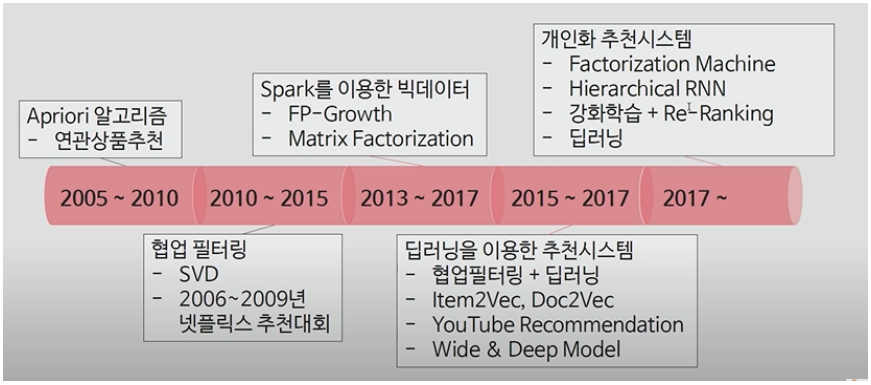
연관규칙분석(룰 기반 모델)

Writer	최윤서(학부학생/공과대학 도시공학)
키워드	
Reference	
주차	1주차

추천시스템이란

사용자에게 상품을 **어떻게 추천**할지에 대한 이해

- 추천시스템 흐름



- 데이터 유형
  - Implicit feedback data
    - 같이 구매한 것은 맞지만 만족 or 불만족했다는 feedback을 얻을 수 없음
    - ex) 유튜브 시청시에 좋아요 or 싫어요 표시하지 않았으면 feedback에 대한 정보 없이 같이 시청했다는 것을 기반으로 추천
  - Explicit feedback data

- 유저가 자신의 선호도를 직접 표현한 데이터(평점에 대한 정보가 포함되어 있음)
- ex) 영화 평점 데이터

연관규칙분석(룰 기반 모델)

상품과 상품 사이의 연관된 규칙을 기반으로 추천 (= 장바구니 분석)
ex) 얼마나 같이 구매가 되는가? / A 아이템을 구매하는 사람이 B 아이템을 구매하는가?

연관 규칙 평가지표

- 지지도 support(A→B) = P(A)
  - 조건절이 발생할 확률. 소프트웨어 상에서 P(A,B) 확률로 계산하기도 함.
  - 조건절이 많이 일어날때 추천해줄 기회가 높게 생기므로, 규칙의 범용성 측면에서 높은 것이 좋음.
- 신뢰도 confidence(A→B) = P(B|A)
  - 조건부 확률을 의미. A를 구매했을때 둘다 구매하는 확률.
- 향상도 lift = P(A,B) / P(A)P(B)
  - 연관 규칙이 독립적으로 구매되는 경우에 비해 얼마나 의미 있는가.
  - lift = 1 ; A와 B가 통계적으로 독립이다 = 우연히 같이 구매된 것이다
  - lift > 1 ; 독립으로 가정했을 때의 빈도보다 둘이 같이 나오는 빈도가 더 높다. positive relationship이 있다
  - lift <1 ; negative relationship이 있다
- 신뢰도 vs 향상도

  B가 기본적으로 구매하는 아이템일 경우에는 confidence(A→B)가 거의 1의 값이 나온다. 따라서 confidence만 가지고 좋은 규칙이라고 판단하기는 어렵고 lift(A→B)와 같이 판단하여야 한다.

모든 경우의 수 연관분석

- How

  주어진 아이템을 통해 만들 수 있는 모든 경우의 수를 나열

지지도, 신뢰도, 향상도가 높은 규칙을 찾아내는 방식

- 단점

  아이템의 증가에 따라 규칙의 수가 기하급수적으로 증가한다

  연관관계≠인과관계. 인과관계는 파악하기 어려움.

Apriori 알고리즘

모든 경우를 보지 않고 **효과적으로 연관 규칙**을 찾기 위한 방식.

Apriori considers **only frequent item sets**. (minimum support를 만족하지 못하는 규칙은 제거)

- How
  1. k개의 item을 가지고 단일항목집단 생성 (아이템 한개씩 구매하는)
  2. 단일항목집단에서 **최소 지지도 이상(하이퍼파라미터)의 항목만** 선택

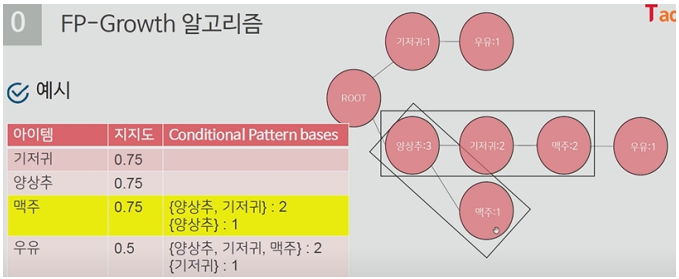
     (현실에서는 좋은 하이퍼 파라미터를 찾는 것이 중요)
  3. 2에서 선택한 항목만을 대상으로 2개 항목집단 생성
  4. 2개 항목 집단에서 최소 지지도 or 신뢰도 이상의 항목만 선택
  5. 위의 과정을 k개의 항목 집단 생성할때까지 반복
- 단점

  이 역시 데이터 클 경우 속도 느리고 연산량 많음.

FP-Growth 알고리즘

트리 구조를 사용하여 Apriori의 속도측면의 단점을 개선한 알고리즘

- How
  1. 각 아이템마다의 지지도를 계산하고, 최소 지지도 이상의 아이템만 선택
  2. 모든 거래에서 빈도가 높은 아이템 순서대로 순서를 정렬
  3. 트리를 생성
     - 새로운 아이템이 나오면 root 노드에 이어주기
     - 기존의 거래에 있는 경우에는 기존 노드에서 확장
  4. 지지도가 낮은 아이템부터 **조건부 패턴**을 생성 (모든 아이템에 대해 반복)



5. 조건부패턴을 통해서 신뢰도, 지지도를 통해서 좋은 연관 규칙들 찾기

- 장점

  속도가 빠르고 후보 Itemset을 생성할 필요가 없음

- 단점

  대용량 데이터셋에서 메모리 한계

  지지도 계산이 트리 만들어지고 나서야 가능