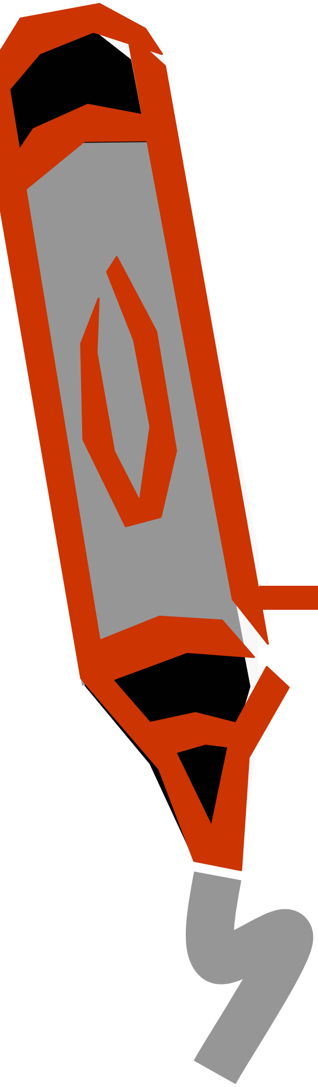


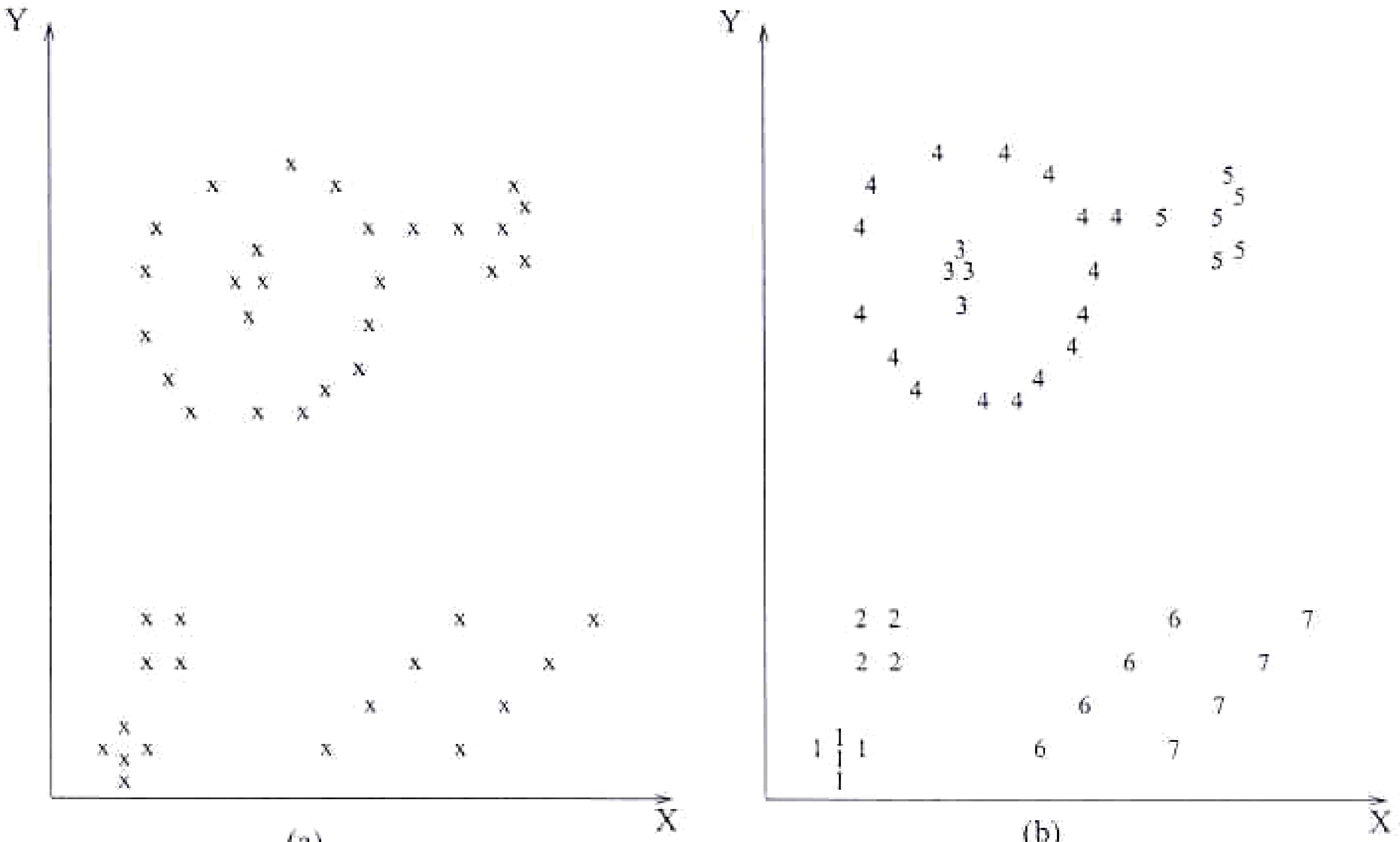
clustering

some slides courtesy James Allan@umass

clustering

- Clustering algorithms are used to group similar objects
  - in IR, typically *documents* and *terms*
- Many different applications for these algorithms
  - studying archaeological sites, classifying plants, pattern recognition
- Many different algorithms and approaches
  - e.g. graph theoretic, nearest means
- Clustering typically based on pair-wise comparisons of objects using a similarity measure
  - e.g. comparing documents, comparing cluster means and documents
- Many possible similarity measures
  - both general and domain-specific

clustering example



(a)

(b)

clustering algorithms

- Relation between properties (terms, features) and classes (clusters)
  - monothetic
  - polythetic

- Relation between objects and classes
  - exclusive
  - overlapping

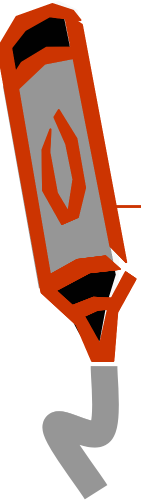
- Relation between classes and classes
  - ordered (hierarchic)
  - flat (simple partition)

cluster types

- Monothetic
    - “ A class is ordinarily defined by reference to a set of properties which are necessary and sufficient for membership in the class.”
    - Aristotelian definition of a class

- Polythetic
    - Define a class in terms of a set G of properties f_1, f_2, ...f_n such that
    - each object in the class possesses a large (but unspecified) number of the properties in G
    - each f in G is possessed by large numbers of these objects
    - no f in G must be possessed by every individual in the class (though in practice it might be)
    - http://149.170.199.144/multivar/hc.htm
    - Examples?

- Focus in IR has been on algorithms that automatically produce polythetic classifications



measures of association

- “Similarity”, “Association”, “Distance”, “Dissimilarity”
  - – pairwise measure
  - – similarity increases as the number or proportion of shared properties increase
  - – typically normalized between 0 and 1
  - – S(X,X) = 1
  - – usually S(X,Y) = S(Y,X) (symmetric)

- Many possibilities
  - – most are normalized versions of simple matching coefficient —X∩Y— or inner product for weighted terms
  - – Dice’s coefficient 2|X∩Y|/(|X|+|Y|)
  - – Jaccard’s coefficient |X∩Y|/|X□Y|
  - – Cosine correlation
  - – EMIM in some applications

- In general, no “best” measure

- Domain-specific features can be important
  - – e.g. *time* in a newspaper database, *symptoms* for patient records, ...

unsupervised learning

- no training data
- for clustering, classes are unknown
- learning underlying structure
- in practice, prior knowledge helps significantly

clustering as representation

- Clustering can be used to transform representations
  - – documents are represented by class membership as well as individual terms

- Can be viewed as dimensionality reduction
  - – especially term clustering
  - – cf. word variant clusters
  - - LSI, Factor Analysis are similar techniques

clustering for efficiency

- Clustering was initially studied by Salton as an efficiency device
  - cluster documents, represent clusters by mean or average document, compare query to cluster representatives
  - faster than sequential search
  - not as fast as optimized inverted file
  - an inverted list is also a form of cluster

- Clustering on disk is still an efficiency device in database systems
  - Though disks that do their own layout partially defeat that...

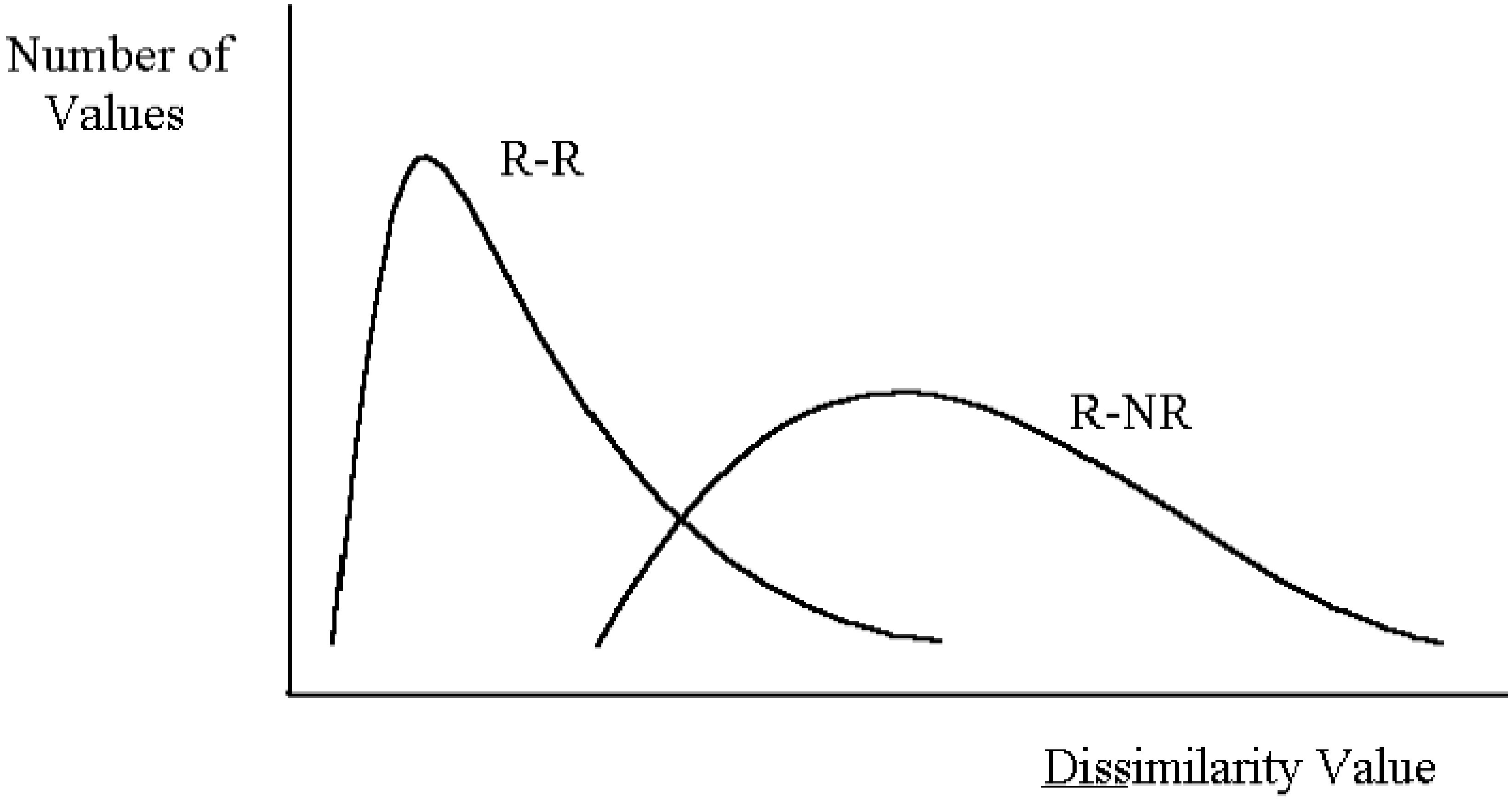
clustering for effectiveness

- By transforming representation, clustering may also result in more *effective* retrieval

- Retrieval of clusters makes it possible to retrieve documents that may not have many terms in common with the query

- “Cluster Hypothesis” proposed that closely associated documents tend to be relevant to the same queries
  - validated for many collections
  - also invalid for many collections
  - can be viewed as a test of the representation (how?)
  - generally holds in *retrieved* set of documents

cluster hypothesis



criteria for clustering methods

1. The method produces a clustering which is unlikely to be altered drastically when further objects are incorporated
   - stable under growth

2. The method is stable in the sense that small errors in the description of objects lead to small changes in the clustering

3. The method is independent of the initial ordering of the objects

- Methods which satisfy these criteria may not, forother reasons, be the best for a particular application

document clustering

- Graph Theoretic
  - – defines clusters based on a graph where documents are nodes and edges exist if similarity greater than some threshold
  - – require at least O(n^2) computations
  - – naturally hierarchic (agglomerative)
  - – good formal properties
  - – reflect structure of data

- Based on relationships to cluster representatives or means
  - – define criteria for separability of cluster representatives or closeness to representatives
  - – typically have some measure of "goodness" of cluster
  - – require only O(n log$n$) or even O(n) computations
  - – tend to impose structure (e.g. number of clusters)
  - – can have undesirable properties (e.g. order dependence)
  - – usually produce partitions (i.e., no overlapping clusters)

graph theoretic clustering

- *single link* clusters *(Connected component)*

  subgraph such that each node is connected to at least one other node in the subgraph and the set of nodes is maximal with respect to that property

- *complete link* clusters(*Maximal complete subgraph)*

  subgraph such that each node is connected to every other node in the subgraph (clique)

- *average link* clusters

  each cluster member has a greater average similarity to the remaining members of the cluster than it does to all members of any other cluster

graph theoretic clustering

- Single-link is provably the only method that satisfies criteria of adequacy

- In practice, single-link produces “long, straggly strings” that are not good clusters
  - only a single-link required to connect

- Complete link produces good clusters, but too few of them
  - many singletons

- For both searching and browsing applications, *average-link clustering* has been shown to produce the best overall effectiveness
  - efficient algorithms are possible, a little more difficult than singlelink

fast partition methods

- Single Pass
  - Assign the document D_1 as the representative (centroid, mean) for C_1
  - For D_i, calculate the similarity S with the representative for each existing cluster
  - If S_{max} is greater than threshold value S_t, add the document to the corresponding cluster and recalculate the cluster representative; otherwise use D_i to initiate a new cluster
  - If a document D_i remains to be clustered, repeat

- Variations of this often used in TDT for cluster detection and new event detection
  - TDT = topic detection and tracking

fast partition methods

- K-means or reallocation methods
  - Select K cluster representatives
  - For i = 1 to N, assign D_i to the most similar centroid
  - For j = 1 to K, recalculate the cluster centroid C_j
  - Repeat the above steps until there is little or no change in cluster membership

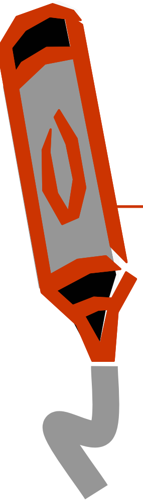
- How should K representatives be chosen?

  -furthest first traversal

- Numerous variations on this basic method
  - cluster splitting and merging strategies
  - criteria for cluster coherence
  - seed selection

Nearest Neighbor clusters

- Cluster each document with its k nearest neighbors
- Produces overlapping clusters
- Called “star” clusters by Sparck Jones
- Can be used to produce hierarchic clusters
- cf. “documents like this” in web search



cluster searching

- *Top-down* searching: start at top of cluster hierarchy, choose one or more of the best matching clusters to expand at the next level
  - tends to get lost

- *Bottom-up* searching: create inverted file of “lowest level” clusters and rank them
  - more effective
  - indicates that highest similarity clusters (such as nearest neighbor) are the most useful for searching

cluster searching

- After clusters are retrieved in order, documents in those clusters are ranked

- Cluster search produces similar level of effectiveness to document search, finds different relevant documents

- Cluster search can be modeled as a Bayesian classification problem with multiple categories
  - rank clusters by $P(C_i|Q)$

human clustering

- Is there a clustering that people will agree on?

- Is clustering something that people do consistently?

- Yahoo suggests there’s value in creating categories
  - Fixed hierarchy that people like

- What about unsupervised clustering (no fixed categories)?

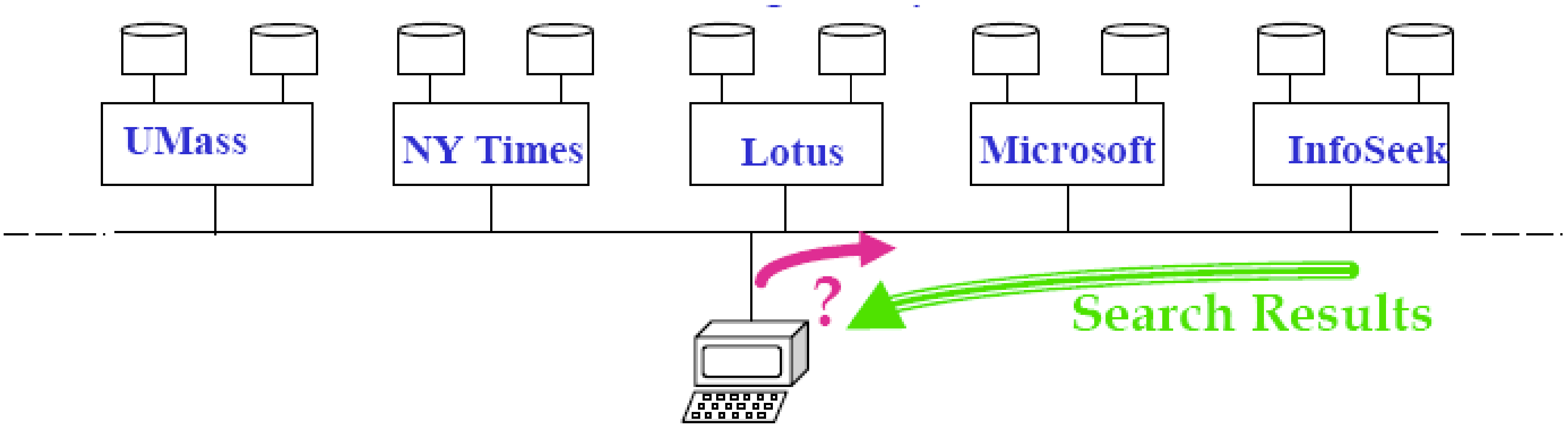
- “Human performance on clustering Web pages”
  - Macskassy, Banerjee, Davison, and Hirsh (Rutgers)
  - KDD 1998, and extended technical report

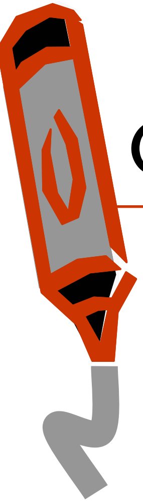
clustering for distributed retrieval

- Will look later at distributed information retrieval
- How are documents assigned to collections?
  - Administrative, legacy, etc
  - By topic if have control
- Collection selection process is “topic selection”
- How to find topics if they’re not provided?
- Document clustering is obvious possibility
  - “Discover” the topics

distributed IR

- IR is usually viewed as searching a single collection of documents
- What is a collection?
  - A single source, e.g., Wall Street Journal? (What time period?)
  - A single location, e.g., the UMass Physical Sciences Library?
  - A set of libraries, e.g., all UMass Amherst libraries?
- Distributed IR: searching when there is more than one collection
  - Local environments, e.g., a large collection is partitioned
  - Wide-area environments, e.g., corporate network, Internet





clustering for distributed retrieval

- Single system (baseline: everything together)
- Heterogeneous collections (typical in distributed IR)
  - Build a model for each existing collection (pre-existing "clusters")
- Global clustering
  - Put everything in a single repository and cluster it
  - Build a model for each cluster
  - Assumes have access to all collections
- Local clustering
  - Cluster each of the heterogeneous collections
  - Maintain models for each cluster within each collection
    - Models point to a cluster within a collection
- Multi-topic representation
  - Cluster each of the heterogeneous collections
  - Gather models for each collection together
    - Models point to a collection