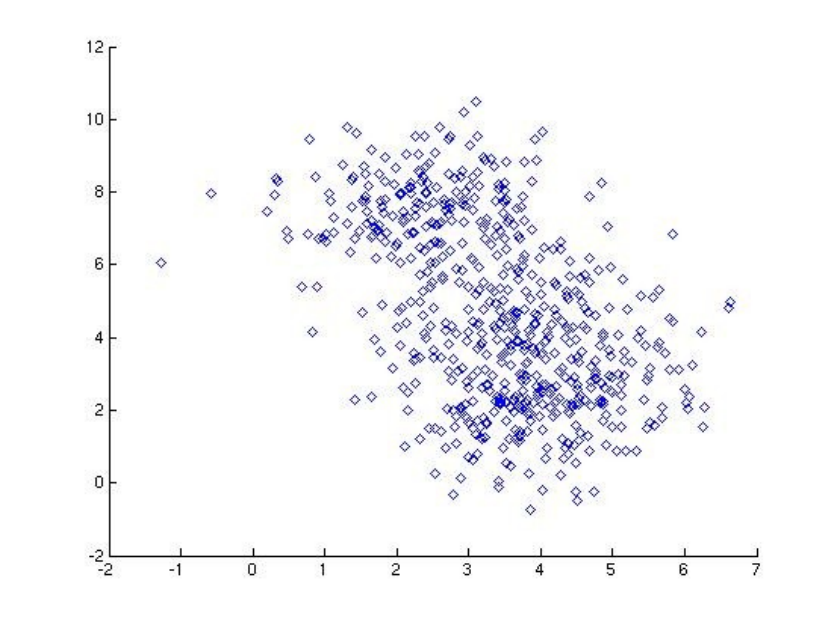
Clustering for IR

- What is clustering?
  - Grouping of objects...
- Why do you need clustering?
  - Information organization
  - Smoothing, Compression, Dimensionality Reduct.
  - Improved retrieval (the cluster hypothesis)
- How do you do clustering?
  - K-means, EM, hierarchical clustering, ...
- How do you know if clustering is successful
  - Look at the clusters
  - Use various formal measures

What is clustering?

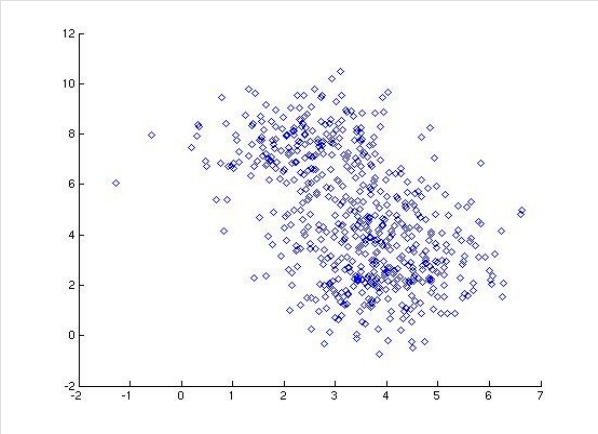
- Partitioning objects into groups such that
  - Objects in the same group are “similar”
  - Objects in different groups are “dissimilar”
  - What about outliers?

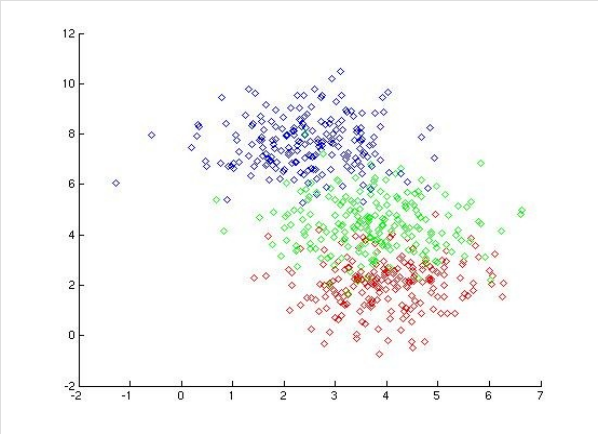


Clustering is about
finding structure
in the data

What is clustering?

- Partitioning objects into groups such that
  - Objects in the same group are “similar”
  - Objects in different groups are “dissimilar”
  - What about outliers?





- Compare to classification.
  - Clustering: labels are unknown.
  - Classification: labels are known

What is clustering?

- Before doing clustering need to decide on
  - How do I represent objects
  - What does similarity/dissimilarity mean
  - How many clusters do I want (trade-off between within cluster similarity and inter-cluster dissimilarity)
    - If each object is its own cluster then ...
    - If all objects are in one cluster ...
    - I want something in between ...
  - Once the representation and the similarity function (or the distance) is fixed, the clustering algorithm can be regarded as a black-box

Example 1: clustering words, the representation

- What are my objects: words
- How do I represent words. For example:
  - Word w is a vector [c_{w1},$c_{w2,}$, ... , c_{wn}] from the term-document matrix, c_{wi} = #times word w occurs in document i
  - Need to apply tf-idf modification
  - Dimensionality reduction through LSI speeds up implementation and may improve quality
- Typically needs to experiment with different possibilities. Analogous to the problem: “what features do I use for classification”

Clustering words. The similarity

- How do I express the similarity between words. For example:
  - cosine
- How do I represent similarity between clusters. For example:
  - Each cluster is a set of words.
  - Each word is a vector
  - Each cluster is the centroid of the word vectors
  - Similarity between two clusters is the cosine between the cluster centroids

Clustering words. Results

- SPENDING TRILLION DEFICIT SHORTFALL DEFICITS BUDGET DARMAN BALANCED BLUEPRINT EARMARKED PROPOSES OUTLAYS FISCAL EXPENDITURES GRAMM RUDMAN
- REGAN SECRETARIES AIDES AIDE OATH SWEARING NOMINATION WARS SDI MIDGETMAN MX WARHEAD CHENEY ADDRESS CASON PEDDLING RESIGNING MONDALE RETIRING RETIRE SUCCEEDING CAMPAIGNS PRESIDENCY ATWATER NIXON LIAISO ...
- CLASS BERKELEY SUPERINTENDENT CAFETERIA TEACHER CLASSMATES KINDERGARTEN SCHOLARSHIP GRADUATED BACHELOR GRADUATING TAUGHT DOCTORATE DAME GRADUATION GRADUATE PRESBYTERIAN TEACHERS GRADES SENIORS DROPOUT EDUCATIONAL EDUCATORS SCHOOLS ...
- ANGELES LOS PITTSBURGH TORONTO PHILADELPHIA LOUIS MONTREAL BOSTON

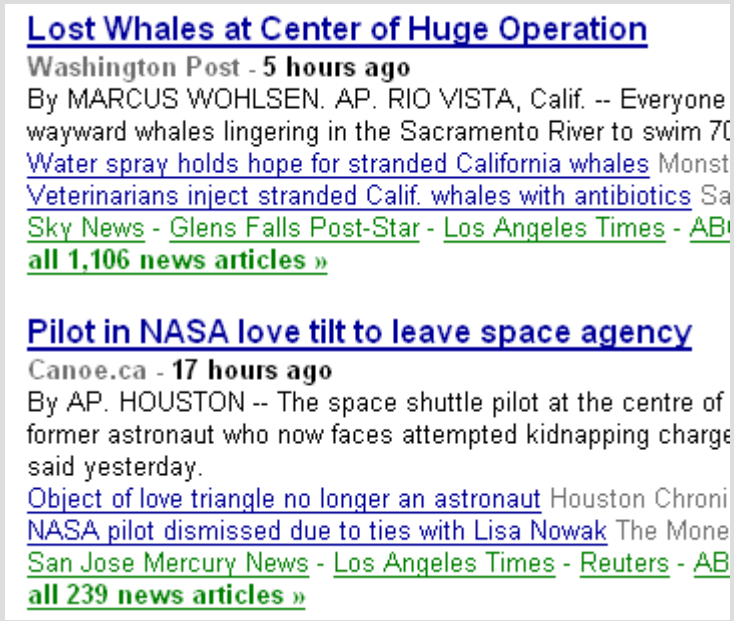
Those results were produced on the Associated Press corpus. LSI representation of words after tf-idf normalization and an agglomerative clustering algorithm were used.

Clustering documents: more difficult than clustering words

- *Similar documents do not use same words*
  - *For example:*
    - *LOTTO 17 20 22 24 25 28*
    - *PLAY 4 7 7 5 6*
    - *WEDNESDAY MEGABUCKS 02 09 23 27 29 36*
    - *BIG 4 LOTTERY 2 8 7 6*
  - ***Solution: apply smoothing***

(Those documents are from real data: Associated Press Corpus)

- *Long document talk about different things*
  - *Break documents into pieces (topics)*
  - *Example: a news page*



Lost Whales at Center of Huge Operation
Washington Post - **5 hours ago**
By MARCUS WOHLSEN. AP. RIO VISTA, Calif. -- Everyone
wayward whales lingering in the Sacramento River to swim 7
Water spray holds hope for stranded California whales Monst
Veterinarians inject stranded Calif. whales with antibiotics Sa
Sky News - Glens Falls Post-Star - Los Angeles Times - AB
all 1,106 news articles »

Pilot in NASA love tilt to leave space agency
Canoe.ca - **17 hours ago**
By AP. HOUSTON -- The space shuttle pilot at the centre of
former astronaut who now faces attempted kidnapping charge
said yesterday.
Object of love triangle no longer an astronaut Houston Chroni
NASA pilot dismissed due to ties with Lisa Nowak The Mone
San Jose Mercury News - Los Angeles Times - Reuters - AB
all 239 news articles »

Example 2: Clustering documents

- How do I represent documents?
  - A document *d* is a probability distribution over words $(p_{d1}, p_{d2}, \ldots, p_{dm})$ where
  - p_{di} = relative freq. of word *i* in document *d*
  - $JS(p, q) \approx \sum w_i(p_i - q_i)^2$ = how close are two distributions p and q
  - $w_i = 1 / p_i + 1 / q_i$: weight
  - called Jensen-Shannon divergence (version given is an approximation (no logs))
  - For explanation: need Information Theory

Document smoothing

- Need more than the usual smoothing for retrieval
- $p_{translation}(w|d) = \sum_k p(w \mid c_k)\, p(c_k \mid d)$
  [called translation prob of word w in document d]
- $p(c_k \mid d)$ = (#terms from cluster k in doc. d) / $[\sum_i p(c_i \mid d)]$
- $P_{final}(w|d) = (1 - \beta)\, p_{translation}(w|d) + \beta\, p(w|d)$
- Instead of clusters c_k can use words or phrases

- If I use $[p(c_1 \mid d), p(c_2 \mid d), \ldots, p(c_K \mid d)]$ instead of $(p_{d1}, p_{d2}, \ldots, p_{dm})$, I achieve dimensionality reduction

Clustering documents. Result*

- TWO MEN ATTEMPTED TO FIRE INTO A CROWD WITH AN UZI SUBMACHINE GUN AND A SEMIAUTOMATIC PISTOL BUT THEIR GUNS JAMMED POLICE SAID . IF THEIR GUNS HADN T JAMMED WE D OF HAD A BLOODBATH A REAL MASSACRE CHIEF INSPECTOR JOHN GRIFFIN SAID FRIDAY . ARTHUR LITTLE 19 AND DANIEL HAMPTON 21 BOTH OF NEW YORK CITY WERE ARRESTED AND CHARGED WITH CRIMINAL ATTEMPT AT MURDER CONSPIRACY TO COMMIT MURDER POSSESSION OF NARCOTICS WITH INTENT ...
- A TRACTOR TRAILER LOADED WITH STEEL COILS REAR ENDED A DISABLED TOURIST BUS ON A ROADWAY SATURDAY INJURING 18 PEOPLE POLICE SAID . THE BUS KNOWN AS THE LOLLY TROLLEY AND DESIGNED TO RESEMBLE A TROLLEY WAS USED TO CARRY VISITORS ON SIGHT SEEING TOURS . IT HAD STOPPED BECAUSE OF MECHANICAL PROBLEMS IN ONE OF THE LANES OF ...
- A TRACTOR TRAILER LOADED WITH STEEL COILS CRASHED INTO A DISABLED TOURIST BUS IN CLEVELAND SATURDAY INJURING 22 PEOPLE . IN IOWA A CAR RAMMED A SCHOOL BUS CARRYING BAND STUDENTS IN AN ACCIDENT THAT INJURED 12 POLICE SAID . NO FATALITIES WERE REPORTED IN EITHER ACCIDENT POLICE SAID .
- A GRAIN ELEVATOR EXPLOSION IN THIS SOUTHWESTERN MINNESOTA TOWN FRIDAY KILLED TWO PEOPLE AND INJURED THREE OTHERS AUTHORITIES SAID . THE BODIES OF TWO EMPLOYEES OF CARGILL INC . WHICH OWNS THE ELEVATOR WERE RECOVERED ABOUT TWO HOURS AFTER THE 2 30 P .M . EXPLOSION SAID TO A DISPATCHER FOR THE COTTONWOOD COUNTY SHERIFF S DEPARTMENT .

*this cluster is from an experiment on Associated press data

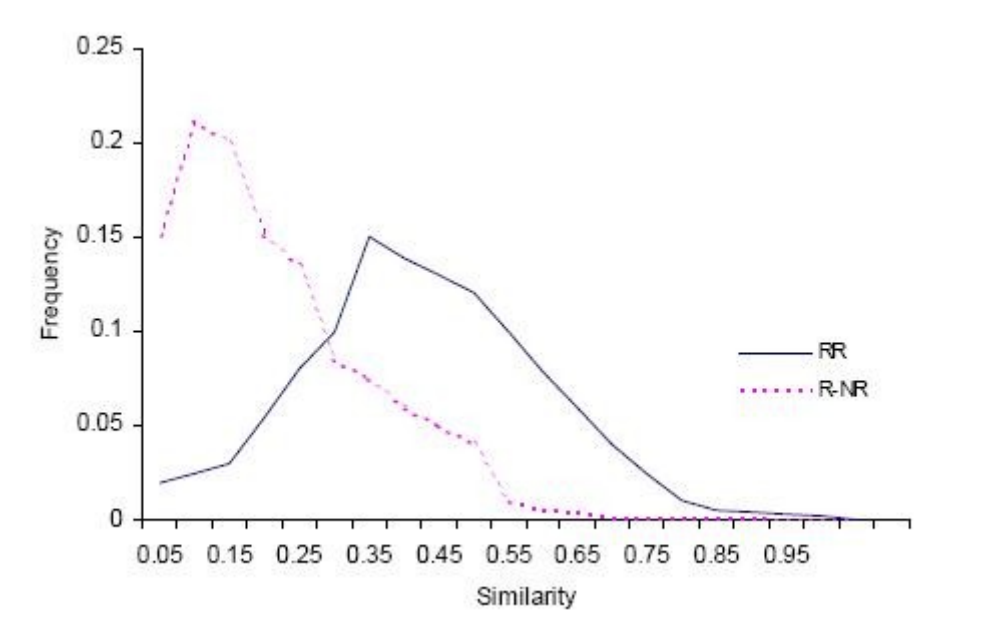
Why do you need clustering?

- Information organization

- Smoothing, Compression
  - IR: query,documents
  - Speech Recognition
  - Machine translation
  - Image Compression
  - ...
- Cluster hypothesis for IR
  - Relevant documents are very similar while non-relevant are dissimilar
  - Use pairwise similarities of retrieved documents to improve retrieval results

The cluster hypothesis for IR

- Observation:
  - Relevant documents: similar
  - Non-relevant documents: dissimilar
- Goal: use observation to improve retrieval
- How? The overlap test (Van Rijsbergen, 1971)



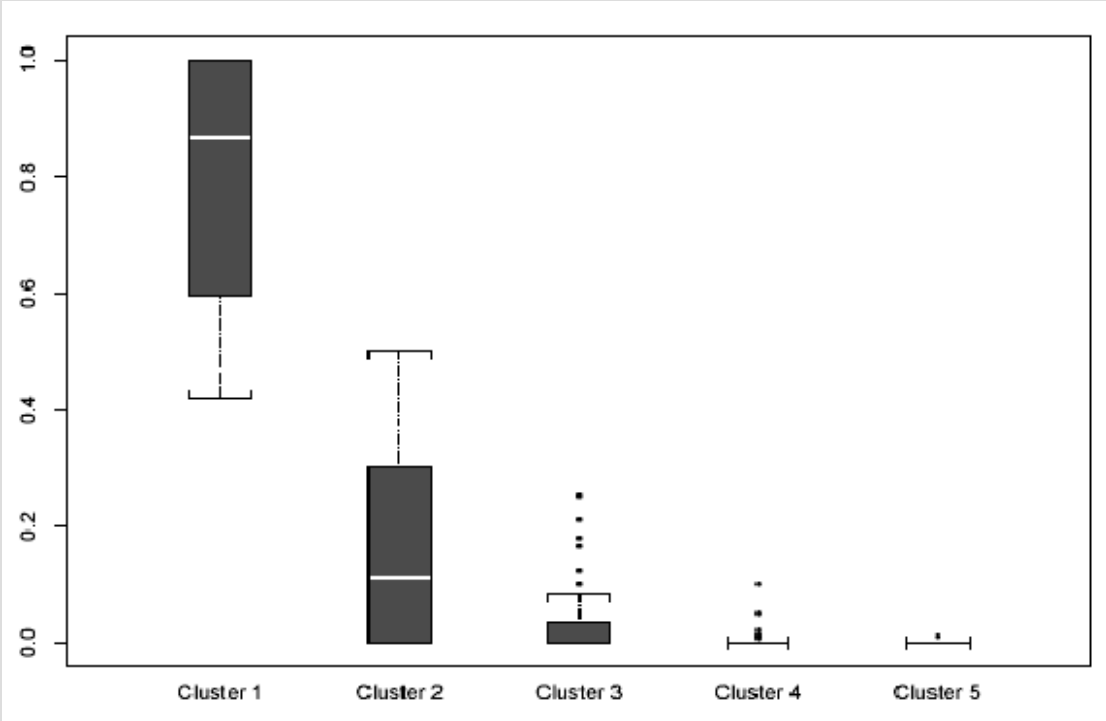
- Look at pairs of documents from a test collection of queries
- The pairs can be classified as
  - Relevant-Relevant (RR)
  - Relevant-Nonrel (R-NR)
  - Nonrel – Nonrel (same dist. as R-NR)
- Plot the distributions RR, R-NR
- In retrieval need similarity between pairs of documents; guess either RR or R-NR

Clustering can improve retrieval

- Input query, retrieve list
- Cluster documents
- Rank clusters according to #rel docs in query
- (Alternative: rank clusters according to similarity to the query)
- Within clusters, rank documents according to similarity to query
- Conclusion: clustering improves retrieval

(Source: http://nlp.stanford.edu/IR-book/ppt/lecture17-hclust.pdf, experiment by Hearst and Pedersen)

Relevance density of clusters



Clustering Algorithms

- Flat clustering
  - K-means, EM
- Hierarchical clustering
  - Top-down (divisive)
  - Bottom-up (agglomerative)
    - Single-link
    - Complete-link
    - Average-link
    - Centroid

K-means

- Initialization: choose k objects as initial cluster centers. This step is critical.
  - Number of clusters *k* must be given by user
  - Heuristic 1: farthest-first traversal
  - Heuristic 2: use agglomerative clustering first
- Repeat until clusters do not change
  - Re-assign each object to the cluster with closest center
  - Re-estimate cluster centers
- Variations: online vs. batch; K-means, K-medoids, K-median
- Beware: K-means is good choice if “natural clusters” have circular shape and roughly the same size; outliers cause problems too

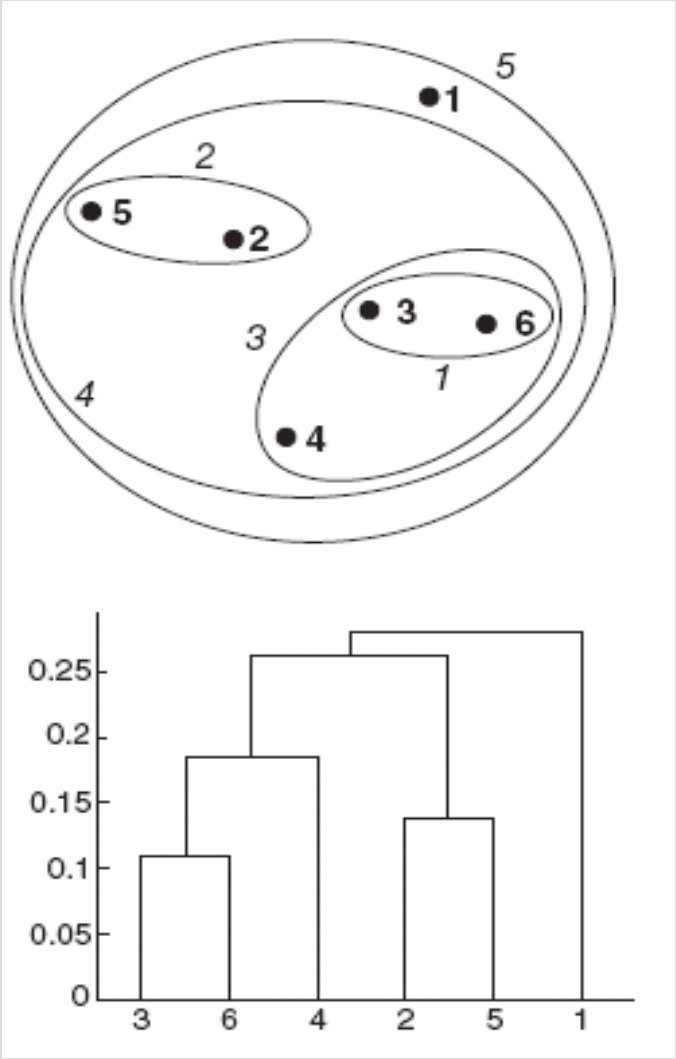
Expectation-maximization (EM) algorithm

- This is a probabilistic version of K-means
  - Each object can belong to different clusters with different probabilities. No hard assignments.
  - Replace: “Re-assign each object ...” by “compute the probability that an object belongs to a cluster” [expectation step]
  - Replace: “re-estimate cluster center” by “re-estimate the parameters of each cluster, typically the mean and the variance” [maximization step]

- The re-estimation formulas depend on the specific probabilistic model
- Can be shown that this procedure maximizes the likelihood of the data; only local optimum guaranteed

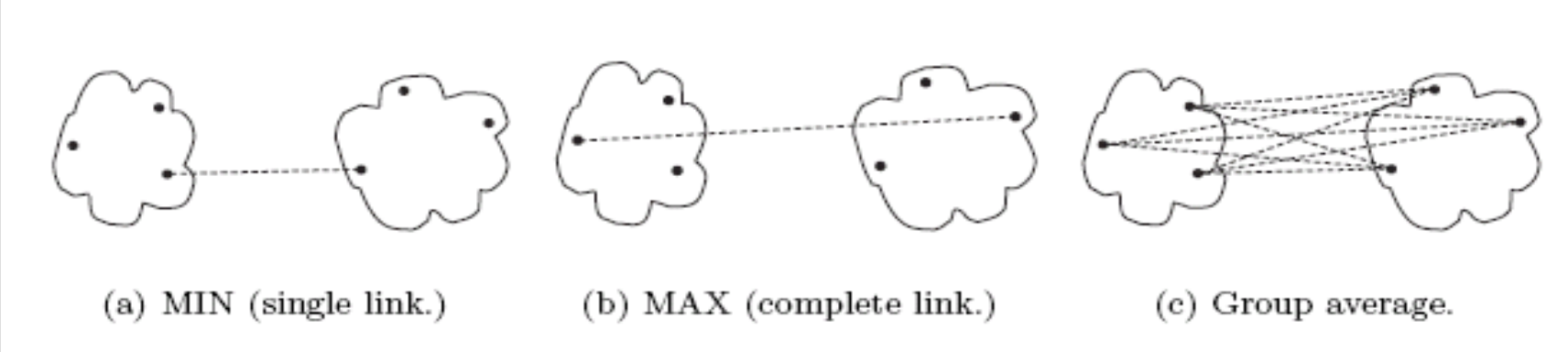
Agglomerative clustering

- Initially each object forms its own cluster
- Pick the two closest clusters and merge them into one (question: distance between clusters; many variants)
- Repeat the above two steps until left with only one cluster
- Results in a binary tree (hierarchy). Can produce partitioning afterwards.
- Note: If an optimization criterion is given for (for example Ward's method) only local maximum... but usually no optimization criterion is given



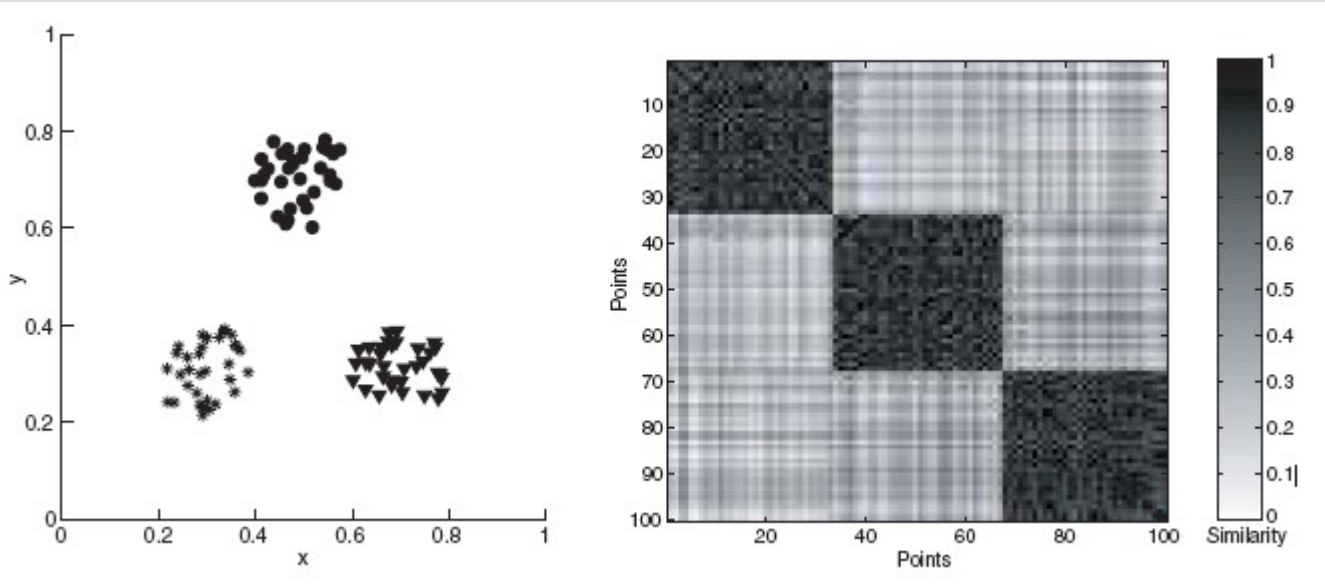
Agglomerative clustering

- **Distance between clusters:**
  - **Single–link: minimum distance between any two points from from the two different clusters. Finds connected components.**
  - **Complete – link: maximum distance between any two points from the two different clusters. Finds cliques.**
  - **Average – link: average of pairwise distances**
  - **Centroid method: distance between the centroids**
  - **Ward's method: distance = increase in squared error if two clusters are merged**
- **Star-clustering – related algorithm, instead of finding cliques find stars (an m-star is a graph with a center connected to m satellites)**



How do you measure the quality of clustering?

- Inspection
  - Look at the produced clusters? Do they make sense.
  - Look at the similarity matrix (typically clusters will be in high dimensional space, so you cannot plot them; but you can plot the pairwise similarities)



Quality of clustering

- Unsupervised measures
  - cohesion(C_i): tightness of cluster i
  - separation(C_i, C_j): distance between clusters i and j
  - validity: linear combination of cohesion and separation
  - trade-off between cohesion and separation
- Supervised measures: require each object to be assigned a category by human judge. Look at clustering as a decision problem.
  - If each cluster predicts its most frequent class, what is the error rate: called purity. There is a related measure called the entropy of clustering.
  - Look at all possible pairs of objects. Report error if two objects from same class are in different clusters or two obj. from different classes are in same cluster. Called the RAND index.
- Application Specific Measures. For example, does my clustering approach improve average precision?

Some theory of clustering...

- So far: clustering algorithms without mentioning any optimality properties
- Theoretical approach: specify optimality criterion and derive clustering alg. from it
  - K-means criterion: minimize sum of squares of distances of points to corresponding cluster. Local maximum only.
  - Farthest-first traversal: minimize the maximum radius. Constant approximation only.
  - EM-algorithm: maximize the likelihood of the data. Local maximum.
  - Divisive clustering: the produced hierarchy has the property that for any given value of k = number of clusters, the clusters are within a constant factor from the optimal. Optimality criterion is to minimize the maximum radius.

What is the best clustering approach?

- There is no perfect clustering algorithm
  - Farthest first traversal is within a constant factor of optimal, but minimization of maximum radius may not be what you need. It forces clusters to have the same radius. Outliers can be a problem too. May need preprocessing.
  - Agglomerative clustering works well in practice but may produce not produce a balanced tree
  - So, one can always use those methods as initialization and then run another algorithm, whose optimization criterion is suitable for the application (for example EM)
  - May need to do post-processing of clusters
- Check that your distance function works. Features, data transformation?
- Watch for smoothing issues.
- Beware: Usually a lot of tuning/hacks required... May try cluster algorithms on synthetic data first to become aware of pathological cases...

We discussed...

- What is clustering?
  - ...
- Why do you need clustering?
  - ...
  - ...
  - ...
- How do you do clustering?
  - ..., ..., ..., ...
- How do you know if clustering is successful
  - ...
  - ...

Additional questions

- How do I choose the number of clusters?
- What's the connection between clustering and compression?
- What's the connection between clustering and classification?

References (if you want to read more)

- Some images are take from
- http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf (Chapter on clustering from a recent Data mining book)
- http://www-csli.stanford.edu/~schuetze/information-retrieval-book.html (Chapters 16 and 17 from a recent Information Retrieval book)
- http://www.dcs.qmul.ac.uk/~tassos/publications/ATombros_PhD.pdf (Phd. thesis on how to use the cluster hypothesis to improve retrieval)