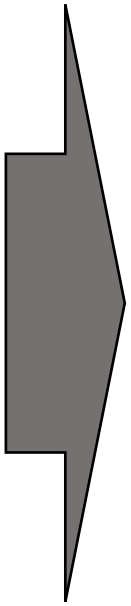
- 품목집합 $\{x_1, x_2, \dots, x_p\}$를 구매하지 않았을 때 품목 Y를 구매할 확률 대비 $\{x_1, x_2, \dots, x_p\}$를 구매했을 때, Y를 구매할 확률의 증가비율
- 이 값이 클수록 품목집합 $\{x_1, x_2, \dots, x_p\}$의 구매여부가 품목 Y를 구매하는 데 큰 영향을 미친다고 해석할 수 있음
- 만약 향상도가 1보다 크면 이 규칙은 결과를 예측함에 있어 우연적 기회(Random chance)보다 좋다고 할 수 있음. 품목 $\{x_1, x_2, \dots, x_p\}$와 Y 간 구매 연관성이 없다면 향상도는 항상 1이 됨

$$lift(\{x_1, x_2, \dots, x_p\} \Rightarrow Y) = \left(\frac{\{x_1, x_2, \dots, x_p\}\text{와 } Y \text{ 동시 구매 건수}}{\{x_1, x_2, \dots, x_p\} \text{ 구매 빈도}}\right) / \left(\frac{Y \text{ 포함 건수}}{\text{전체 구매 건}}\right) = \textbf{신뢰도/P(Y)}$$

예시#1 : 장바구니 영수증

장바구니 구매 내역을 바탕으로 신뢰도 및 지지도를 계산해보자.

장바구니 Item 집합	주문 수
라면	100
우유	150
계란	200
{라면, 우유}	400
{라면, 계란}	300
{우유, 계란}	200
{라면, 우유, 계란}	100
구매안함	550
전체 거래 수	2,000



장바구니 Item 집합	해당 Item 집합 포함된 주문 수	확률
라면	900	0.450
우유	850	0.425
계란	800	0.400
{라면, 우유}	500	0.250
{라면, 계란}	400	0.200
{우유, 계란}	300	0.150
{라면, 우유, 계란}	100	0.050

예시#1 : 장바구니 영수증

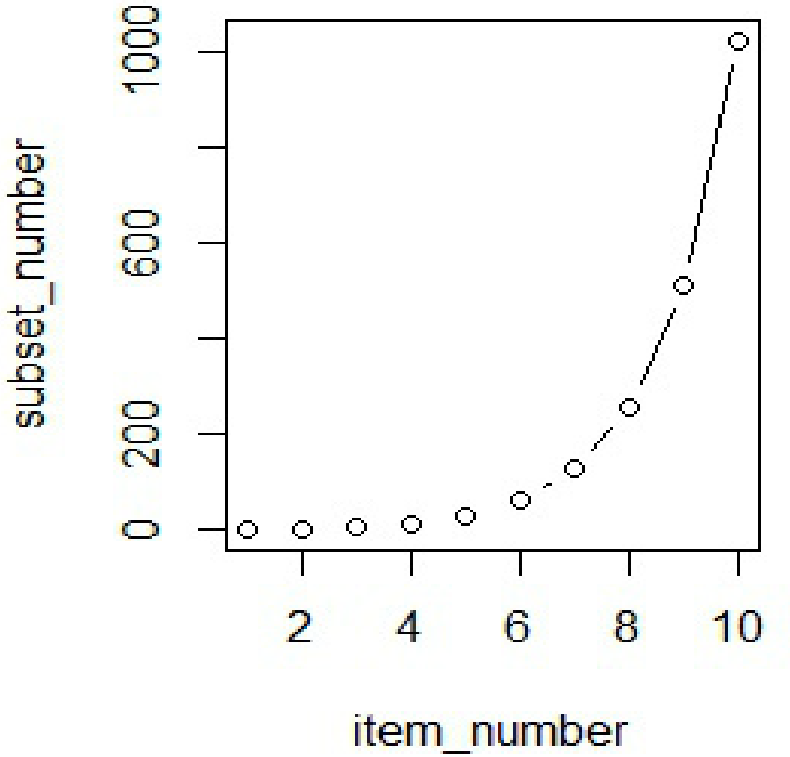
규칙 X => Y	지지도(Support) P(X∩Y)	P(X)	신뢰도(Confidence) P(Y|X)=P(Y∩X)/P(X)	향상도(Lift) P(Y|X)/P(Y)
라면 => 우유	0.250	0.450	0.556 = 500/900	1.308
우유 => 라면	0.250	0.425	0.588 = 500/850	1.307
계란 => 우유	0.150	0.400	0.375	0.882
우유 => 계란	0.150	0.425	0.353	0.883
라면 => 계란	0.200	0.450	0.444	1.110
계란 => 라면	0.200	0.400	0.500	1.111
{라면, 우유} => 계란	0.050	0.250	0.200 = 100/500	0.500
{우유, 계란} => 라면	0.050	0.150	0.333	0.740
{라면, 계란} => 우유	0.050	[illegible]	0.250	[illegible]588

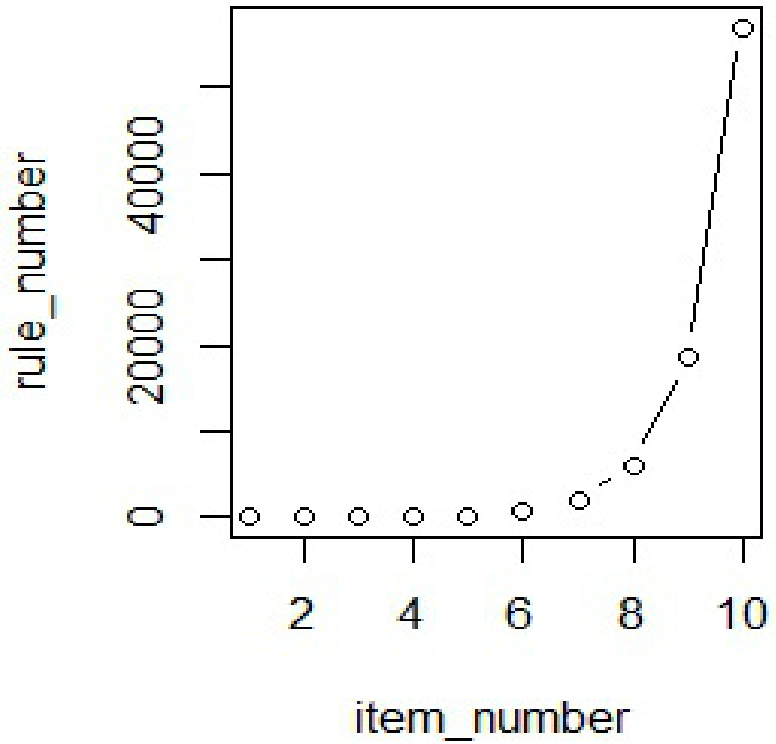
라면, 우유, 계란 모두 포함된 집합 중 가장 신뢰도(Confidnece)가 높은 집합은 '우유, 계란' 구매한 고객이 라면도 구매할 규칙임
하지만, 해당 집합의 지지도는 0.05로 발생확률이 매우 낮아 해당 규칙에 의미를 부여하기가 제한적임

이러한 경우를 위해 향상도(Lift) 지표를 활용함.
하지만, 예시에서는 향상도(Lift)도 낮아 해당 규칙이 유용한 규칙은 아님을 알 수 있음

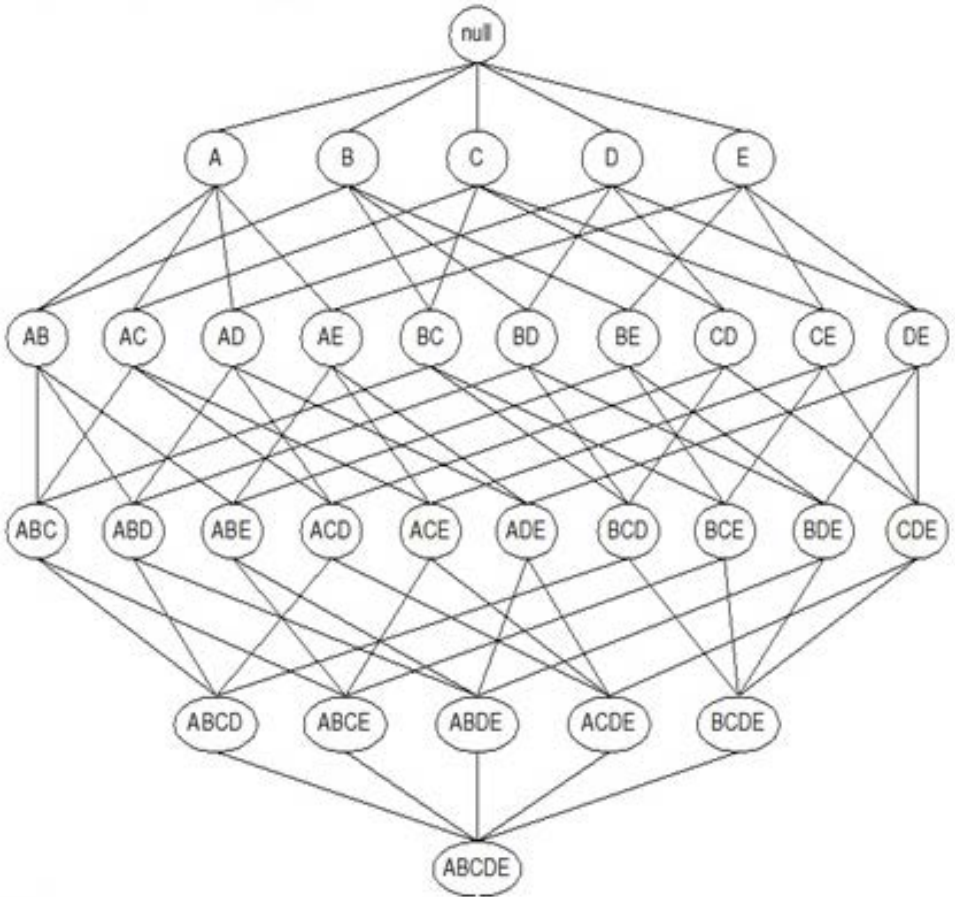
어떻게 연관규칙을 찾을 것인가? – “A Priori Algorithm”

연관규칙을 찾을 때, 장바구니에 들어갈 수 있는 품목의 수가 증가하면, 고려해야할 규칙(Rule)의 가지 수가 기하급수적으로 증가함





어떻게 연관규칙을 찾을 것인가? – “A Priori Algorithm”



아무것도 사지 않는 경우

5개 중 1개만 사는 경우

5개 중 2개만 사는 경우

5개 중 3개만 사는 경우

5개 중 4개만 사는 경우

5개 중 5개 모두 사는 경우

<u>31가지 경우의 수</u>가 나옴

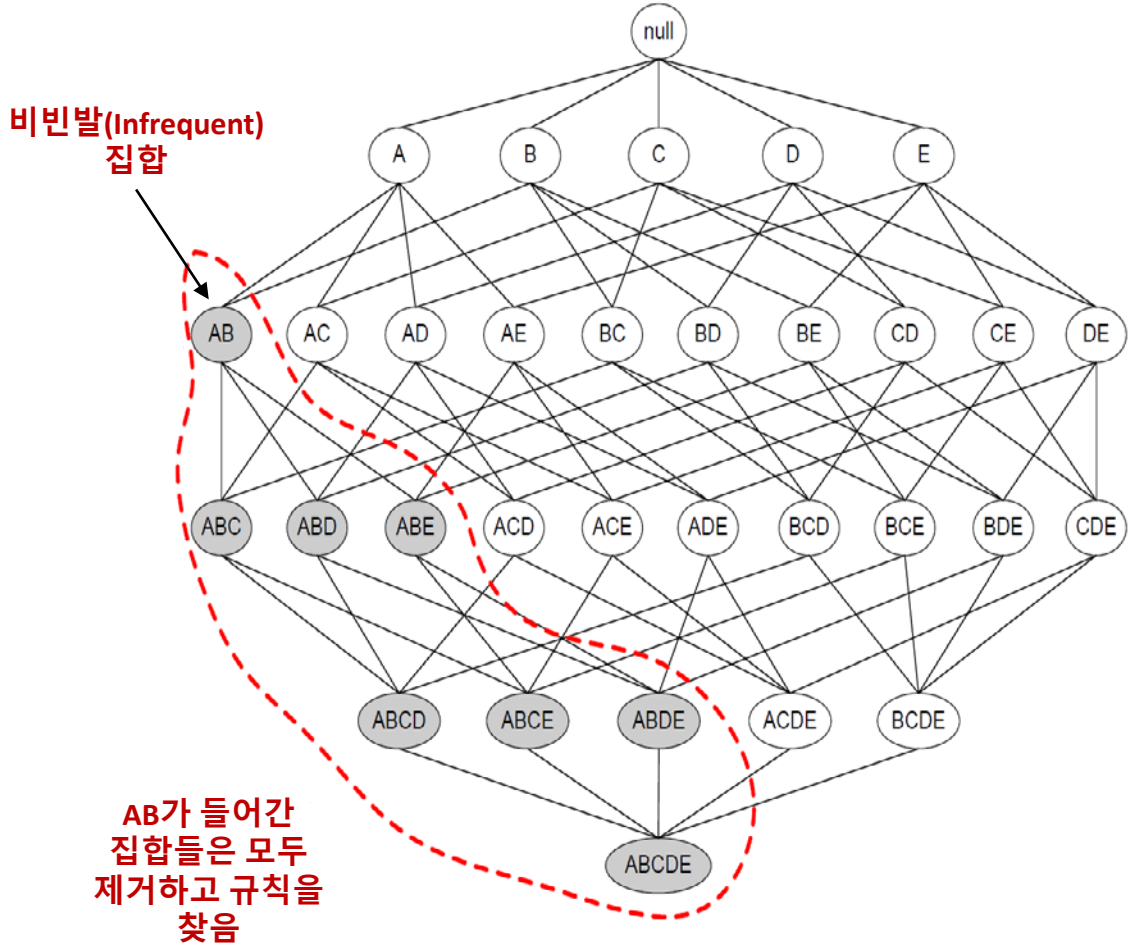
※ 참고
k개 아이템이 존재하는
장바구니의 부분집합의 수 = 2^k

Source : https://rfriend.tistory.com/192?category=706118

어떻게 연관규칙을 찾을 것인가? – “A Priori Algorithm”

A Priori Algorithm은 빈발집합(Frequent sets)만을 고려함으로써 계산 비용을 줄이기 위해, 비빈발집합(Infrequent sets)은 적절하게 제거(가지치기; Prunning)하고 규칙을 찾음

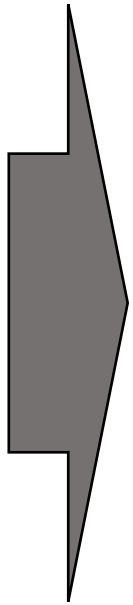
- 가령, 아이템 집합 {A,B}의 지지도(support)가 0.1 이라고 가정함
- {A,B}가 부분집합으로 들어가는 {A,B,C}, {A,B,D},…, 등의 지지도는 아무리 높아도 0.1를 넘지 못함
- 따라서, 임의의 아이템 집합의 지지도가 일정 수준을 넘지 못하면 해당 아이템의 부분집합 지지도도 기준보다 명백히 작아질 것임
- 이런 방식으로 **일정 기준을 넘지 못하는 규칙을 제거**하고 “유용한 규칙”을 찾아가는 방식을 “A Priori Algorithm”이라 함



Transaction Data

Transaction Data는 일반적으로 희소 행렬(Sparse Matrix)가 되므로 R에서 정의하는 “Transaction” 데이터 타입으로 변환해서 사용하도록 함

ID	Items
1	달걀, 라면, 참치캔
2	라면, 햇반
3	라면, 콜라
4	달걀, 라면, 햇반
5	달걀, 콜라
6	라면, 콜라
7	라면, 햇반
8	달걀, 라면, 콜라, 참치캔
9	달걀, 라면, 콜라
10	양파



ID	달걀	라면	참치캔	햇반	콜라	양파
1	1	1	1	0	0	0
2	0	1	0	0	1	0
3	0	1	0	0	0	1
4	1	1	0	0	1	0
5	1	0	0	0	0	1
6	0	1	0	0	0	1
7	0	1	0	0	1	0
8	1	1	1	0	0	1
9	1	1	0	0	0	1
10	0	0	0	0	0	0

비거래 데이터의 Transaction Data화

컨텐츠 필터링은 거래형 데이터가 아닌 데이터를 희소행렬로 만들어 거래데이터로 변환함으로써 이를 통해 연관규칙을 만들어낼 수 있음

범주형 데이터(Categorical Data) → 이항변수화(Binarization)

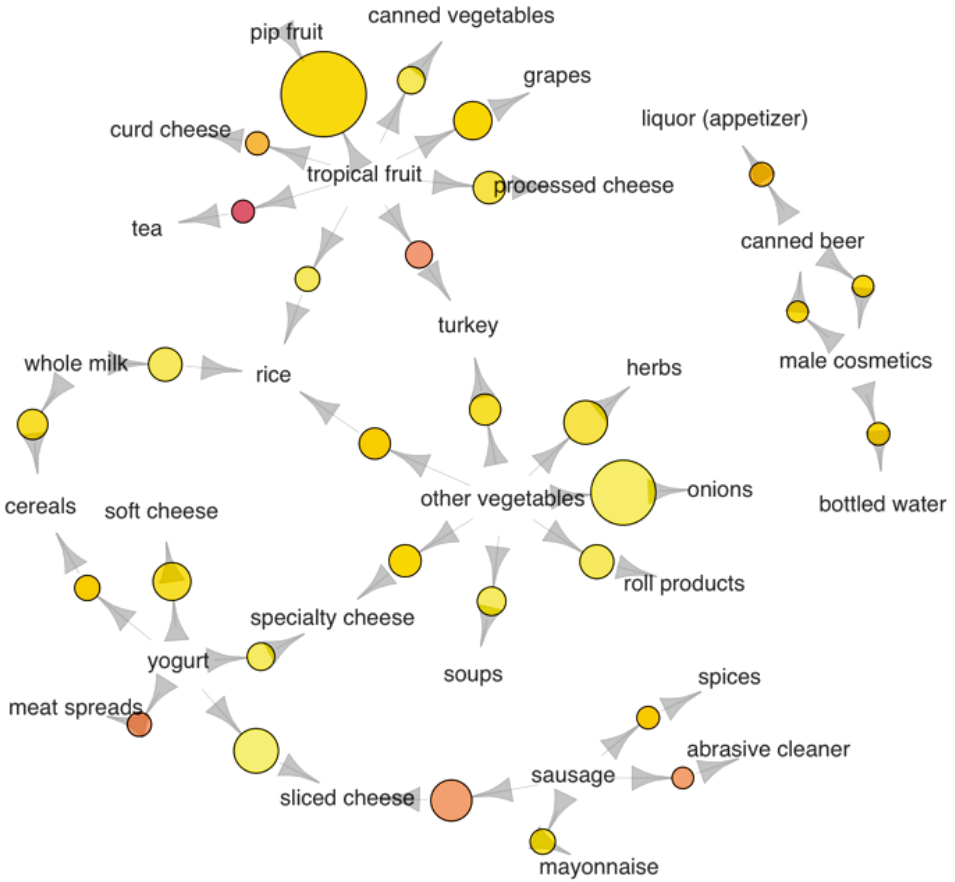
CUST_ID	GENDER	AGE	CHILD_PRD_YN	MOBILE_APP_USE	RE_ORDER
1	FEMALE	23	NO	YES	YES
2	MALE	28	NO	YES	NO
3	FEMALE	42	NO	NO	NO
4	FEMALE	34	YES	YES	YES
5	MALE	45	NO	NO	NO
6	FEMALE	36	YES	YES	YES

CUST_ID	GENDER=MALE	GENDER=FEMALE	AGE=20	AGE=30	AGE=40	CHILD_PRD_YN=YES	CHILD_PRD_YN=NO	MOBILE_APP_USE=YES	MOBILE_APP_USE=NO	RE_ORDER=YES	RE_ORDER=NO
1	0	1	1	0	0	0	1	1	0	1	0
2	1	0	1	0	0	0	1	1	0	0	1
3	0	1	0	0	1	0	1	0	1	0	1
4	0	1	0	1	0	1	0	1	0	1	0
5	1	0	0	0	1	0	1	0	1	0	1
6	0	1	0	1	0	1	0	1	0	1	0

Source : https://rfriend.tistory.com/194?category=706118

연관규칙의 시각화(Visualization)

연관규칙을 찾은 다음 아래와 같이 규칙들 간 관계를 네트워크 맵으로 시각화하여 표현할 수 있음



- 원의 크기는 규칙의 **<u>지지도(Support)</u>**를 나타냄
- 가령, {Tropical Fruit}을 구매한 고객이 {Pip Fruit}을 구매할 규칙에 대한 **지지도(Support)**가 가장 높고, {Tropical Fruit}은 다른 품목과의 연관구매도 많이 발생하는 것을 알 수 있음
- 또한, {Spices}를 구매하는 고객은 {Sausage}를 구매했을 경우에만 {Spices}를 구매하는 것을 알 수 있음

Source : https://algobeans.com/2016/04/01/association-rules-and-the-apriori-algorithm/

연관규칙분석의 장점 및 한계점

장점 및 현업으로의 적용방안

- 거래량을 몰라도 단순히 거래의 여부만 가지고도 연관거래패턴을 파악할 수 있음
- 비교적 연산이 간단하고, 쉽게 적용할 수 있음
- 간단한 분석에 비해 많은 Insight을 줄 수 있음
  - 제품 진열(연결 진열)
  - 세트상품 구성 및 연결판매
  - 누수시장에 대한 제품 개발
  - 연관제품 추천

한계점

- 거래량을 알더라도 거래의 여부만 고려되고, 거래량이 거래패턴 분석에 고려되지 않음
- 유용한 규칙이 생각보다 많지 않음
- 거래 Data는 공개되지 않는 경우가 많음

응용 : 연관규칙과 클러스터링의 결합

기대효과 (Expected Effect)

1	"누수시장 발견으로 시장기회 창출"	• 세분시장 별 구매 패턴, 동일 세분시장 내 구매 패턴의 Profiling을 통해 누수시장을 발견하고, 누수 시장을 커버해 새로운 시장 기회를 발견할 수 있는지를 판단할 수 있음
2	"Bundling 상품개발"	• 연관규칙이 발견된 품목 Category를 바탕으로 Bundling 상품을 개발할 수 있음
3	"Bundling Pricing"	• 연관규칙이 발견된 품목 Category에 대해 현재 제공하는 각 제품에 대한 Price 대비 저렴한 Bundle Price를 결정해 수익성 제고 및 고객 유지를 도모할 수 있음
4	"수요예측"	• 시차연관분석(Sequential association rule)을 통해 현재의 구매행위 다음 이뤄질 구매에 대한 수요를 예측할 수 있음
5	"제품 추천"	• 동일 고객 群의 특정 고객이 구매하지 않는 품목에 대해 해당 고객 群의 품목 Category에 맞게 Bundle price와 함께 추천할 수 있음