UNIVERSITY of DE

# Text Clustering

# - Similarity based Methods

(CPEG 657: Search and Data Mining)

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

UNIVERSITY of DE

# Similarity-based Clustering

- Given a similarity function that can be used to measure similarity between two documents,
- Find a partition to
  - Maximize intra-cluster similarity
  - Minimize inter-cluster similarity
- How to find the partition?
  - Hierarchically group similar documents
  - Search by starting at a random partition

UNIVERSITY of DE

# Method 1 (Similarity-based):

# Agglomerative Hierarchical Clustering

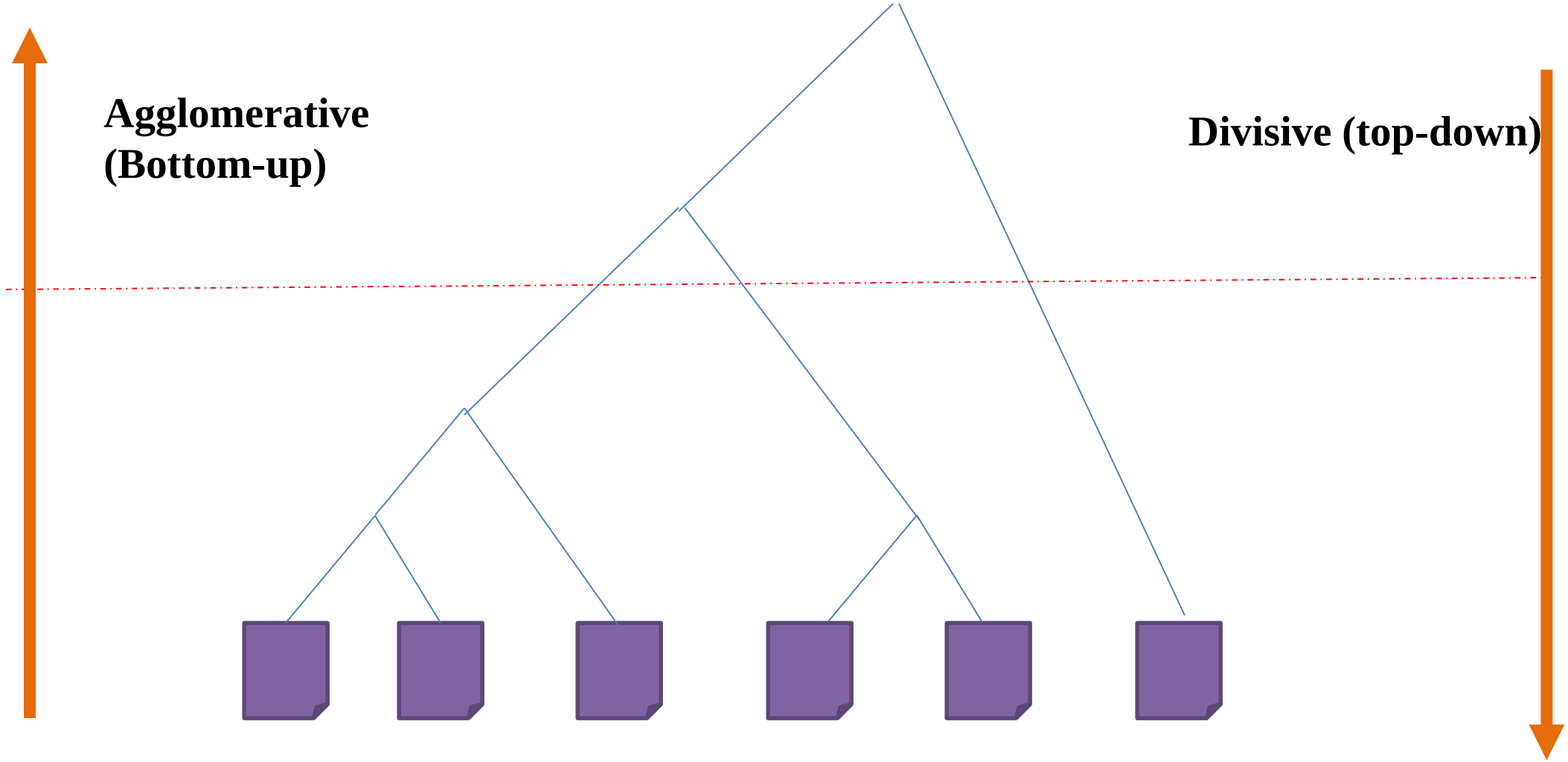
UNIVERSITY of DE

# Agglomerative Hierarchical Clustering

- Given a similarity function to measure similarity between two documents
- Gradually group similar documents together in a bottom-up fashion
- Stop when some stopping criterion is met

UNIVERSITY of DE

# Agglomerative Hierarchical Clustering



UNIVERSITY of DE

# How to Compute Group Similarity?

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

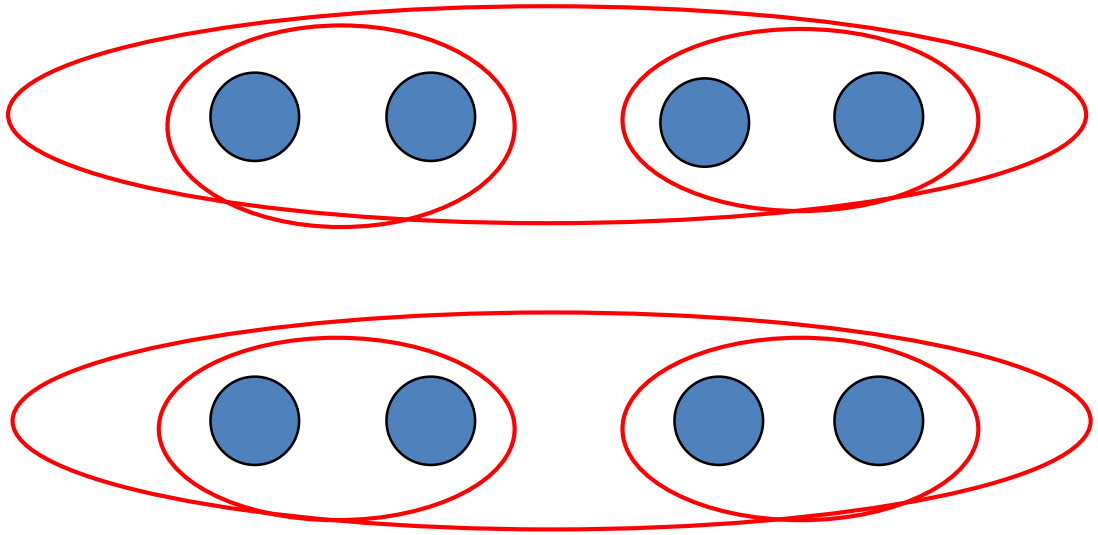
**average-link algorithm**: s(g1,g2)= average of similarity of all pairs

UNIVERSITY of DE

# Three Methods Illustrated

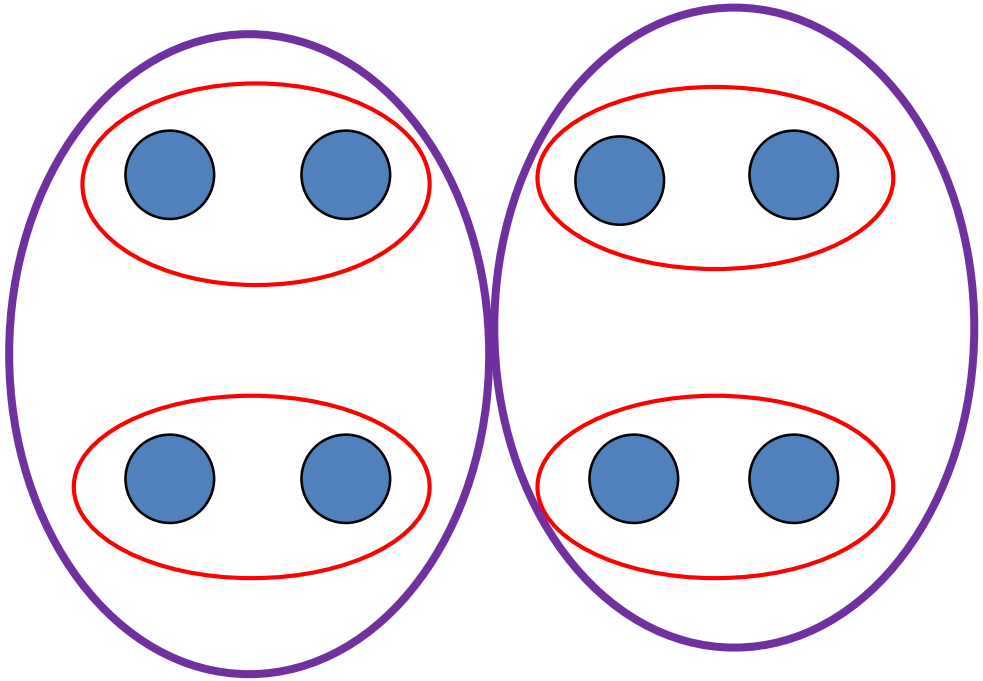
UNIVERSITY of DE

# Single Link Example



UNIVERSITY of DE

# Complete Link Example



UNIVERSITY of DE

# Comparison of the Three Methods

- Single-link
  - Loose clusters
  - Individual decision, sensitive to outliers
- Complete-link
  - Tight clusters
  - Individual decision, sensitive to outliers
- Average-link
  - Group decision, insensitive to outliers

UNIVERSITY of DE

# Method 2 (similarity-based):

# K-Means

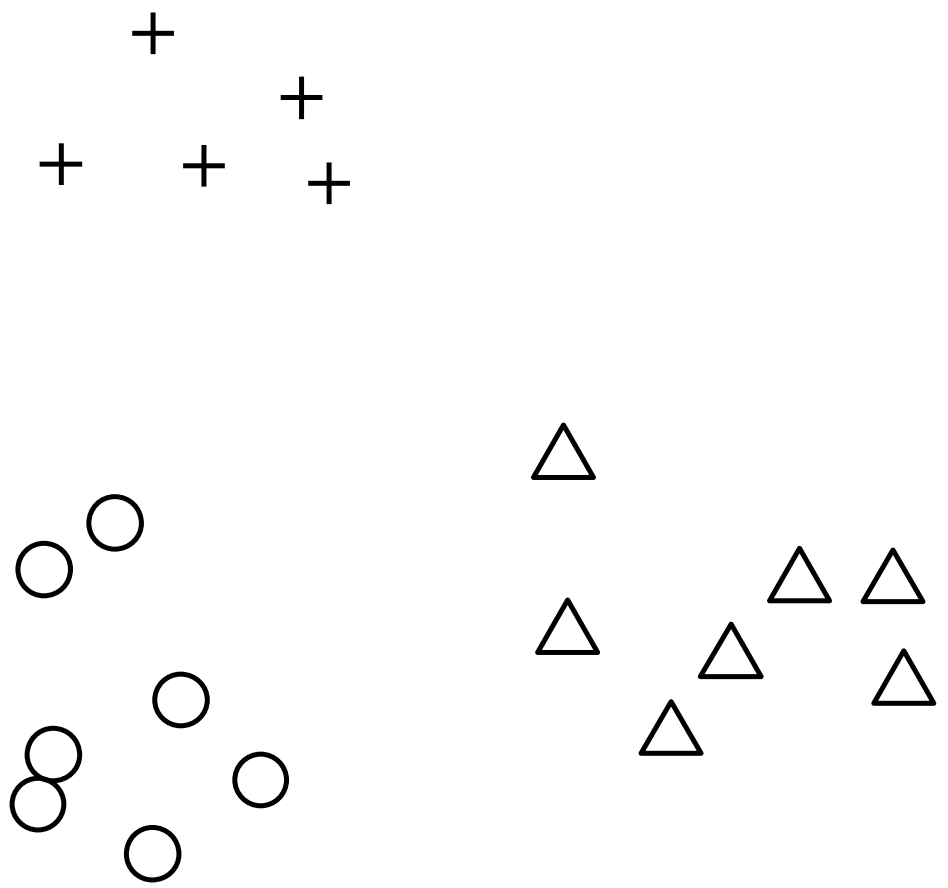
UNIVERSITY of DE

# K-Means Clustering

- Start with k randomly selected documents, and assume they are the centroids of k clusters
- Assign every document to a cluster whose centroid is the closest to the document
- Re-compute the centroid for each cluster
- Repeat this process until the centroids converge

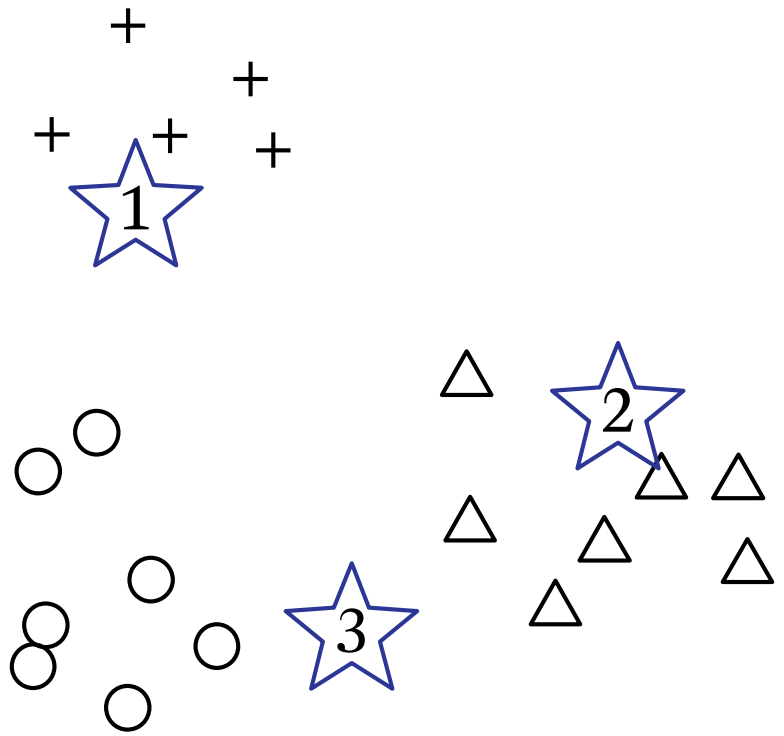
UNIVERSITY of DE

# An Example of K-Means



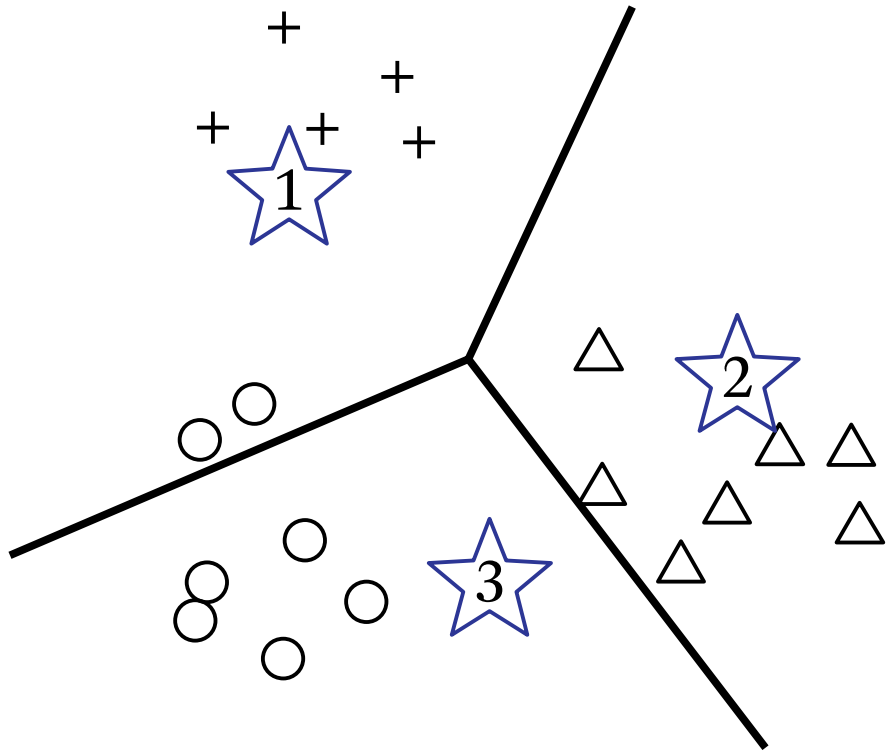
UNIVERSITY of DE

# An Example of K-Means



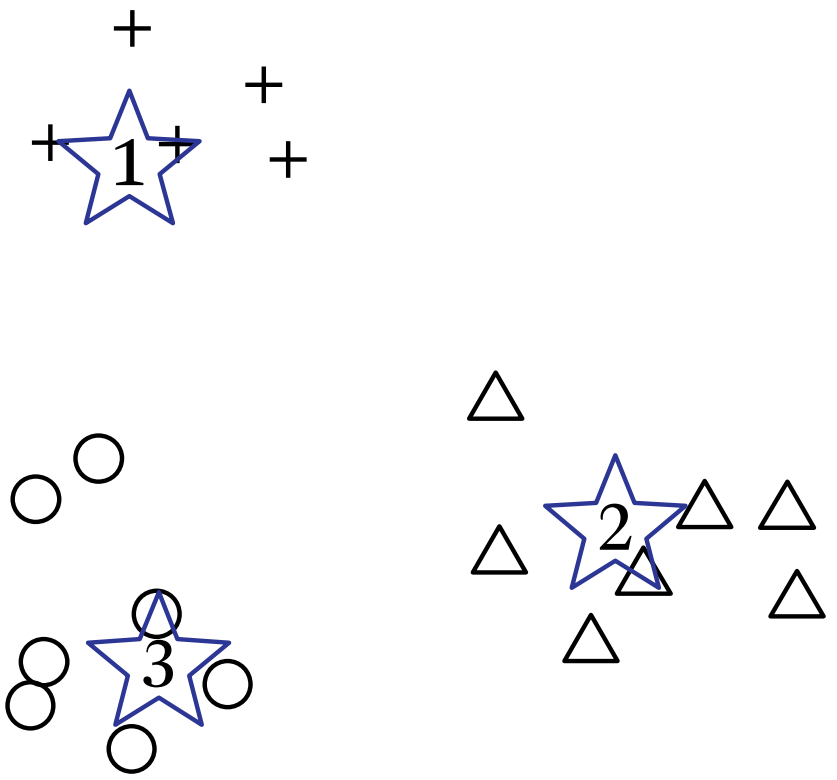
UNIVERSITY of DE

# An Example of K-Means



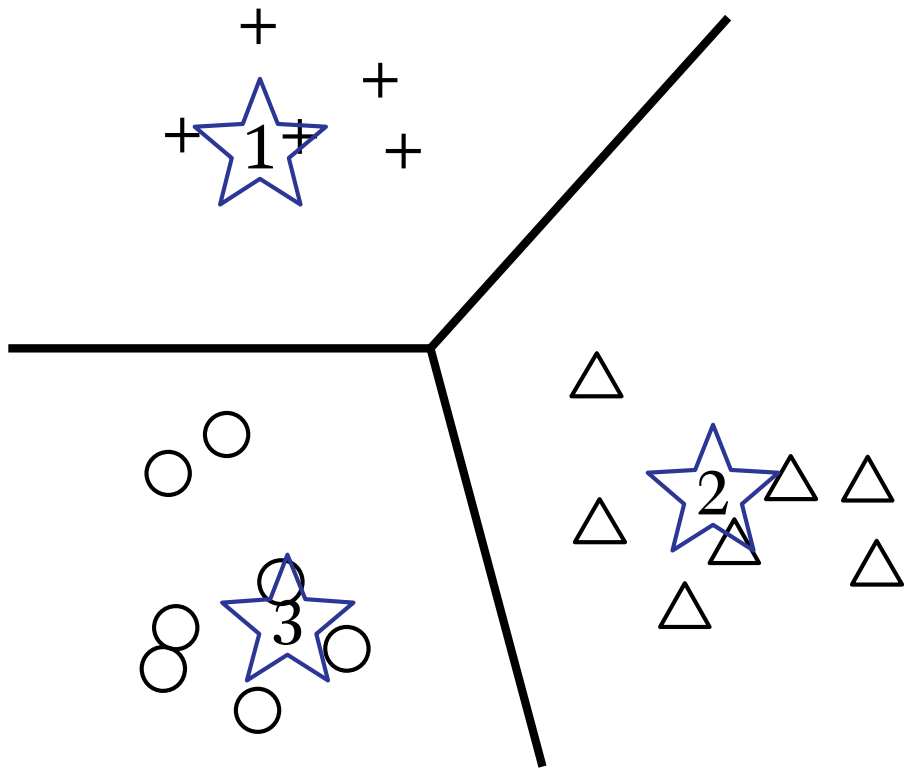
UNIVERSITY of DE

# An Example of K-Means



UNIVERSITY of DE

# An Example of K-Means



UNIVERSITY of DE

# Results can vary based on random seed selection.