# Reading Summary IR: Ch 16

Xin Qian
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
xinq@cs.cmu.edu

## 1. CH 16

Clustering algorithms group documents into clusters. It is the most common form of unsupervised learning while classification is a form of supervised learning. Distance measure is an important input that can influence the outcome of clustering. *Flat clustering* creates a flat-structured set of clusters. *Hierarchical clustering* creates a hierarchical set of clusters. *Hard clustering* has a hard assignement while |*soft clustering* has a soft assignment. Cluster hypothesis assumes contingency - documents in the same cluster share similar behavior w.r.t. relevance to information needs. Search result clustering, Scatter-Gather, Collection clustering, clustering to improve search results, language model clustering are five applications of clustering. Clustering also improves search efficiency.

Hard flat clustering aims at an assignment that minimizes the objective function. K-means clustering minimizes the average distance between documents and their centroids. We usually use topic similarity or high values on the same dimensions in the vector space model for document similarity. *Partitional clustering* has each document belongs to only one cluster. *Non-exhaustive clustering* has some documents assigned to no cluster. *Cardinality* of a clustering is the number of clusters. Most flat clustering algorithms starts at an initial partitioning point. It is important to choose a favorable initial point.

The internal criterion to evaluate a clustering is to check its high intra-cluster similarity and low inter-cluster similarity. Purity, Normalized mutualinformation, Rand index and F measure are four external criteria of clustering quality. Purity is the number of correct assignment divided by N. Normalized Mutual Information measures the normalized amount of knowledge gained about what the classes are given what the clusters are. Rand Index measures the percentage of correct decision on assigning two documents either into different clusters or the same cluster. F measure penalize false negatives more than false positives by a parameter, to emphasis on recall.

K-means has an objective to minimize the average squared Euclidean distance from documents to its centroid. Residual sum of squares measures how representative the centroids are to the members of their clusters. Minimizing RSS is done iteratively through two steps: reassigning documents to the cluster dominated by a closet centroid and updating centroid based on the current member of its cluster. Termination conditions can be in many forms. However, K-means does not guaranteed to converge to a global minimum. It would be terrible to choose an outlier as an initial seed, resulting in a singleton cluster. We need effective heuristics for seed selection: excluding outliers; trying out multiple starting points or obtaining seeds from hierarchical clustering. K-means is linear towards number of iterations(I), number of clusters(K), number of vectors(N) and dimensionality of the vector space(M). K-medoids is a variant of K-means that defines the medoid of a cluster as the document vector that is closest to the centroid. Since document vector are sparse, distance computations are more efficient. We can determine the cluster cardinality, the value of K, through either a heuristic method, find a point where successive decreases in $RSS_{min}$ become drastically smaller, or a penalty-based approach with distortion and model complexity. A justification for the second approach is the Akaike Information Criterion.

Model-based clustering incorporates our knowledge about a domain and assumes that the data was generated by a model. We choose a cluster that maximizes the likelihood of generating a given set of documents. One common algorithm in this kind is the Expectation-Maximization algorithm. EM has an expectation step and a maximization step (which computes the prior and lexical parameters). It is more critical to find good seeds for EM than for K-means. Sometimes we use a hard K-means clustering for an initial assignment, and apply the EM algorithm to find a "softer" assignment.


*11741'F16 Pittsburgh, Pennsylvania USA*

DOI: 10.1145/1235