Document Clustering

- Clustering is to divide objects (documents) into groups.
- Clustering is based on the similarity between two objects (documents).
- Objects in the same cluster should be similar to each other and objects in different clusters should be dissimilar.
- Clustering is unsupervised learning and classification is supervised learning
- Many different clustering algorithms
- Many different similarity measures
Applications of Documents Clustering

- Clustering has been used in IR since 60's.
- Clustering can be used for transforming document representation
  - Documents are by class membership (topics)
  - Dimensionality reduction (a group of documents share a small set of terms.)
- Clustering for efficiency
  - Document clusters are represented as their representative documents.
  - Retrieval compares the representative documents
  - An inverted file is a form of cluster
Applications of Documents Clustering

• Clustering for effectiveness
  – closed associated documents tend to be relevant to the same queries
  – retrieval of clusters makes it possible to retrieve documents that do not have many common terms with the query

• Clustering for browsing
  – clustering is used to organize documents for browsing.
Cluster Representation

• Instance-Based
  – a cluster is represented by its instances.

• Representative (Centroid) Based
  – a cluster is represented by its center (mean)

• Description Based
  – a cluster is represented by logic descriptions
Clustering Methods

• Nonhierarchical methods
  – single pass method
  – reallocation method (k-means)

• Hierarchical methods (Graph Theorectic)
  – Single link
  – complete link
Single Pass Method

1. Form the first cluster $C_1$ with the first document $D_1$
2. For $D_i$, calculate the similarity $S$ with all existing clusters
3. If $S_{\text{max}}$ is larger than a threshold, add $D_i$ to the most similar cluster. Otherwise, form a new cluster with $D_i$.
4. Repeat step 2 and 3 until all documents are clustered.
Nearest Neighbor vs. Centroid

• Nearest Neighbor
  the similarity between a document and a cluster is the similarity between the document and the nearest document in the cluster.

• Centroid
  the similarity between a document and a cluster is the similarity between the document and the cluster centroid.
K-Means

1. Select $k$ documents as $k$ initial clusters
2. For each remaining document $D_i$, assign it to the most similar cluster.
3. For all $k$ clusters, recalculate the cluster centroid.
4. Repeat the step 2 and 3 until there is no change in all clusters
Single Pass vs. K-Means

- Single pass is very fast.
- It is hard to select the similarity threshold and a small threshold results in a small number of clusters.
- Single pass is order dependent (depends on the order of the documents)
- K-Means is slower than the single pass method, but it is fast enough for a large set of documents.
- The number of clusters is predefined.
Graph Theorectic Methods

• Generate clusters from a graph
• Documents are nodes in the graph
• Two nodes (documents) are connected if their similarity is larger than a given threshold.
• Single link method
  – each connected component of the graph is a cluster
• Complete link method
  – each complete subgraph of the graph is a graph.
Graph Theoretic Clustering Algorithm

1. Form a cluster for each document
2. Find two most similar clusters $C_i$ and $C_j$
3. Merge $C_i$ and $C_j$ to form a new cluster
4. Repeat Step 2 and 3 until the stopping condition is satisfied.

Stopping Conditions

1. a given number of clusters left
2. all the similarities of all clusters are smaller than a given threshold
Similarity Measures of Clusters

- Sim($C_i$, $C_j$) = 1 - Dis($C_i$, $C_j$)
- Single Link
\[ \text{Dis}(C_i, C_j) = \text{MIN}_{x \in C_i, y \in C_j} ||x - y|| \]
- Complete Link
\[ \text{Dis}(C_i, C_j) = \text{MAX}_{x \in C_i, y \in C_j} ||x - y|| \]
- Average Link
\[ \text{Dis}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x \in C_i} \sum_{y \in C_j} ||x - y|| \]
Comparison

- Single link produces long straggly strings that are not good clusters

- Complete link produces good clusters, but not many of them (a lot of singletons)

- Average link produces the best overall results for many applications
Criterion for Clustering

- Criterion measures the clustering quality.

- Maximize the similarity between objects in the same cluster and minimize the similarity between objects from different clusters.

- The Sum-of-square-error criterion

\[
\sum_{i=1}^{c} \sum_{x \in C_i} \|x - m_i\|
\]