

데이터마이닝팀

4팀

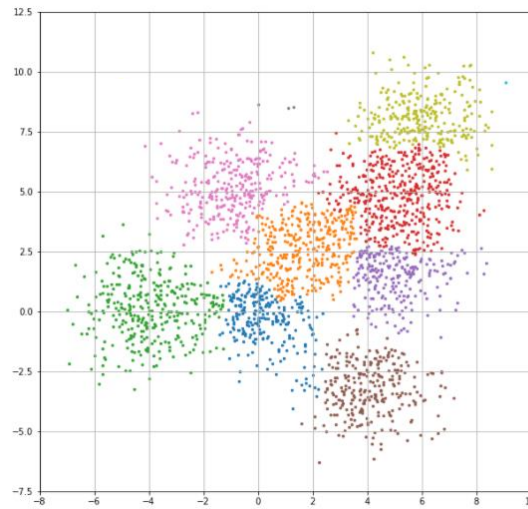
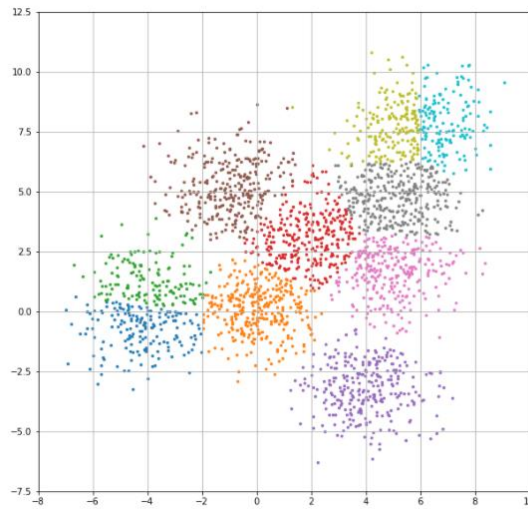
장이준
이선민
김영호
김현우
박시언

1

클러스터링

Clustering

클러스터링



데이터 내에서 **그룹**을 찾아내는 것이 목표



Silhouette Method



실루엣 방법

각 데이터 별로 실루엣 계수를 확인하는 방법

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}$$

$a(i)$

군집 내 거리 **Intra-cluster variance**

객체 i 와 객체 i 와 같은 군집 안에 속하는 나머지 객체들 간의 거리의 평균

$b(i)$

군집 간 거리 **Inter-cluster variance**

객체 i 와 객체 i 와 다른 군집에 속하는 나머지 객체들 간 거리의 평균의 최솟값

Elbow point method



Elbow Point Method


클러스터 내 RSS가 **최소**가 되도록 클러스터의 중심을 결정해 나가는 방법



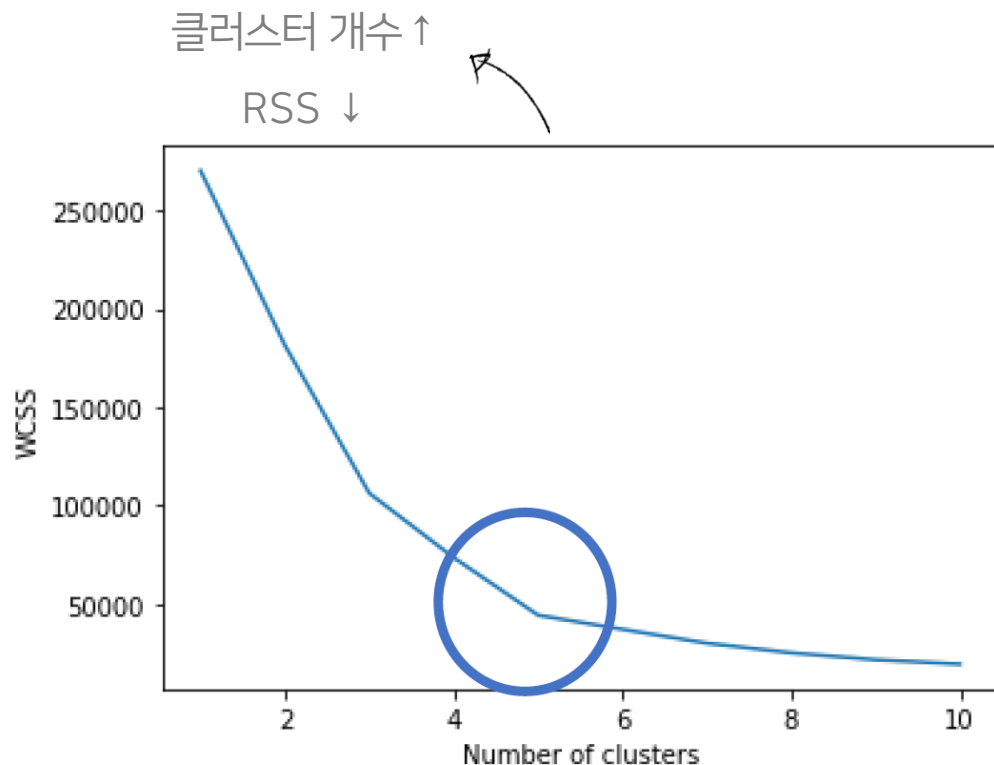
클러스터 내 중심점과 객체들 간의 거리, 즉
RSS가 **최소**가 되게 하는 중심점을 고르는 문제



클러스터 개수 ↑ 

RSS ↓ 

Elbow point method



“ Elbow point ”

오차의 합이
급격하게 감소하는 지점



해당 지점에서 클러스터 개수 결정

Non-Hierarchical clustering | K-means

K-means clustering



WCSS

$$WCSS = \sum_{k=1}^K n_k \sum_{C(i)=k} \|X_i - \bar{X}_k\|^2 \quad \blacktriangleright \text{클러스터 내 분산}$$

$$n_k = \sum_{i=1}^N I(C(i) = k) : k\text{번째 클러스터 point 개수}$$

$$\bar{X}_{jk} = \frac{1}{n_k} \sum_{C(i)=k} X_{ij} : k\text{번째 클러스터의 } j\text{번째 속성의 평균}$$

$$\bar{X}_k = (X_{1k}, X_{2k}, \dots, X_{pk})$$

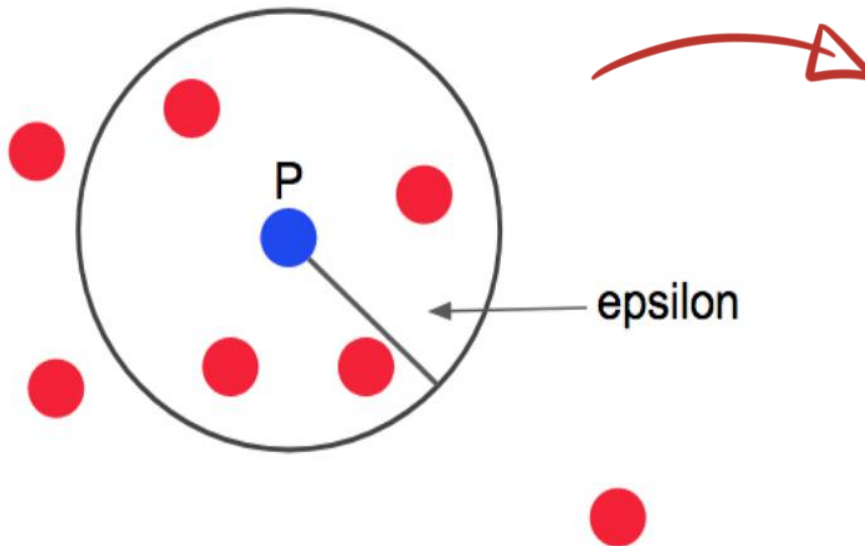


각각의 obs에서 그 obs가 포함된 클러스터의 중심점과의
거리의 합으로 WCSS 계산

DBSCAN

Terminology ϵ -Neighborhood of a point

점 p 의 ϵ -neighborhood q 는
 p 와의 거리가 ϵ 보다 작거나 같은 점들의 집합으로 정의



점 p 를 중심으로 군집을 이루려면
 p 의 ϵ -neighborhood q 들이
최소 minPts 이상 있어야 함!

Hierarchical Clustering

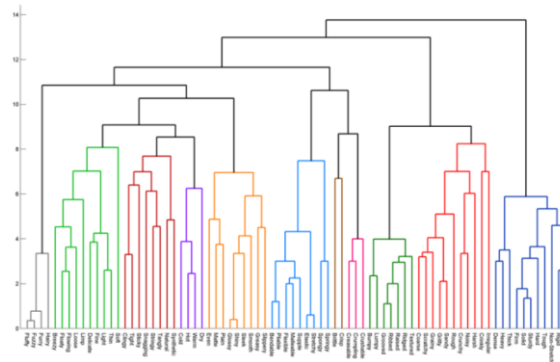


계층적 클러스터링

트리 모델을 이용하여 개별 개체들을 **순차적이고 계층적으로 유사한** 개체 혹은 그룹과 함께 클러스터를 만들어주는 알고리즘



클러스터의 개수를 사전에 정하지 않고도 학습 수행이 가능



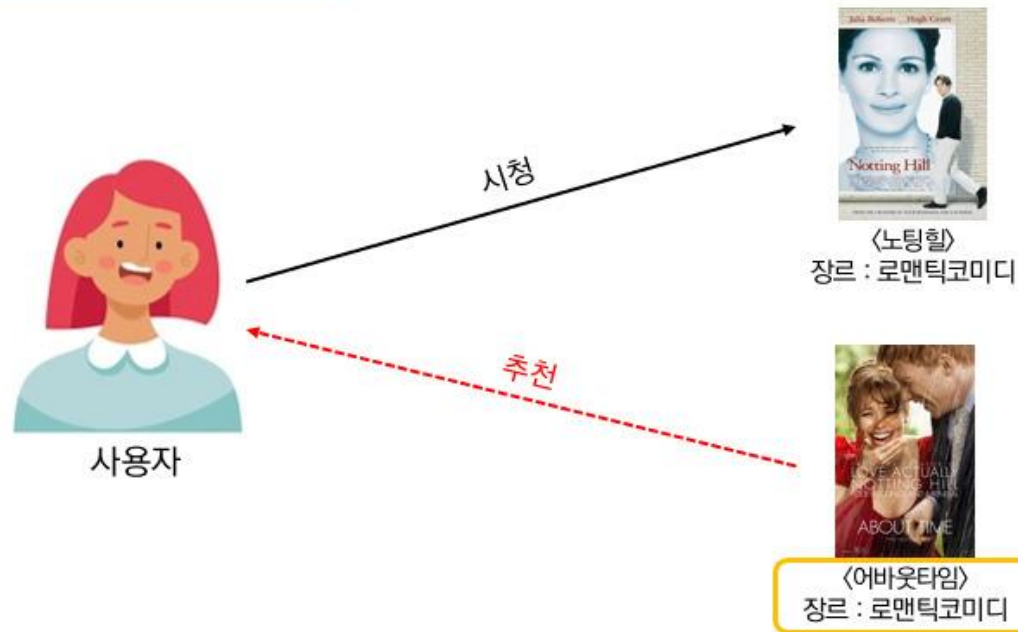
2

추천 시스템

추천 시스템 | 콘텐츠 기반 필터링

콘텐츠 기반 필터링(Content-Based Filtering)

내용 기반 필터링
(Content-based Filtering)



사용자가 과거에 소비했던 콘텐츠 특성을 분석하여
유사한 특성을 지닌 콘텐츠를 사용자에게 추천해주는 시스템

추천 시스템 | 콘텐츠 기반 필터링

TF-IDF(Term Frequency – Inverse Doc Frequency)



TF-IDF

어느 문서이든지 많이 나오는 단어에 대한 페널티를 부여함과 동시에
그 문서를 대표할 수 있는 주요 단어를 추출해내는 방법

$$TF - IDF = TF \cdot \log \frac{n_D}{1 + n_t}$$

TF : 하나의 문서 내에서 단어 t 가 나온 빈도수

n_D : 전체 문서 수 n_t : 단어 t 가 나온 문서 수

추천 시스템 | 콘텐츠 기반 필터링

코사인 유사도(Cosine Similarity)



코사인 유사도

두 벡터 A와 B의 코사인 각도를 이용하여 구하는 유사도

A_i : 벡터 A의 i번째 원소 B_i : 벡터 B의 i번째 원소

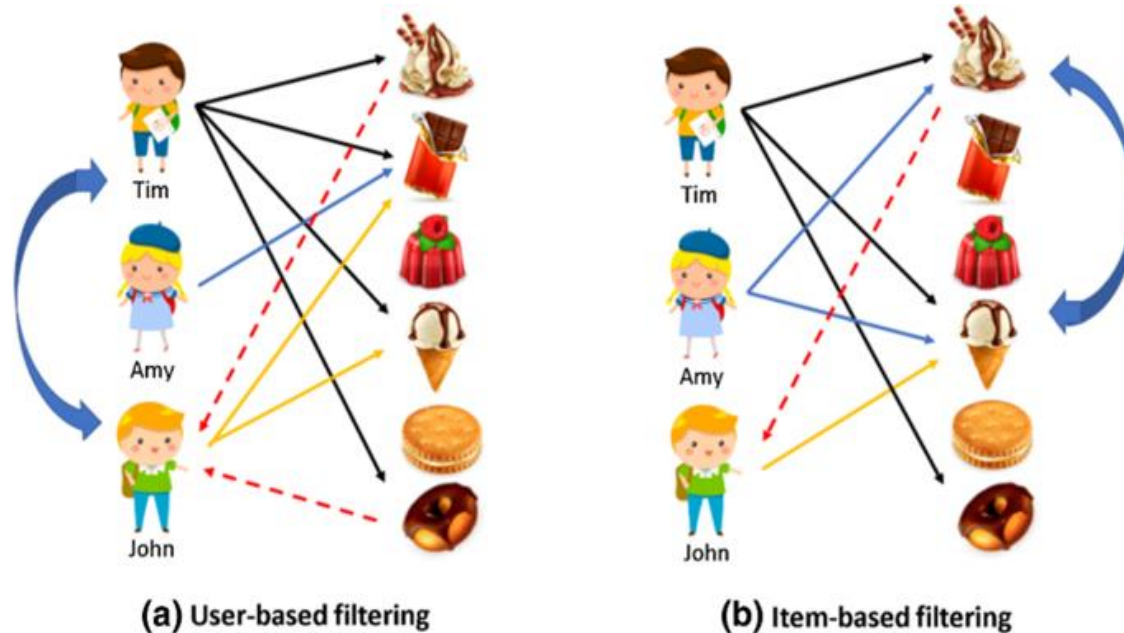
$$\text{cosine similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n (A_i \times B_i)}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

값이 1에 가까울수록 유사도가 높음!



추천 시스템 | 협업 필터링

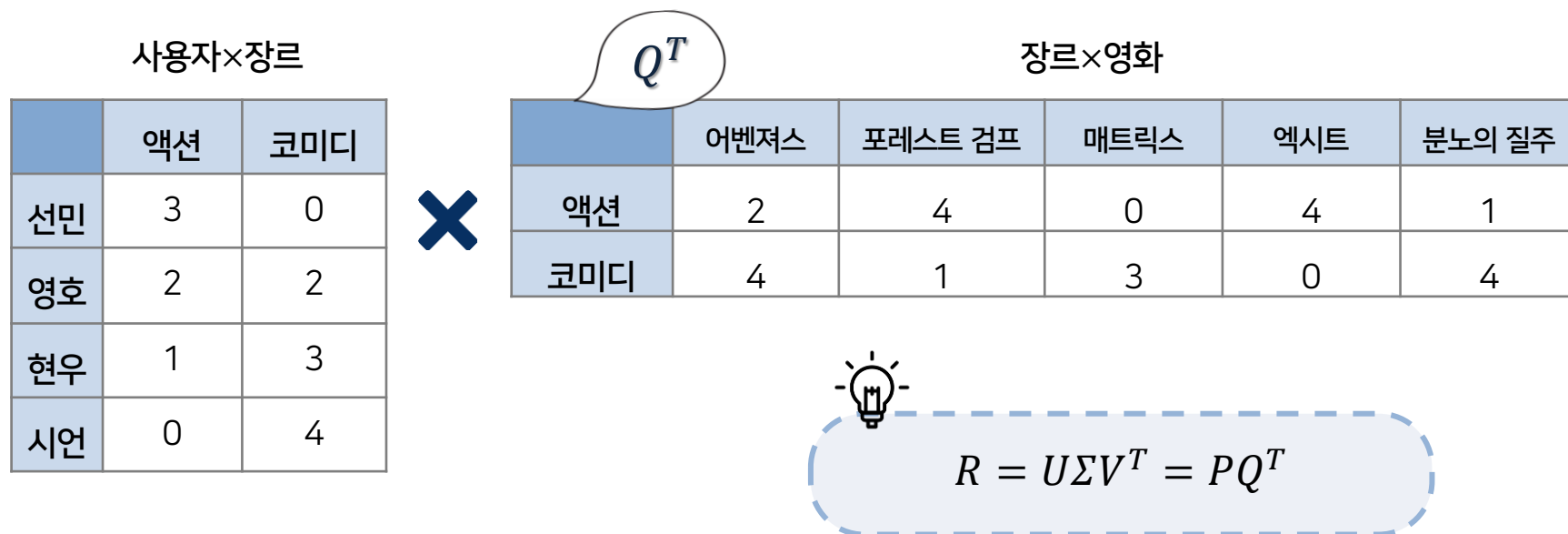
협업 필터링(Collaborative Filtering)이란?



구입내역, 선호도, 만족도를 기반으로 사용자 혹은 아이템 간의 협업(상호 작용 데이터)를 통하여
 비슷한 성향을 가진 사용자 선호하는 아이템,
 혹은 소비한 아이템과 유사한 아이템을 추천하는 시스템

추천 시스템 | 잠재 요인 협업 필터링

예시를 통한 잠재 요인 기반 협업 필터링



평점행렬 R 을 $R = PQ^T$ 의 형태로 분할하면
 사용자×장르, 장르×영화 2개의 행렬이 생성됨