9주차 강의

연관성 분석

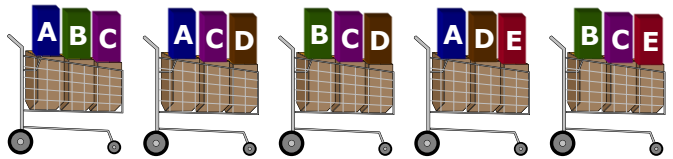
(Association Analysis)

2022.05.03

Institutor: JS LEE

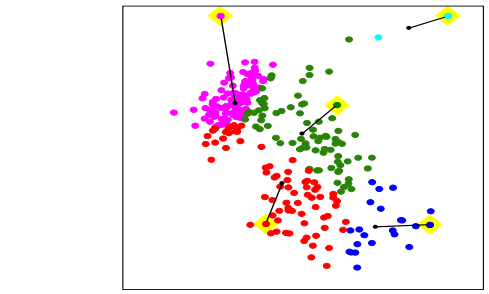
비지도 학습

	비지도학습 (자율학습) (unsupervised learning)
의미	•데이터 포인트들 간의 관계를 기반으로 데이터에서 패턴을 찾아내는 작업
특징	•타겟(출력, 결과)변수 존재하지 않음
분석 기법	•군집분석, 연관성 분석, …



Rule	Support	Confidence
A ⇒ D	2/5	2/3
C ⇒ A	2/5	2/4
A ⇒ C	2/5	2/3
B & C ⇒ D	1/5	1/3

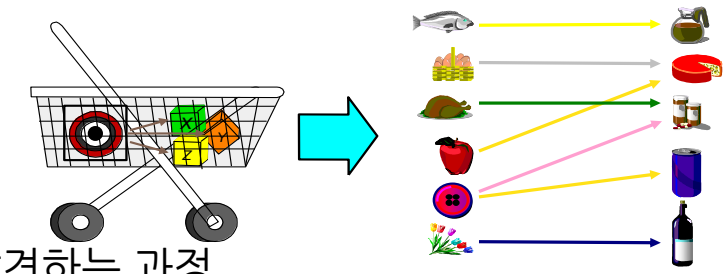
연관성 분석 (Association Analysis)



군집분석 (Clustering)

연관성 분석 - 개요

연관성 분석이란?



- 트랜잭션 집합이 주어지면, 트랜잭션에서 항목들의 발생을 기반으로 연관규칙을 발견하는 과정
- 연관규칙
  - 일련의 거래나 사건들의 연관성에 대한 규칙
  - “if-then” 형식으로 표현

{Item A} → {Item B}

Item A: if절, 조건절, 선행(antecedent), 전제(premise)
Item B: then절, 주절, 후행(consequent), 결론(conclusion)

  - 인과관계를 의미하지 않음
  - 상품 혹은 서비스간의 관계를 살펴보고 이로부터 **유용한 규칙**을 찾아내고자 할 때 이용되는 기법
  - 장바구니 분석 또는 친화성 분석이라고도 함
  - 각각의 거래(Transaction)에 대한 상품들의 항목을 분석

    * 일반적으로 하나의 거래는 한 고객에 의한 구매를 의미하며, 상품은 그 구매를 통해 구입된 물건 및 서비스를 의미함

연관성 분석 - 개요

Figure: (Han & Kamber, 2001)

Transaction	Item 1	Item 2	Item 3	Item 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

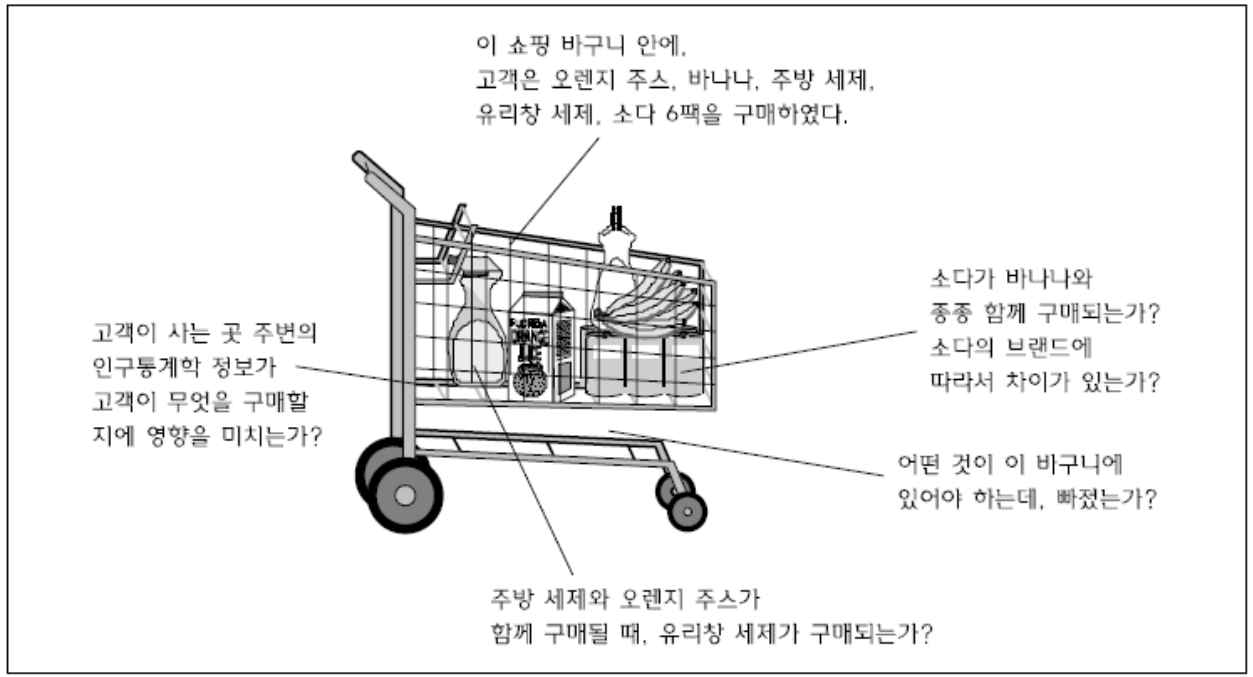
ⓒ 고려대 DSBA 유튜브

- Customer 1, 2, 3,….,n (각각의 장바구니) => 거래 (transation)
- 각 장바구니에 포함된 상품 => 상품/ 또는 항목 (item) ,
- 항목집합Itemset : ▪ 하나 이상의 항목 모음
  - 예: {Milk, Bread, Diaper}
  - k-항목집합k-itemset: k개 항목들이 포함된 항목 집합

연관성 분석 - 개요

- 바코드 스캐너, 재고관리 시스템, 카드 정보 등으로 엄청난 거래 데이터가 저장이 되고 이러한 분석하여 구매패턴을 파악함
- 이러한 것을 장바구니 분석이라 함.
- 이 기법은 쇼핑 데이터로 창안 되었으나, 다른 경우에도 유용하게 사용됨
- 일반적으로 아래와 같은 작업들이 포함됨
  - 거대한 데이터베이스에서 간단한 성능 측정치를 이용하여 연관성 찾기
  - 거래 데이터의 특이점 찾기
  - 유용하고 실행가능한 (actionable) 패턴 식별방법 알아 두기
- 실행 가능한 패턴을 찾는 것이 목표임
- 연관규칙은 분석기술과 비즈니스 이해 필요
  - 특정한 목표 없이 데이터의 패턴을 나타냄 (무방향성 데이터마이닝)
  - 이 패턴들이 잘 맞는지에 대한 판단은 사람의 해석에 달려 있음

연관성 분석 - 개요



〈그림 9-1〉 장바구니 분석은 어떤 물품들이 함께 구매되는지와 고객에 대하여 이해하는 것을 도와준다.

- 이러한 정보는 쉽게 행동에 옮길 수 있음(actionable)
  - 새로운 상점의 상품 배치, 패키지 상품 개발, 교차판매전략 등
  - 특정한 상품을 구입한 고객이 어떤 부류 (특성)에 속하는지=> 상품기획 도움

연관성 분석 - 개요

연관규칙의 세가지 범주

- 목요일, 식료품 가게를 찾는 고객은 아기 기저귀와 맥주를 함께 구입하는 경향이 있다.
  - 일반적으로 알아낼 수 없는 규칙 → 실제 적용으로 좋은 결과 기대 (실행 가능한 : actionable)
- 한 회사의 전자제품을 구매하던 고객은 전자제품을 살 때 같은 회사의 제품을 사는 경향이 있다.
  - 상식적으로 널리 알려진 관련성 → 이의 발견은 큰 의미가 없다. (사소한 : trivial)
  : 너무 명확하여 유용하지 않은 규칙
- 새로 연 건축 자재점에서는 변기 덮개가 많이 팔린다.
  - 타당한 근거가 없는 연관성 → 설득력 부족으로 일반화에 무리 (설명하기 어려움 : inexplicable)
  : 단순한 데이터에 있는 랜덤 패턴 일 수 있음

행동 가능한 규칙(Actionable Rules) : 높은 질의 행동 가능한 정보를 포함
사소한 규칙(Trivial Rules) : 해당 분야에 익숙한 사람이라면 누구나 이미 알고 있는 규칙
설명 불가능한 규칙(Inexplicable Rules) : 설명이 되지 않고, 실제 행동을 취할 수도 없는 규칙

연관성 분석 - 적용

- 수퍼마켓, 대형 할인마트에서 소비자의 구매패턴 분석

  - 소비자가 구매할 가능성이 높은 상품들을 같이 배열

  - 특별히 함께 많이 구매되는 상품의 재고를 효과적으로 관리

  - 판촉상품이나 경품의 선정 등에도 유용하게 적용

- 백화점, 호텔에서 고객들이 다음에 원하는 서비스를 미리 파악
- 신용카드, 대출 등의 은행서비스 내역으로부터 특정한 서비스를 받을 가능성이 높은 고객 탐지
- 의료보험금이나 상해보험금 청구가 특이할 때 보험사기 징조, 추가적인 조사
- 환자의 의무기록에서 여러 치료가 같이 이루어진 경우 합병증 발생의 징후 탐지
- 온라인 추천시스템

  - 구매할 가능성이 있는 항목을 검토하는 고객에게 그 항목과 자주 같이 구매되는 다른 항목을 추천

- 언론 웹사이트에 접속한 독자들이 방문하는 페이지들의 연관성을 분석

연관성 분석 - 개요

- 암 데이터에서 흥미로우며 빈번히 발생하는 DNA패턴과 단백질 서열 검색
- 고객이 휴대폰 서비스 중단 또는 결합상품 업그레이드를 할 때 선행되는 행동
- 수강과목 추천
- 올비브영 : “ 음료 매출 향상시켜라”
  : 유독 매출이 저조한 음료.. MD는 소비자 매출데이터 분석해 봄

  여성이 아침시간 음료 소비가 많으며, 음료를 살 때, 헬스 쵸코바와 같은 스낵류 함께 구매

  MD는 음료의 구색을 늘리는 대신, 간단하게 아침대용으로 먹을 만한 스낵류를 늘였음

.


중앙일보 중앙SUNDAY
뉴스 머니 오피니언 S MAGAZINE

당뇨·심장병·뇌졸중 부르는 잇몸병

[중앙선데이] 입력 2016.07.10 00:39 487호 24면 지면보기

치주질환, 전신질환 위험 얼마나 높이나	
당뇨병 6배	폐렴 4.2배
뇌졸중 2.8배	심장질환 2.7배
심혈관계질환 2배	치매 1.7배
성기능장애 1.5배	골다공증 1.21배
류머티스관절염 1.17배	

자료: 대한치주과학회

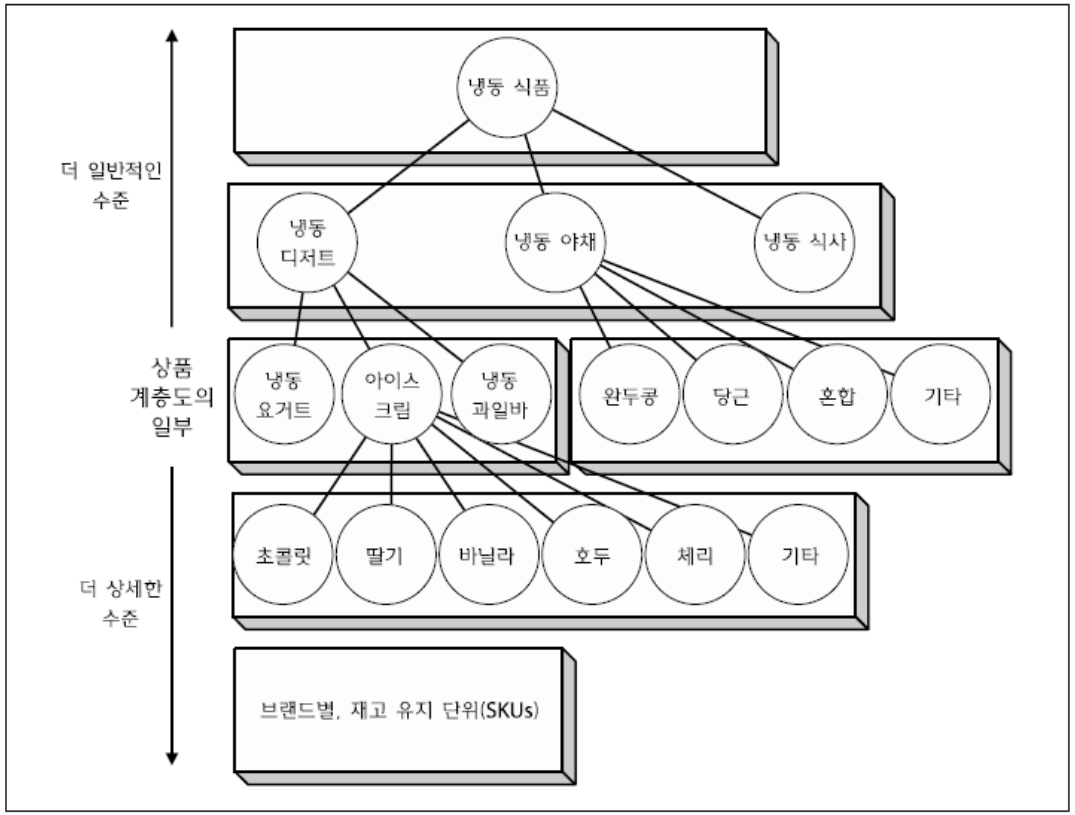
연관성 분석 -개요

- 연관성 분석을 활용한 대표적인 마케팅 방법
  - 교차판매(Cross Sell)와 상향판매 (Up Sell)
  - 이미 확보된 고객에게 더 많은 서비스를 판매함으로써 한 고객으로부터 더 많은 매출을 발생하게 하는 것이 주 목표이다

교차판매 (Cross Sell)	상향판매 (Up Sell)
또 다른 구매의 결과로 고객에게 다른 제품 또는 서비스를 판매 하는 행위를 의미함	기존 고객들이 이윤을 많이 내는 제품들을 구매하도록 유도하는 판매하는 행위를 의미함

연관성 분석 - 고려사항

상품을 어떻게 구분할 것인가?



〈그림 9-10〉 상품 계층도는 가장 일반적인 것에서 시작하여 점차 상세한 것으로 이동한다.

- **보다 구체적인** 결과일수록 결과들이 행동으로 옮겨지기 쉬울 가능성이 높음
- **상품 계층을 올라가면** 올라갈수록 물품의 수는 줄어듦
- **일반화된 물품**들은 충분한 지지도를 가진 규칙을 찾는데 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

연관규칙 작업 기준

- 트랜잭션 집합이 주어질 때, 연관 규칙 마이닝의 목표는 다음 조건을 갖는 모든 규칙을 찾는 것
  - support ≥ minsup 임계값
    -> 빈발 항목집합Frequent Itemset :minsup 임계값보다 크거나 같은 항목집합
  - confidence ≥ minconf 임계값

- 무차별 대입 접근 : Brute-force approach:
  - 가능한 모든 연관 규칙 나열
  - 각 규칙에 대한 지지도와 신뢰도 계산
  - minsup와 minconf 임계값에 포함 안된 규칙 삭제 ->계산 금지Computationally prohibitive

연관규칙 절차

연관규칙을 절차

(예) 언론 웹사이트에 접속한 독자의 페이지 방문내역

카테고리: 뉴스, 정치, 금융, 연예, 스포츠, 예술

세션(또는 거래): 한 사용자가 특정 기간 내에 이 카테고리들의 콘텐츠에 접근하는 것

오프라인과 비교: 세션 ↔ 거래, 엑세스한 페이지(분야) ↔ 구매한 항목

단계 1: 거래(transaction) 형태의 데이터 준비

특별한 형태의 입력 데이터 필요: 거래 유무를 나타내는 이진형

[온라인 뉴스 사이트에서 엑세스한 미디어 데이터]

Session ID	List of media categories accessed
1	{News, Finance}
2	{News, Finance}
3	{Sports, Finance, News}
4	{Arts}
5	{Sports, News, Finance}
6	{News, Arts, Entertainment}

[연관성 분석 데이터 형태]

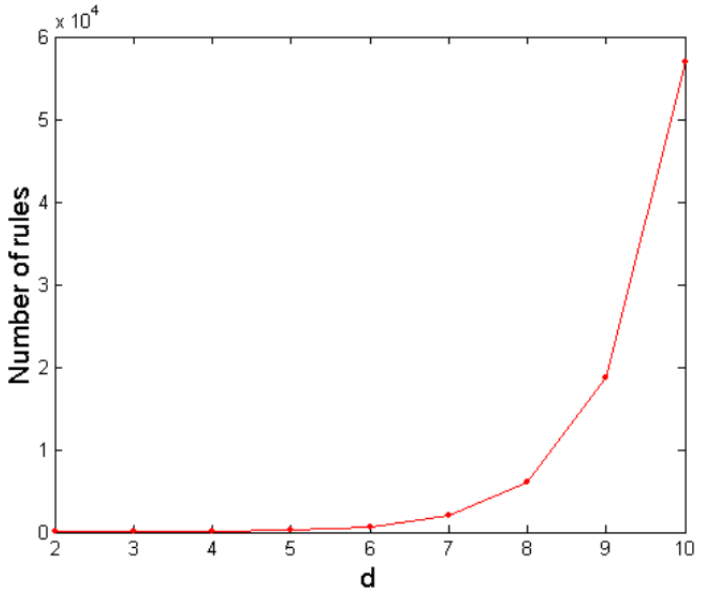
Session ID	News	Finance	Entertain	Sports	Arts
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	1	0
4	0	0	0	0	1
5	1	1	0	1	0
6	1	0	1	0	1

연관규칙 생성 절차

단계 2: 자주 발생하는 항목집합(빈발항목집합)의 후보자 목록 생성

- 가장 자주 발생하는 항목들로 분석을 제한하므로 마지막 단계에서 추출된 최종 규칙집합들이 좀 더 의미가 있음
- d개 항목에 대하여 $(2^d - 1)$개의 항목집합 생성 (null 제외)
- d개 항목인 경우 가능한 연관 규칙수

$$R = \sum_{k=1}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$
$$= 3^d - 2^{d+1} + 1$$

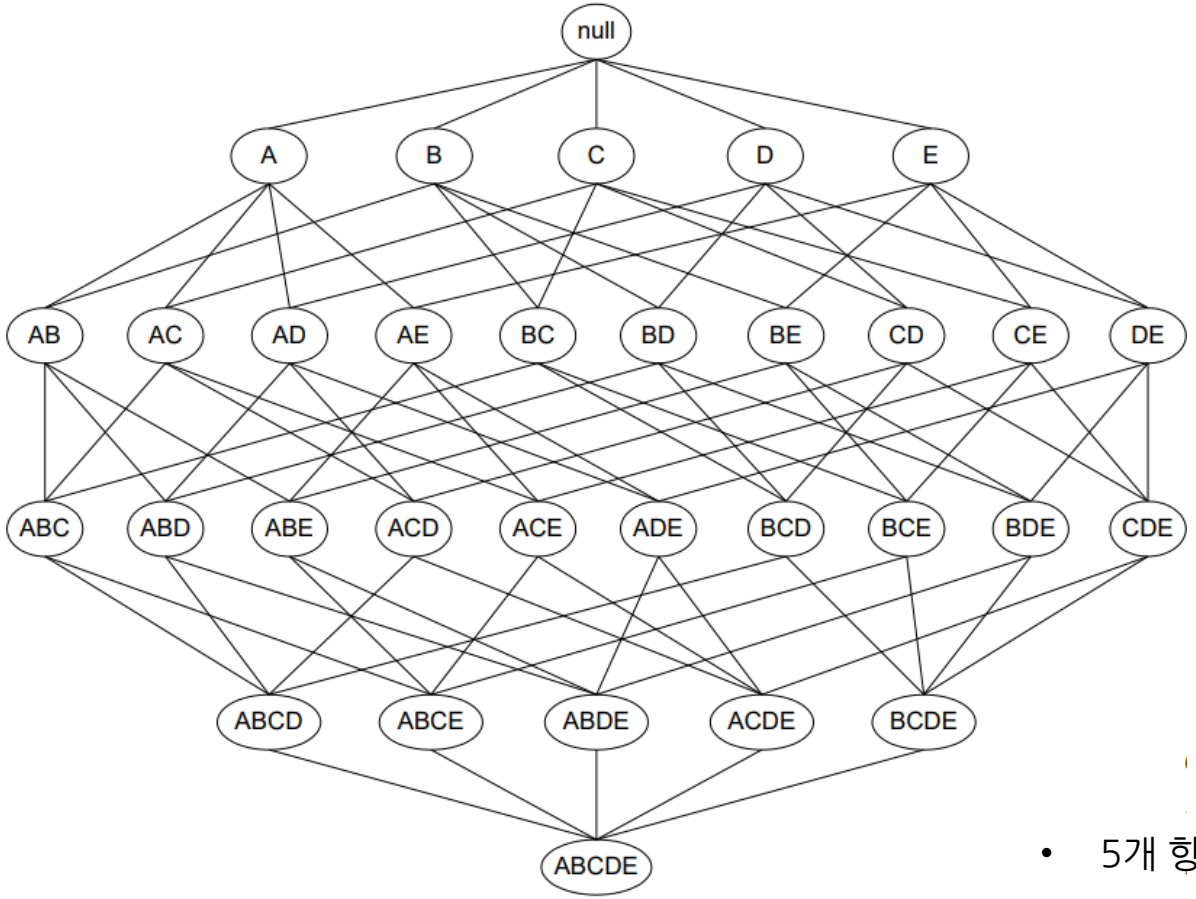


- 항목집합 수가 6개 : 연관규칙 수 602개
- 항목집합 수가 5개 : 연관규칙수 180개
- 항목집합 수가 4개 : 연관규칙수 50개

- 빈도가 작은 항목을 정리하기 위해 지지도의 최소 임계치를 정의
- 항목집합 트리(또는 격자 형태): 빈발항목 집합을 쉽게 찾아냄
- 빈발항목 집합을 찾는 효과적인 알고리즘으로는 선험적(Apriori) 알고리즘과 FP(Frequent Pattern) 성장 알고리즘이 가장 대표적

연관규칙 생성 절차

단계 2: 자주 발생하는 항목집합(빈발항목집합)의 후보자 목록 생성

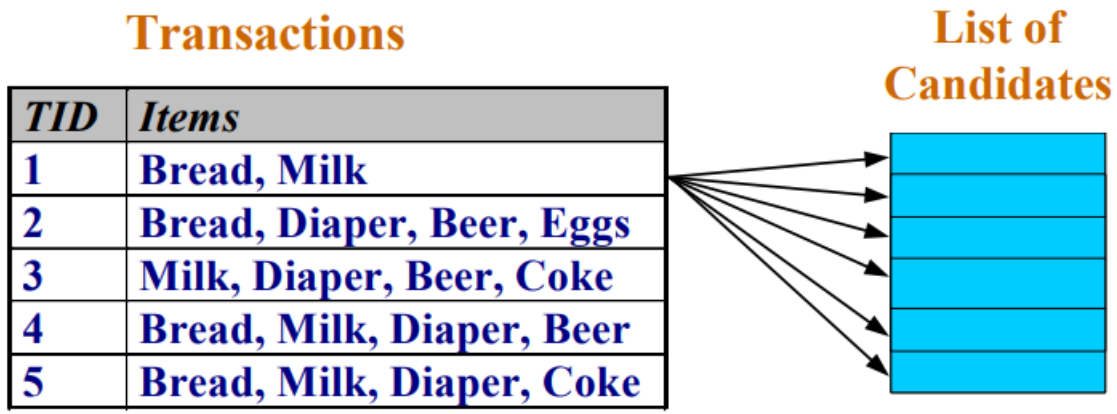


- 5개 항목에 대하여 $(2^5 - 1)$개의 항목집합 생성 (null 제외)
  :31개 항목 생성

연관규칙 생성 절차

단계 2: 자주 발생하는 항목집합(빈발항목집합)의 후보자 목록 생성

- 무차별 대입 접근Brute-force approach:
  - 격자의 각 항목집합은 빈발 항목집합 후보
  - 데이터베이스를 스캔하여 각 후보의 지지 횟수 계산
  - 모든 후보와 각 트랜잭션을 매치
  - 계산 비용이 큼



- 가지치기를 통해 빈발항목집합이 후보자 목록을 감수시키기

연관규칙 생성 절차

단계 3: 항목집합에서 관련성이 있는 연관규칙들을 생성

- 측정 척도에 따라 규칙들을 생성하고 선택
- 대표적인 측정척도 : 지지도, 신뢰도, 리프트
- 신뢰도의 최소 임계치보다 높은 값을 가지는 모든 규칙들을 추출

연관규칙 생성 절차

측정 척도: **지지도(support)**

지지도(X) = X의 거래수 / 전체 거래수

- 거래 집합에서 해당 항목집합의 상대적 발생빈도
- 발생빈도가 높은 항목들이어서 활용 및 조사가치가 있는 패턴을 발견함
- 낮은 지지도의 규칙은 드물게 발생하는 항목들이거나 우연히 항목 관계가 생성된 것이므로, 허위 규칙을 만들어 낼 수 있음
- 발생빈도가 드문 규칙을 제외하기 위해 지지도의 임계치를 명시
- 임계치 이상인 규칙에 대해서만 추가 분석

Session ID	News	Finance	Entertain	Sports	Arts
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	1	0
4	0	0	0	0	1
5	1	1	0	1	0
6	1	0	1	0	1

지지도({스포츠})= 2/6 = 0.33

지지도({뉴스.금융}) = 4/6 = 0.67

연관규칙 생성 절차

측정 척도: **신뢰도(confidence)**

신뢰도(X→Y) = 지지도(X∩Y) / 지지도(X)

- 선행조건을 포함하는 모든 거래들 중 규칙의 결과가 발생할 가능성
- 규칙이 믿을 만한지 측정

Session ID	News	Finance	Entertain	Sports	Arts
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	1	0
4	0	0	0	0	1
5	1	1	0	1	0
6	1	0	1	0	1

$$\text{신뢰도}(\{\text{뉴스}, \text{금융}\} \rightarrow \{\text{스포츠}\}) = \frac{\text{지지도}(\{\text{뉴스}, \text{금융}, \text{스포츠}\})}{\text{지지도}(\{\text{뉴스}, \text{금융}\})} = \frac{2/6}{4/6} = 0.5$$

뉴스와 금융 페이지를 방문하는 사용자들 중 50%는 스포츠 페이지도 방문한다는 것을 의미

연관규칙 생성 절차

측정 척도: **향상도(lift)**

향상도(X→Y) = 신뢰도(X→Y) / 지지도(Y)
= 지지도(X∩Y) / (지지도(X) x 지지도(Y))

- 규칙 X→Y가 의미가 있다면 전체 거래에서 항목 Y를 포함하고 있는 거래의 비율보다 항목 X가 포함된 거래 내에서 항목 Y를 포함하고 있는 거래의 비율이 더 클 것이다. 즉, P(Y) < P(Y|X)

$$향상도(\{뉴스, 금융\} \rightarrow \{스포츠\})$$

$$= \frac{지지도(\{뉴스, 금융, 스포츠\})}{지지도(\{뉴스, 금융\}) \times 지지도(\{스포츠\})}$$

$$= \frac{\mathbf{0.333}}{\mathbf{0.667 \times 0.333}} = \mathbf{1.5}$$

- 향상도 ≈ 1 : 두 항목은 거의 독립적, 흥미롭지 않음
- 향상도 > 1 : 두 항목은 양의 연관관계
- 향상도 < 1 : 두 항목은 음의 연관관계, '이다' 대신에 '아니다'를 사용하여 결과를 역으로 나타내는 것이 좋다. 그러나 원래의 연관규칙만큼 유용하지 않을 수 있음
  (예) "B와 C이면 A이다"의 신뢰도가 33%이면
  "B와 C이면 A가 아니다"의 신뢰도는 67%가 된다.
- 향상도 절대값이 클수록 좀 더 흥미로운 규칙

지지도(Y): 전체 거래에서 항목 Y를 포함하고 있는 거래의 비율

신뢰도(X → Y): 항목 X가 포함된 거래 내에서 항목 Y를 포함하고 있는 거래의 비율

향상도(X→Y): 연관규칙 (X→Y)로 인하여 항목 Y를 포함하고 있는 거래의 비율이 증가한 정도

연관규칙 생성 절차

Exercise

항목	거래의수
버섯	100
페페로니	150
치즈	200
버섯+페페로니	400
버섯+치즈	300
페페로니+치즈	200
버섯+페페로니+치즈	100
추가도핑안함	550
합계	2000

(버섯+페페로니) ➔ 치즈

(버섯+치즈) ➔ 페페로니

(페페로니+치즈) ➔ 버섯

버섯 ➔ 페페로니

에 대하여,

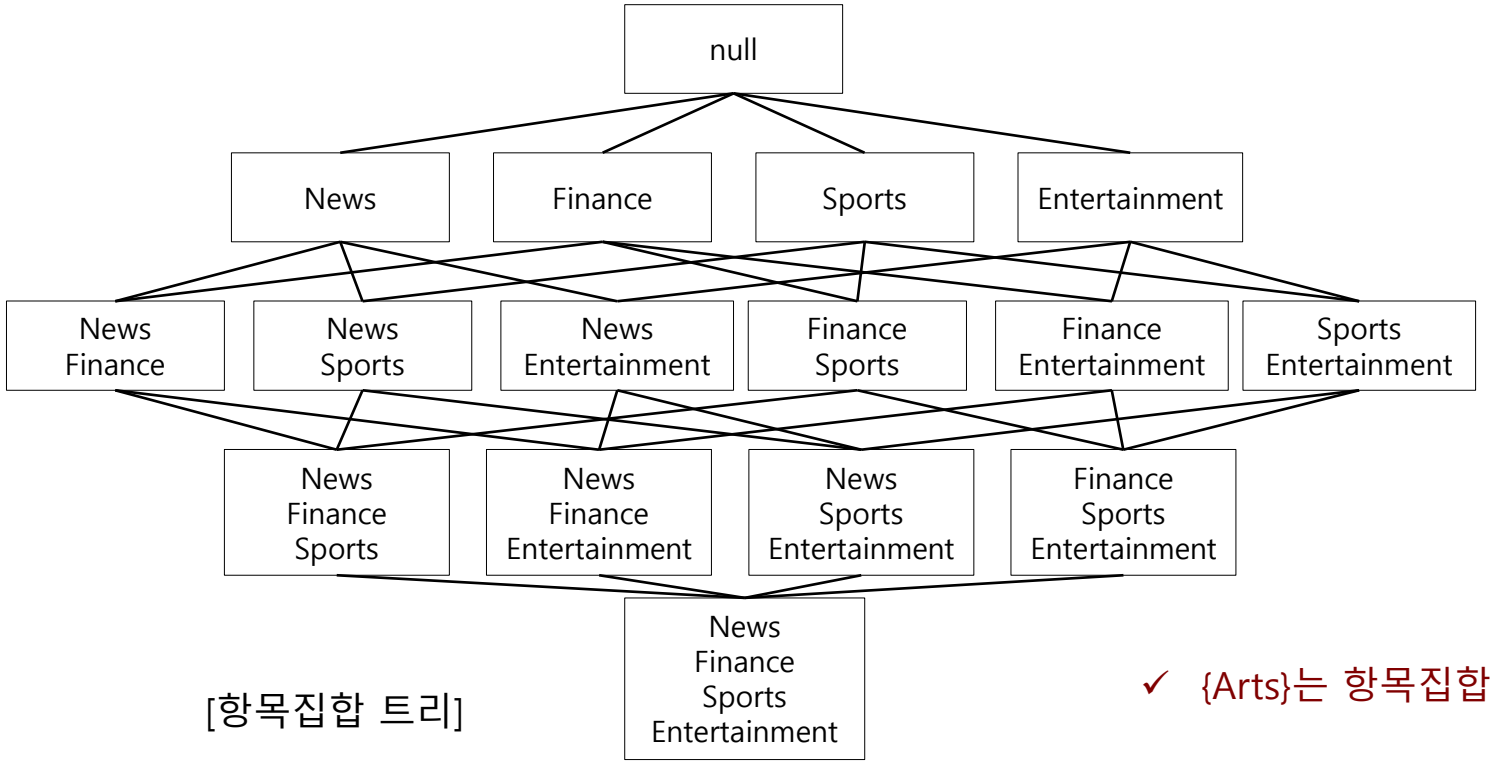
지지도 신뢰도 향상도를 구하십시오.

규칙(A➔B)	P(A)	P(B)	Support P(A∩B)	Confidence P(A∩B) /P(A)	Lift P(A∩B)/P(A)*P(B)
(버섯+페페로니) ➔ 치즈					
(버섯+치즈) ➔ 페페로니					
(페페로니+치즈) ➔ 버섯					
버섯 ➔ 페페로니					

연관규칙 생성 절차

빈발항목집합 찾는 알고리즘 : Apriori 알고리즘

- 항목집합의 지지도는 부분집합의 지지도를 결코 초과하지 않음
  - 한 항목집합이 빈발하다면, 이 항목집합의 모든 부분집합은 역시 빈발 항목집합이다.
  - 한 항목집합이 비반발하다면, 이 항목집합을 포함하는 모든 집합은 비빈발항목집합이다.



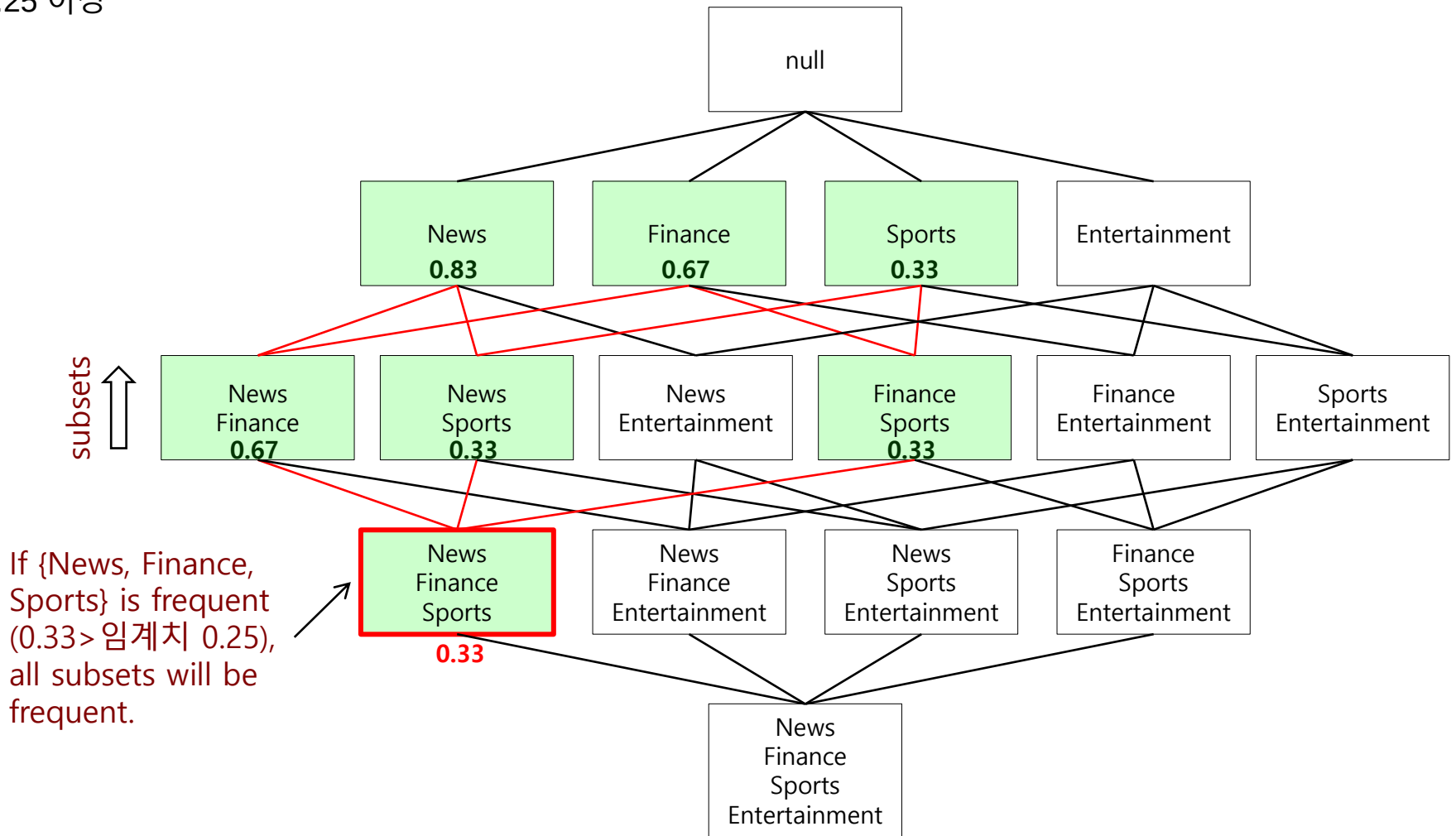
[항목집합 트리]

✓ {Arts}는 항목집합 생성에서 제외

연관규칙 생성 절차

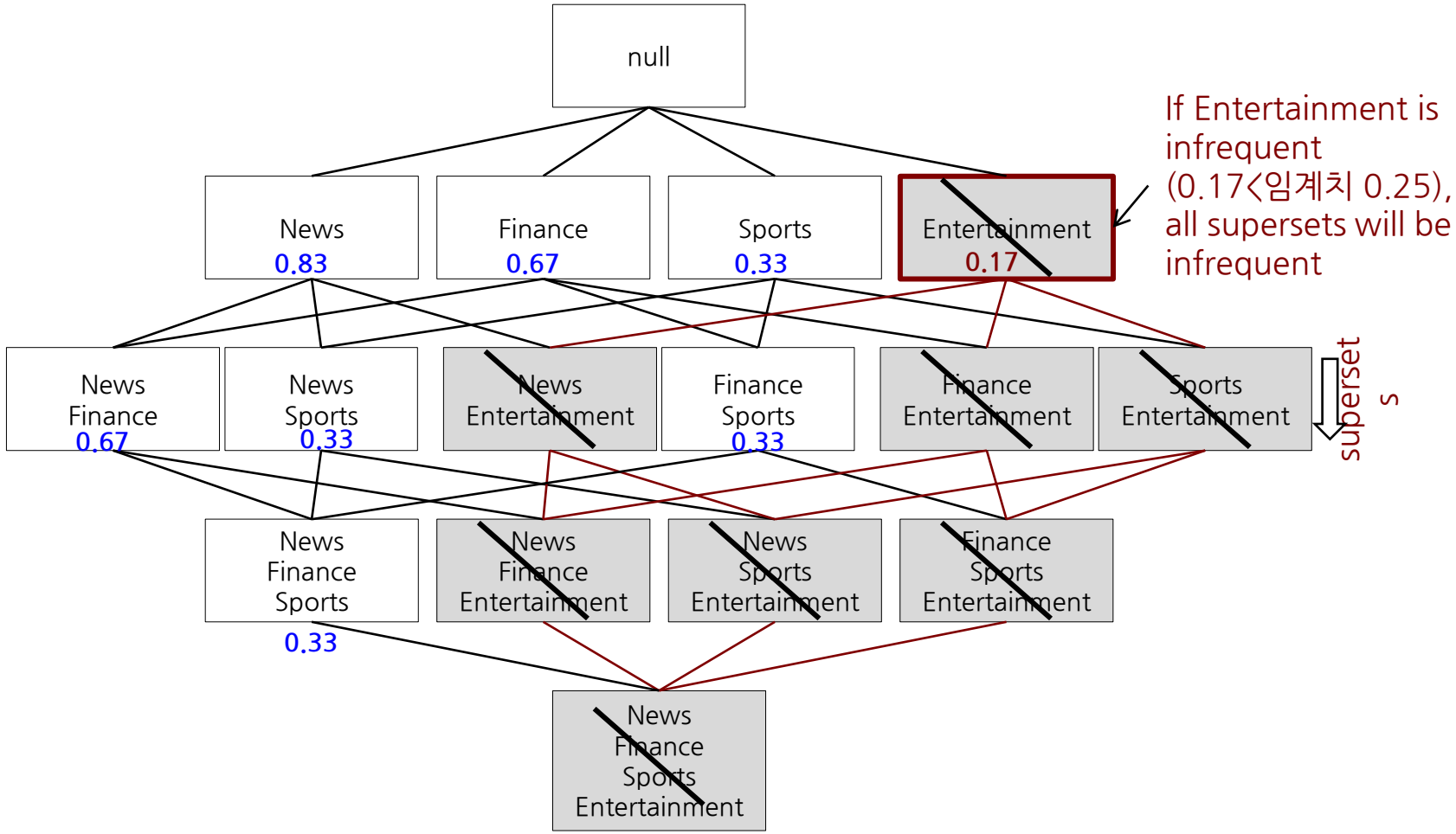
빈발항목집합 찾는 알고리즘 : Apriori 알고리즘

지지도 임계치 0.25 이상



연관규칙 생성 절차

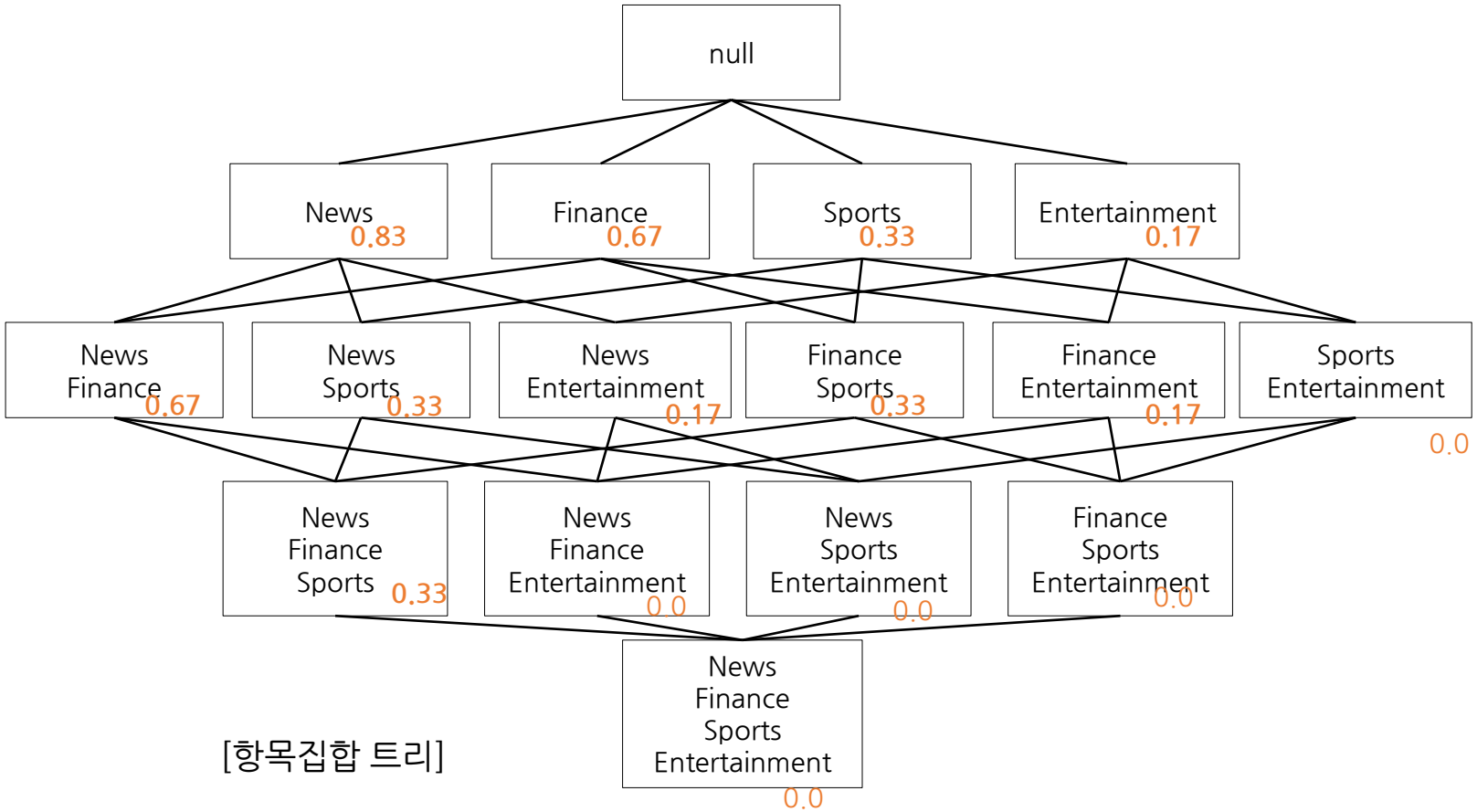
빈발항목집합 찾는 알고리즘 : Apriori 알고리즘



현재 가능한 항목집합은 15개 ($2^4 - 1$)
지지도가 임계치(0.25)보다 작은 {Entertainment} 항목 제외 → 가능한 항목집합은 7개 ($2^3 - 1$)

연관규칙 생성 절차

빈발항목집합 찾는 알고리즘 : Apriori 알고리즘



[항목집합 트리]

연관규칙 생성 절차

빈발항목집합 찾는 알고리즘 : 빈발패턴 성장 알고리즘 (FP-Growth Algorithm)

- 빈발패턴 나무(Frequent-Pattern tree, FP-tree)라고 불리는 특별한 그래프 데이터구조를 사용하여 거래 레코드들을 압축함으로써 빈발항목 집합을 계산
  - 먼저 빈발패턴 나무를 만들고 압축된 나무를 이용하여 빈발항목 집합을 생성
  - 알고리즘의 효율성은 거래 레코드가 얼마나 압축될 수 있는가에 따라 결정됨

- 빈발항목들 사이의 관계를 그래프를 이용하여 매핑하기 때문에, 연관성 분석 이외에도 문서 군집화, 텍스트 마이닝, 감성(sentiment) 분석의 전 처리 과정 등에 사용

- 실행방법에 차이가 있지만, 빈발패턴 성장 알고리즘과 선험적 알고리즘의 결과는 유사
  - 빈발항목 집합으로부터 규칙을 생성하는 것이 선험적 알고리즘과 유사하기 때문
  - 선험적 알고리즘과 다른 점은 패턴을 찾기 위해 후보를 만들지 않는 것임

- 빈발패턴 성장 알고리즘은 개념과 설명에 있어 그래프와 부분 그래프 분석이 포함되어 있긴 하지만, 프로그램 언어로 쉽게 개발될 수 있음

연관규칙 생성 절차

FP(Frequent-Pattern) Tree 생성

- 나무 다이어그램 형식으로 시각화하기 위해서는 거래 리스트를 나무 맵(tree map)으로 변환해야 함
- 나무 맵은 모든 정보를 보유하는 빈발경로들(frequent paths)을 나타냄

(예) 클릭 스트림의 데이터셋을 이용한 FP-tree의 생성 과정

- 1단계

- 각 거래에서 모든 항목들을 전체 빈도수의 내림차순으로 정렬
- 각 항목의 전체 빈도수: News 5 > Finance 4 > Sports 3 > Entertainment 1

Session ID	List of media categories accessed
1	{News, Finance}
2	{News, Finance}
3	{Sports, Finance, News}
4	{Sports}
5	{Sports, News, Finance}
6	{News, Entertainment}

빈도수 정렬

Session ID	Items
1	{News, Finance}
2	{News, Finance}
3	{News, Finance, Sports}
4	{Sports}
5	{News, Finance, Sports}
6	{News, Entertainment}

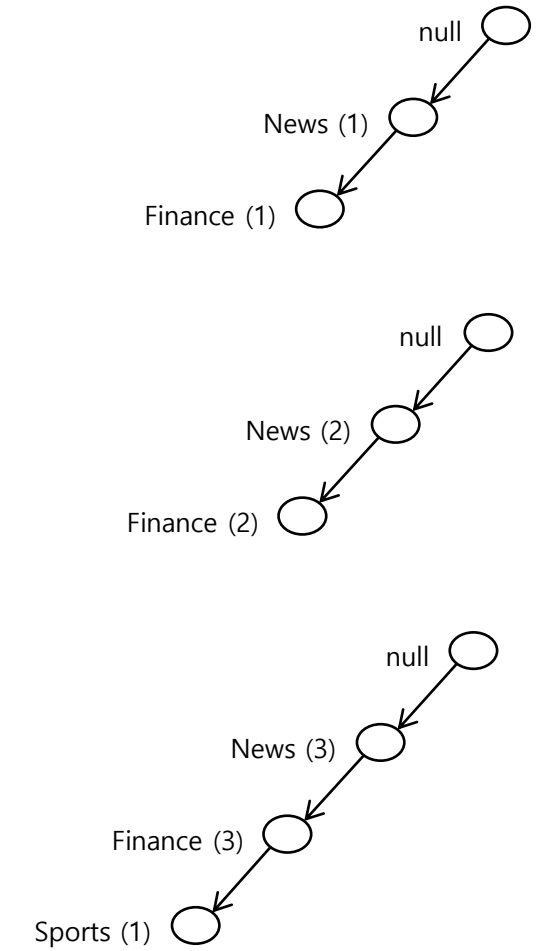
연관규칙 생성 절차

FP-Tree 생성

2단계
Session 1: {News, Finance}
빈발나무에 맵핑하기
널(null) 노드에서 시작
항목명 옆의 괄호 안의 숫자는 해당
경로를 따르는 거래의 개수

3단계
Session 2: {News, Finance}
첫 번째와 동일 → 같은 경로를 따르며
괄호 안의 숫자만 증가

4단계
Session 3: {News, Finance, Sports}
이 나무는 {Sports}로 확장된 것이며,
경로마다 숫자가 하나씩 증가



[빈발패턴 나무: Session 1~3]

Session ID	Items
1	{News, Finance}
2	{News, Finance}
3	{News, Finance, Sports}
4	{Sports}
5	{News, Finance, Sports}
6	{News, Entertainment}

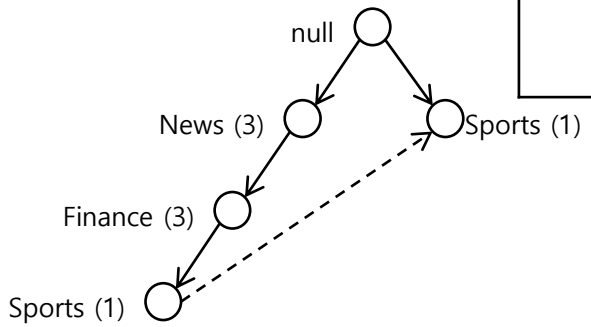
연관규칙 생성 절차

Session ID	Items
1	{News, Finance}
2	{News, Finance}
3	{News, Finance, Sports}
4	{Sports}
5	{News, Finance, Sports}
6	{News, Entertainment}

FP-Tree 생성

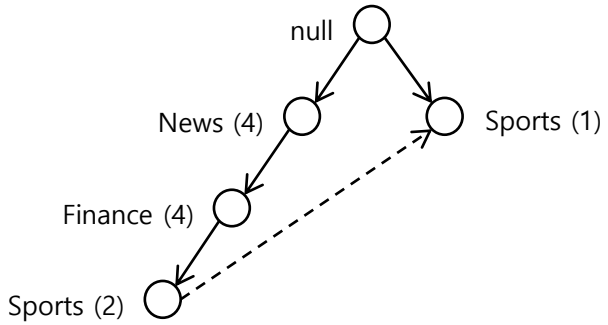
5단계
Session 4: {Sports}
한 항목만을 포함
→ 널(null) 항목에서 시작하여 새로운 경로를 생성

이 {Sports} 노드는 {News}와 {Finance} 다음에 오는
{Sports} 노드와는 다르지만, 두 노드는 동일 항목을
나타내므로 점선으로 연결

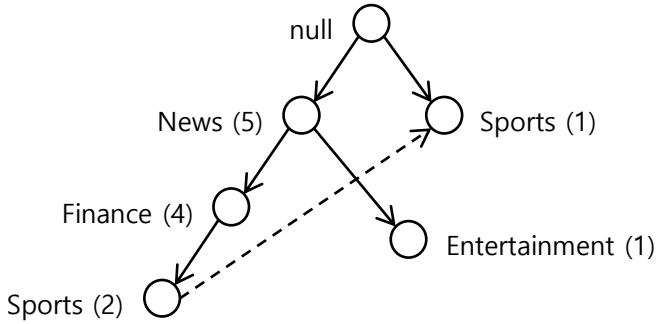


[빈발패턴 나무: Session 1~4]

6단계
위의 프로세스를 모든 거래들이 스캔될 때까지 계속
마지막 결과로 모든 Session 레코드들은 축약된 빈발패턴 나무로 표현됨



[빈발패턴 나무: Session 1~5]



[빈발패턴 나무: Session 1~6]

연관규칙 생성 절차

빈발항목 집합 생성

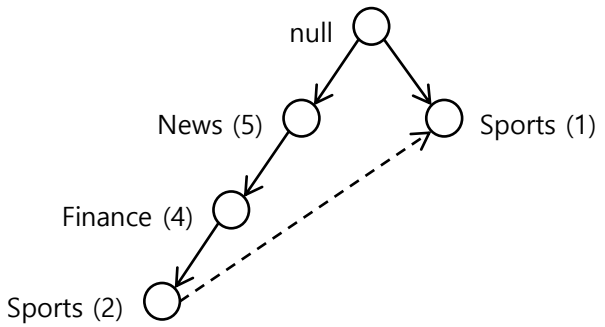
빈발항목 집합을 생성하기 위해서 최소 빈발항목에서 시작하여 모든 항목집합을 생성하는 상향식(bottoms-up) 접근방법을 사용

나무 구조는 지지도 개수(전체 빈도수)로 정렬되기 때문에, 최소 빈발항목은 나무의 잎 (leaf)에서 발견 됨

(예) {Entertainment}가 빈발항목이 아니므로 (임계치 조건 불만족) 무시

{Sports}로 끝나는 모든 가능한 항목집합들인 {Sports}, {Finance, Sports},

{News, Sports}, {News, Finance, Sports}를 찾음



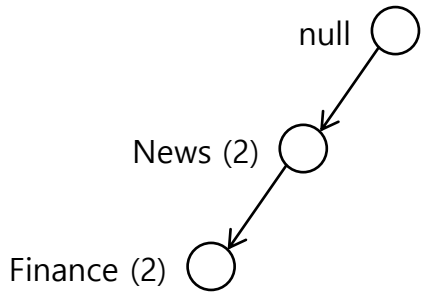
[정리된 빈발패턴 나무]

연관규칙 생성 절차

빈발항목 집합 생성

조건부 빈발패턴 나무

- 특정 항목으로 끝나는 항목집합 전부를 찾기 위해 그 항목의 접두 경로와 조건부 빈발패턴 나무를 생성한다.
- 항목의 접두 경로: 그 항목을 포함하는 경로만을 가지는 부분 나무
- 항목의 조건부 빈발패턴 나무: 빈발패턴 나무에서 그 항목을 제외한 것
  (예) {Sports} 항목의 조건부 빈발패턴 나무



[{Sports}의 조건부 빈발패턴 나무]

연관규칙 -요약

- 항목집합(item sets):규칙의 선행 또는 결과 부분에 나타나는 항목들 (예) {Sports}, {News, Finance} 등
- 생성된 규칙들은 지지도, 신뢰도, 향상도 등의 척도들로 평가
- 발생빈도가 적고 관련이 적은 규칙들을 걸러낼 수 있도록 지지도와 신뢰도의 적정한 임계치 설정 중요
- 빈발항목 집합을 찾는 효과적인 알고리즘으로는 선험적(Apriori) 알고리즘과 FP(Frequent Pattern) 성장 알고리즘이 가장 대표적

연관규칙 -요약

- 장점:
  - 결과가 분명하다. (if-then 규칙)
  - 이해하기 쉽고 실제 적용이 용이하다.
  - 방대한 자료 분석의 시작점으로 적합하다.
  - 사전에 분석방향이나 목적이 특별히 없는 경우 매우 유용하다.
  - 데이터의 자료구조가 간단하다.
  - 변수의 개수가 많은 경우에 쉽게 사용될 수 있다.
  - 계산이 용이하다.

- 단점:
  - 품목 수의 증가에 따라 계산량이 폭증한다.
  - 자료의 속성에 제한이 있다.

    예) 구매자의 개인정보 중 나이 등의 연속형 변수를 사용할 수 없다.
  - 적절한 품목을 결정하기가 어렵다.
  - 거래가 드문 품목에 대한 정보를 찾기가 어렵다.

순차(Sequential) 연관규칙

- 시계열 자료와 같이 사건들이 어떤 순서로 일어나고, 이 사건들 사이의 연관성을 알아내는 것.

  - (예) 카드 결제 청구가 평소보다 많다는 것과 다음 달에 현금서비스를 받는 것은 연관성이 있다. vs. 연관성 분석은 동시에 일어나는 품목에 대해서 다룬다

- 같은 고객의 구매패턴을 기반으로 구매패턴이 시간에 따라 연관이 있는지를 알려고 할 때 사용된다.

  - (예) 슈퍼마켓에서 고객들의 물품구매 행태

    - 고객 A가 처음에는 담배와 술을, 다음 날에는 담배와 신문을, 그 다음 날에는 음료수와 과자를 구매했을 때

    - 시간적인 순서를 고려하게 되면 ({담배, 술}, {담배, 신문}, {음료수, 과자})와 같이 거래의 순열로 표현 가능 → "사용자 순열"

- 모든 사용자 순열 중에서 몇 % 이상 공통으로 나타나는 순열을 찾는 것