Text Clustering: Similarity-based Approaches

ChengXiang “Cheng” Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Overview

- What is text clustering?
- Why text clustering?
- How to do text clustering?
  - Generative probabilistic models
  - **Similarity-based approaches**
- How to evaluate clustering results?

Similarity-based Clustering: General Idea

- Explicitly define a similarity function to measure similarity between two text objects (i.e., providing "clustering bias")
- Find an optimal partitioning of data to
  - maximize intra-group similarity and
  - minimize inter-group similarity
- Two strategies for obtaining optimal clustering
  - Progressively construct a hierarchy of clusters (hierarchical clustering)
    - Bottom-up (agglomerative): gradually group similar objects into larger clusters
    - Top-down (divisive): gradually partition the data into smaller clusters
  - Start with an initial tentative clustering and iteratively improve it ("flat" clustering, e.g., k-Means)

Similarity-based Clustering Methods

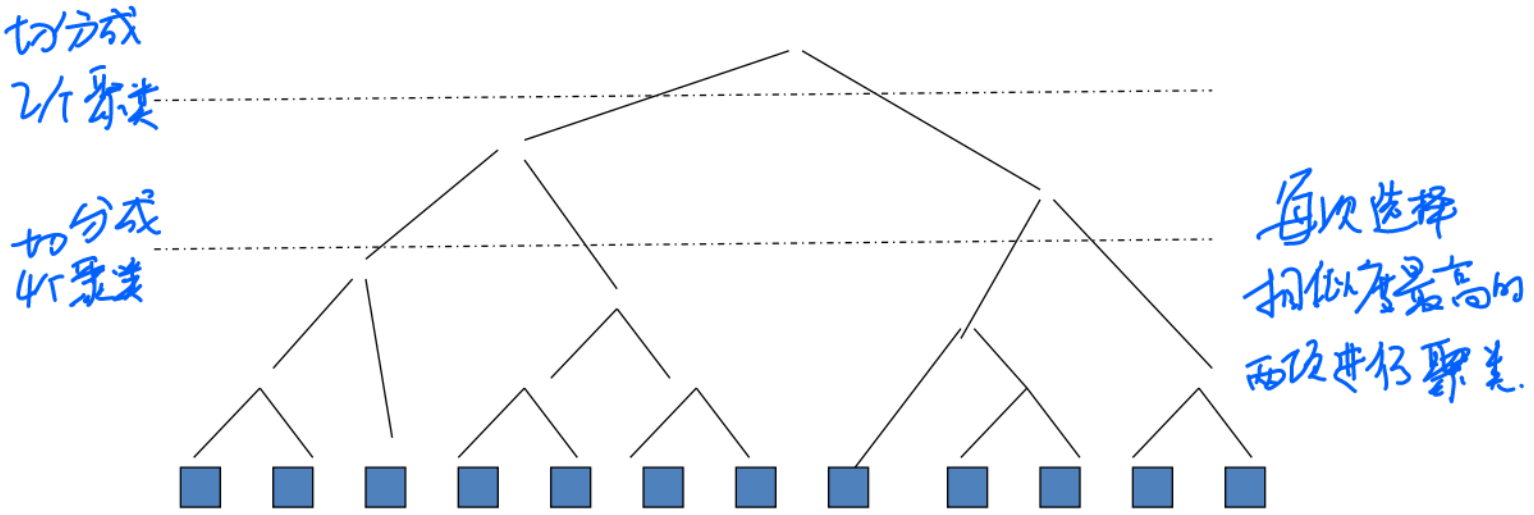
- Many general clustering methods are available!

- Two representative methods
  - Hierarchical Agglomerative Clustering (HAC)
  - k-means

Agglomerative Hierarchical Clustering (AHC)

- Given a similarity function to measure similarity between two objects
- Gradually group similar objects together in a bottom-up fashion to form a hierarchy 从小到大.
- Stop when some stopping criterion is met
- Variations: different ways to compute group similarity based on individual object similarity

Similarity-induced Structure



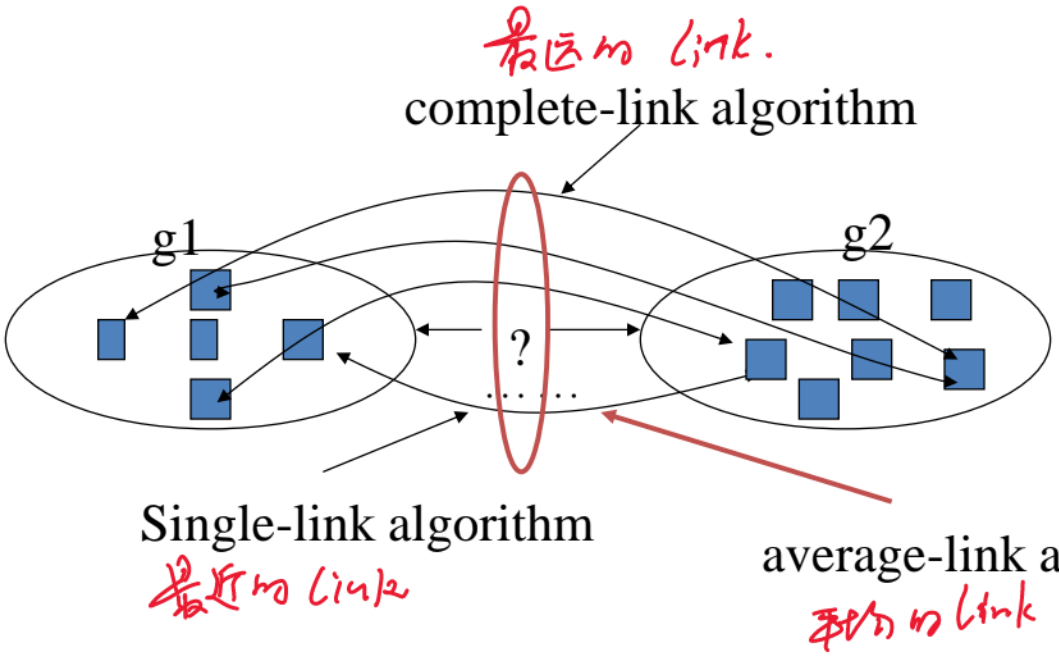
How to Compute Group Similarity

Three popular methods:

Given two groups g1 and g2,

- Single-link algorithm: s(g1,g2)= similarity of the closest pair
- Complete-link algorithm: s(g1,g2)= similarity of the farthest pair
- Average-link algorithm: s(g1,g2)= average of similarity of all pairs

Group Similarity Illustrated



Comparison of Single-Link, Complete-Link, and Average-Link

- Single-link
  - "Loose" clusters
  - Individual decision, sensitive to outliers
- Complete-link
  - "Tight" clusters
  - Individual decision, sensitive to outliers
- Average-link
  - "In between"
  - Group decision, insensitive to outliers
- Which one is the best? It depends on what you need!

K-Means Clustering

- Represent each text object as a term vector and assume a similarity function defined on two objects
- Start with k randomly selected vectors and assume they are the centroids of k clusters (initial tentative clustering) ➔ **Initialization**
- Assign every vector to a cluster whose centroid is the closest to the vector ≈ **E-step** **difference?**
- Re-compute the centroid for each cluster based on the newly assigned vectors in the cluster ≈ **M-step** **difference?**
- Repeat this process until the similarity-based objective function (i.e., within cluster sum of squares) converges (to a local minimum)

Very similar to clustering with EM for mixture model!

Summary of Clustering Methods

- Model based approaches (mixture model)
  - Uses an implicit similarity function (model ➔ clustering bias)
  - Cluster structure is “built” into a generative model
  - Complex generative models can discover complex structures
  - Prior can be leveraged to further customize the clustering algorithm
  - However, no easy way to directly control the similarity measure
- Similarity-based approaches
  - Allows for direct and flexible specification of similarity
  - Objective function to be optimized is not always clear
- Both approaches can generate both term clusters and doc clusters