Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
Άνοιξη 2009

HY463 - Συστήματα Ανάκτησης Πληροφοριών
Information Retrieval (IR) Systems

Ομαδοποίηση Εγγράφων (Document Clustering)

Γιάννης Τζίτζικας

Διάλεξη :
Ημερομηνία :

Clustering (ομαδοποίηση)

- ***Clustering*** is the process of grouping similar objects into naturally associated subclasses.
- This process results in a set of "clusters" which somehow describe the underlying objects at a more abstract or approximate level.
- The process of clustering is typically based on a "similarity measure" which allows the objects to be classified into separate natural groupings.
- A *cluster* is then simply a collection of objects that are grouped together because they collectively have a strong internal similarity based on such a measure.
- A *similarity measure* (or *dissimilarity measure*) quantifies the conceptual distance between two objects, that is, how alike or disalike a pair of objects are.
  - Determining exactly what type of similarity measure to use is typically a domain dependent problem.

Clustering

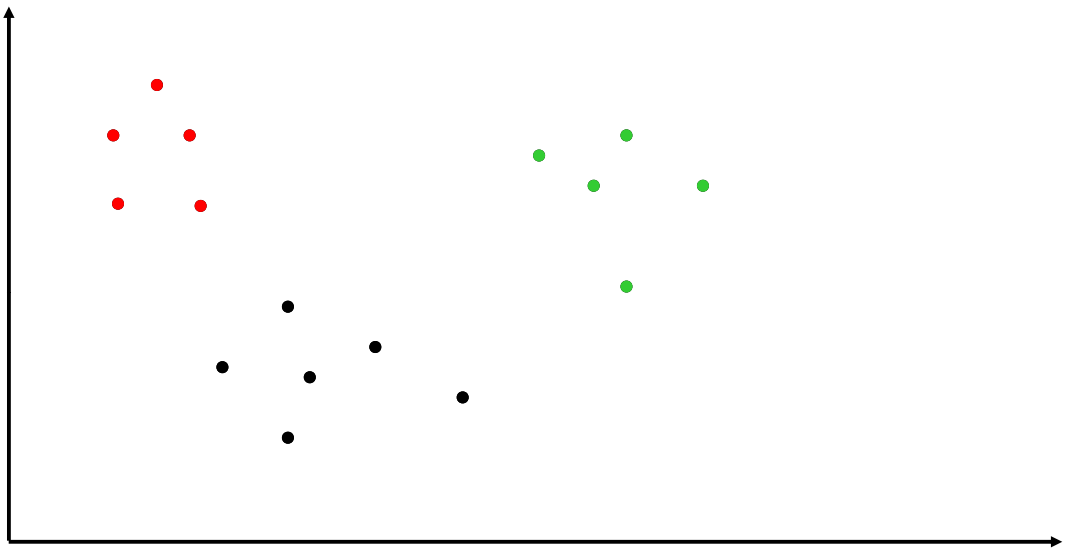
A clustering of a set N is a partition of N, i.e. a set $C_1,\ldots, C_k$ of subsets of N, such that:

$$C_1 \cup \ldots \cup C_k = N \quad \text{and} \quad C_i \cap C_j = \varnothing, \text{ for all } i \neq j.$$

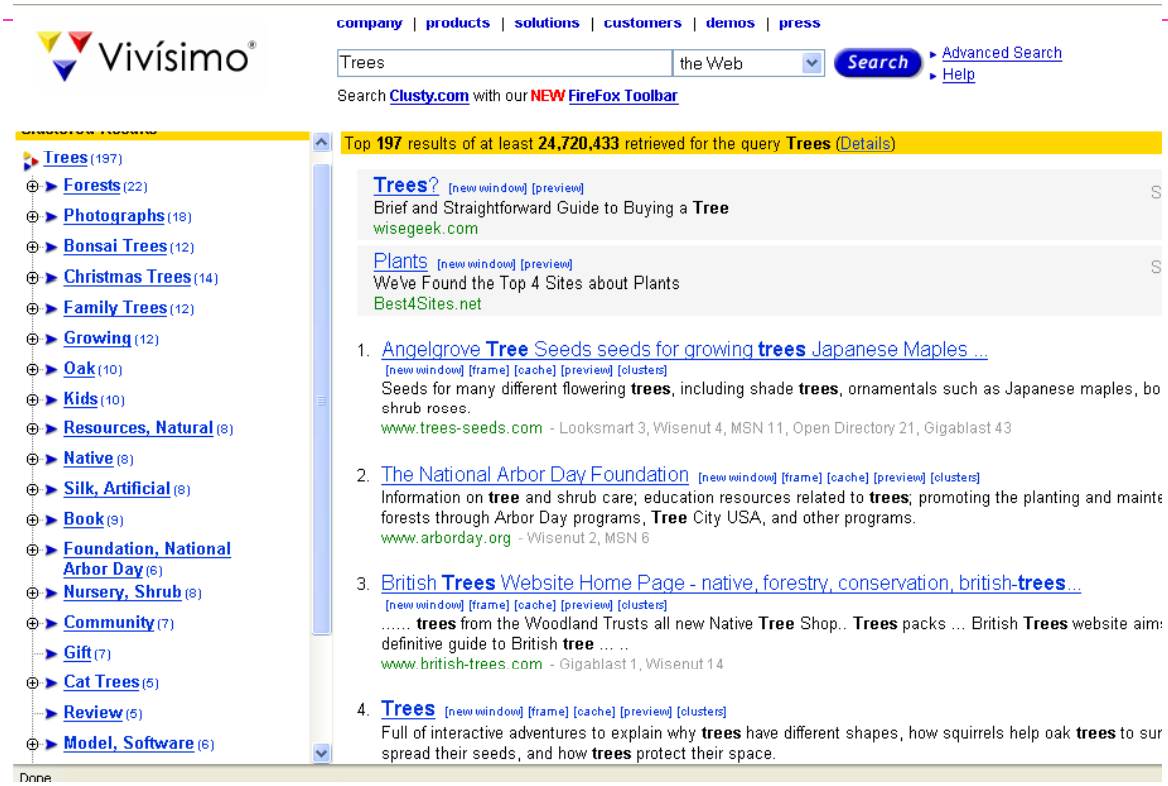
- Variation: the clusters are not disjoint. This is somethimes called *overlapping clustering.*

- Clustering is used in areas such as:
  - medicine, anthropology, economics, data mining
  - software engineering (reverse engineering, program comprehension, software maintenance)
  - information retrieval
- In general, any field of endeavor that necessitates the analysis and comprehension of large amounts of data may use clustering.

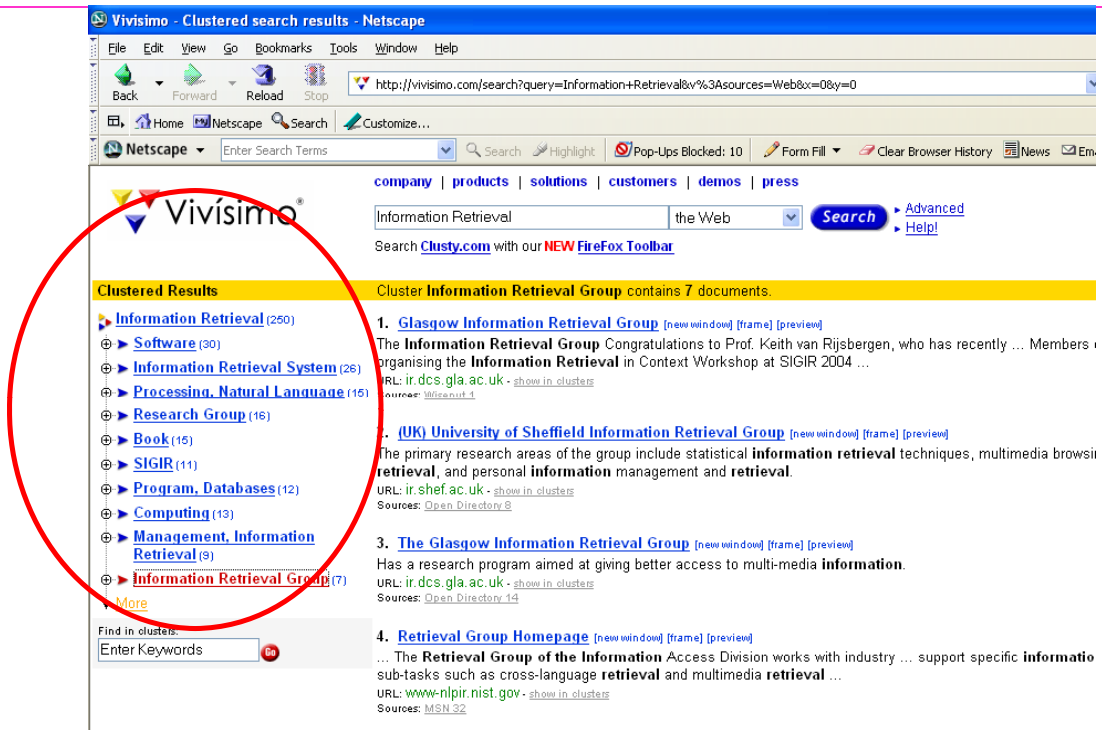
Παράδειγμα ομαδοποίησης (στο δισδιάστατο χώρο)



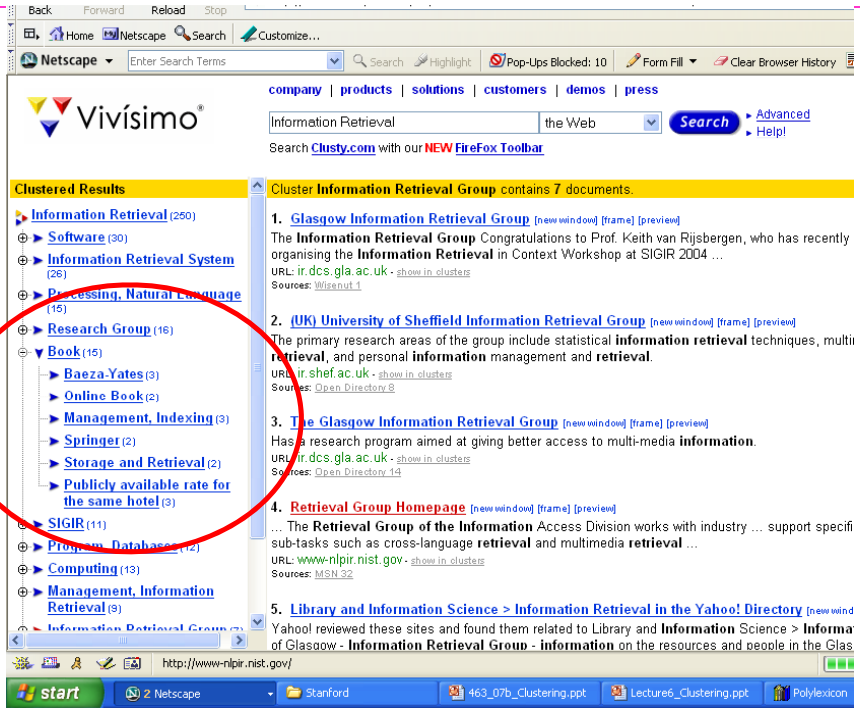
Παράδειγμα ομαδοποίησης αποτελεσμάτων www.vivisimo.com



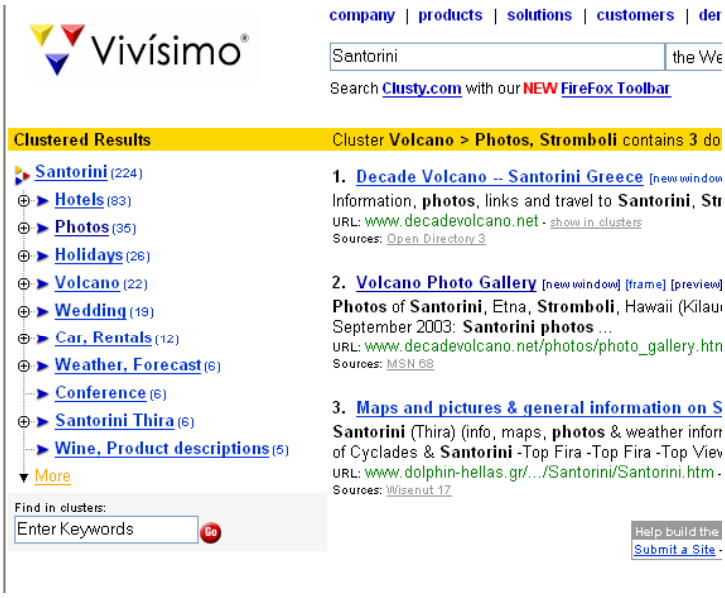
Παράδειγμα ομαδοποίησης αποτελεσμάτων www.vivisimo.com



Παράδειγμα ομαδοποίησης αποτελεσμάτων



q=Santorini



Τύποι Ομαδοποίησης

- Ανάλογα με τη σχέση μεταξύ Ιδιοτήτων και Κλάσεων
  - Monothetic clustering
  - Polythetic clustering
- Ανάλογα με τη σχέση μεταξύ Αντικειμένων και Κλάσεων
  - Αποκλειστική (exclusive) ομαδοποίηση
  - Επικαλυπτόμενη (overlapping) ομαδοποίηση
    - Ένα αντικείμενο μπορεί να ανήκει σε παραπάνω από μία κλάση
- Ανάλογα με τη σχέση μεταξύ Κλάσεων
  - Χωρίς διάταξη: οι κλάσεις δεν συνδέονται μεταξύ τους
  - Με διάταξη (ιεραρχική): υπάρχουν σχέσεις μεταξύ των κλάσεων

Monothetic vs. Polythetic

- Monothetic
  - Μια κλάση ορίζεται βάσει ενός συνόλου ικανών και αναγκαίων ιδιοτήτων που πρέπει να ικανοποιούν τα μέλη της (Αριστοτελικός ορισμός)
- Polythetic
  - Μια κλάση ορίζεται βάσει ενός συνόλου ιδιοτήτων Φ =φ1,...,φn, τ.ω.
    - Κάθε μέλος της κλάσης πρέπει να έχει ένα μεγάλο αριθμό των ιδιοτήτων Φ
    - Κάθε φ του Φ χαρακτηρίζει πολλά αντικείμενα
    - Δεν είναι αναγκαίο να υπάρχει μια φ που να ικανοποιείται από όλα τα μέλη της κλάσης
- Στην ΑΠ, έχει δοθεί έμφαση σε αλγόριθμους για αυτόματη παραγωγή polythetic classifications.

Monothetic vs. Polythetic

	A	B	C	D	E	F	G	H
1	+	+	+					
2	+	+		+				
3	+		+	+				
4		+	+	+				
5					+	+	+	
6					+	+	+	
7					+	+		+
8					+	+		+

Figure 3.1. An illustration of the difference between monothetic and polythetic

- 8 individuals (1-8) and 8 properties (A-H).
- The possession of a property is indicated by a plus sign. The individuals 1-4 constitute a polythetic group each individual possessing three out of four of the properties A,B,C,D.
- The other 4 individuals can be split into two monothetic classes {5,6} and {7,8}.

Μέτρα Συσχέτισης (Association)

- Μετρικές συναρτήσεις ομοιότητας, συσχέτισης (απόστασης):
  - It is a pairwise measure. Similarity increases as the number or proportion of shared properties increases
  - Typically normalized between 0 and 1
  - S(X,X)=1, S(X,Y)=S(Y,X)
- Παραδείγματα μετρικών ομοιότητας
  - Οι περισσότερες είναι κανονικοποιημένες εκδόσεις του $|X \cap Y|$ ή του εσωτερικού γινομένου (εάν έχουμε βεβαρημένους όρους)
  - **Dice's coefficient** $2\,|X \cap Y| / |X| + |Y|$
  - **Jaccard's coefficient** $|X \cap Y| / |X \cup Y|$
  - **Cosine correlation**
- Δεν υπάρχει το «καλύτερο» μέτρο (που να δίνει τα καλύτερα αποτελέσματα σε κάθε περίπτωση)
  - Βέβαια οι αλγόριθμοι ομαδοποίησης είναι ανεξάρτητοι από το πως ακριβώς ορίζεται το μέτρο

Παραδείγματα Μέτρων Ομοιότητας για Έγγραφα

- Dice's coefficient $2\,|X \cap Y| / (|X| + |Y|)$
- Jaccard's coefficient $|X \cap Y| / |X \cup Y|$

Μέτρα για την περίπτωση που τα βάρη δεν είναι δυαδικά:

$$\text{DiceSim}(dj, dm) = \frac{2\sum_{i=1}^{t}(w_{ij} \cdot w_{im})}{\sum_{i=1}^{t} w_{ij}^{2} + \sum_{i=1}^{t} w_{im}^{2}}$$

$$\text{JaccardSim}(dj, dm) = \frac{\sum_{i=1}^{t}(w_{ij} \cdot w_{im})}{\sum_{i=1}^{t} w_{ij}^{2} + \sum_{i=1}^{t} w_{im}^{2} - \sum_{i=1}^{t}(w_{ij} \cdot w_{im})}$$

$$\text{CosSim}(dj, dm) = \frac{\vec{d}_j \cdot \vec{d}_m}{|\vec{d}_j| \cdot |\vec{d}_m|} = \frac{\sum_{i=1}^{t}(w_{ij} \cdot w_{im})}{\sqrt{\sum_{i=1}^{t} w_{ij}^{2} \cdot \sum_{i=1}^{t} w_{im}^{2}}}$$

Ομαδοποίηση ως τρόπος Αναπαράστασης (Clustering as Representation)

- Η ομαδοποίηση είναι μια μορφή μη επιτηρούμενης μάθησης (unsupervised learning)
  - Για εκμάθηση της υποκείμενης δομής και κλάσεων
- Η ομαδοποίηση είναι μια μορφή μετασχηματισμού της αναπαράστασης (representation transformation)
  - Τα έγγραφα παριστάνονται όχι μόνο βάσει των όρων αλλά και βάσει των κλάσεων στις οποίες μετέχουν
- Η ομαδοποίηση μπορεί να θεωρηθεί ως μια τεχνική για μείωση των διαστάσεων (dimensionality reduction)
  - Ειδικά το term clustering
  - Latent Semantic Indexing, Factor Analysis είναι παρόμοιες τεχνικές

Ομαδοποίηση για βελτίωση της απόδοσης (Clustering for Efficiency)

- Μπορούμε να αξιοποιήσουμε τις τεχνικές ομαδοποίησης προκειμένου να επιταχύνουμε **την αποτίμηση επερωτήσεων**.
- Παράδειγμα:

Τρόπος:

1. Cluster all documents of the collection
   - (very time consuming but it can be done once)
2. Represent each cluster by its mean or average document
3. Whenever a new query is submitted, compare the query to each cluster representative
   - The cluster representatives are less than the documents
4. Return to the user the documents of the most similar cluster(s)

Ομαδοποίηση για βελτίωση της Αποτελεσματικότητας (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
  - E.g. LSI (Latent Semantic Indexing)

Two Document Clustering Approaches

- Graph Theoretic
  - Defines clusters based on a graph where documents are nodes and edges exist if *similarity* greater than some threshold
  - Requires at least $O(n^2)$ computation
  - 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
  - Typically have some measure of goodness of cluster
  - Require only O(n logn) 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 (no overlapping clusters)

Criteria of Adequacy for Clustering Methods

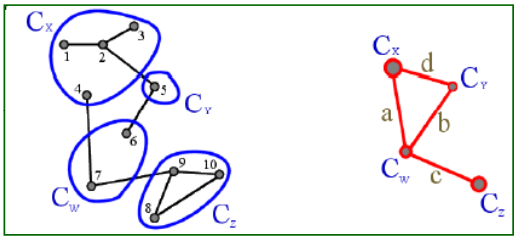
Desired properties:

- The method produces a clustering which is unlikely to be altered drastically when further objects are incorporated
  - in other words: stable under growth
- The method is stable in the sense that small errors in the description of objects lead to small changes in the clustering
- The method is independent of the initial ordering of the objects

Graph Theoretic **Clustering** Algorithms

Graph Clustering

- Graph clustering deals with the problem of clustering a graph
  - grouping similar nodes of a graph into a set of subgraphs

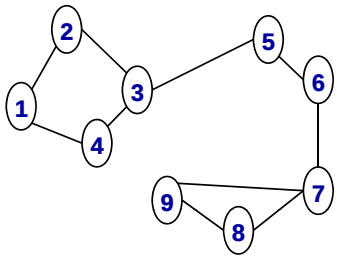


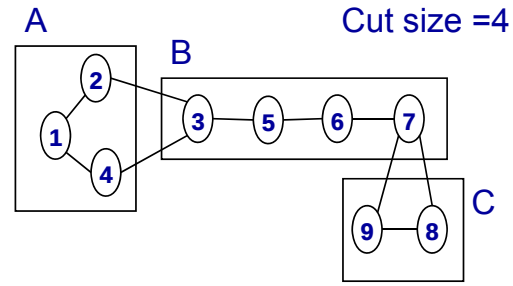
Quality criteria for graph clustering methods

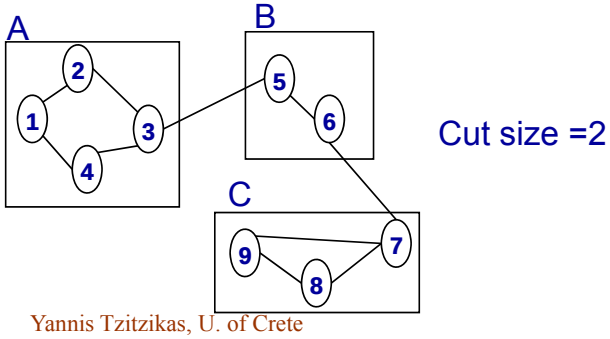
Graph clustering methods should produce clusters with high cohesion and low coupling

- high cohesion:
  - there should be many internal edges
- low "cut size":
  - The cut size (else called *external cost)* of a clustering measures how many edges are external to all sub-graphs, that is, how many edges cross cluster boundaries.
- Uniformity of cluster size is also often desirable.
  - A uniform graph clustering is where $|C_i|$ is close to $|C_j|$ for all i,j in {1..k}

Example







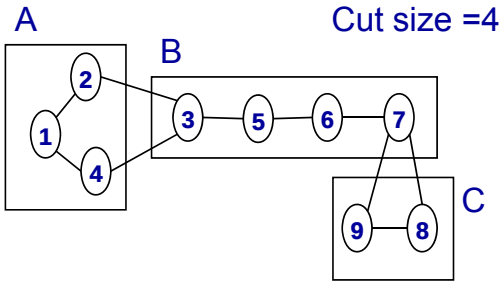
Quality Measures for Graph Clustering

- There are several. One well known is the **CC measure** (Coupling-Cohesion measure)

$$CC = \frac{|E^{in}| - |E^{ex}|}{|E|}$$

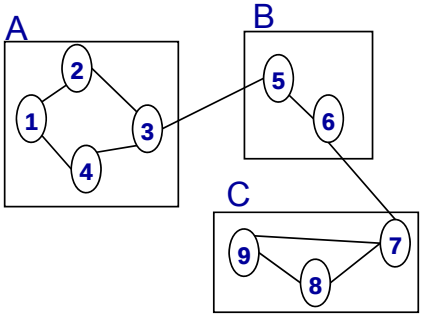
- E^{in}: the "internal" edges: those that connect nodes of the same cluster
- E^{ex}: the "external" edges: those that cross cluster boundaries
- maximum value of CC: 1
  - when all edges are internal
- minimum value of CC: -1
  - when all edges are external

Example



Cut size =4

$$CC = \frac{6-4}{10} = 0.2$$

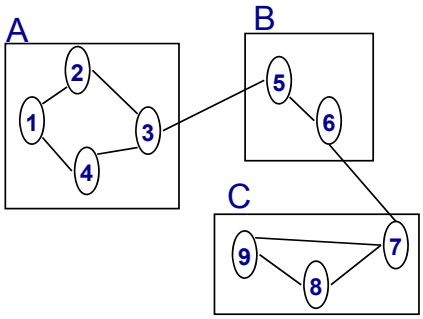


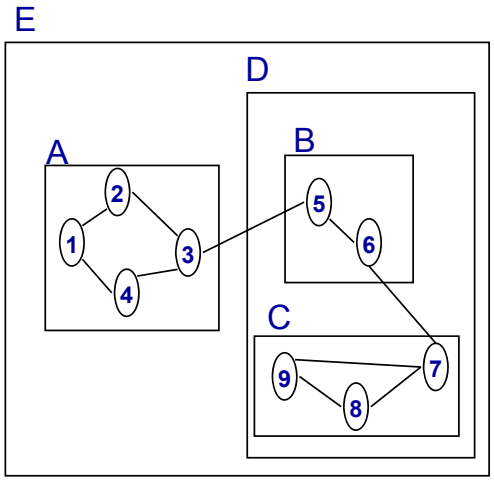
Cut size =2

$$CC = \frac{8-2}{10} = 0.6$$

Hierarchical Graph Clustering

- The clusters of the graph can be clustered themselves to form a higher level clustering, and so on.
- A hierarchical clustering is a collection of clusters where any two clusters are either disjoint **or** nested.





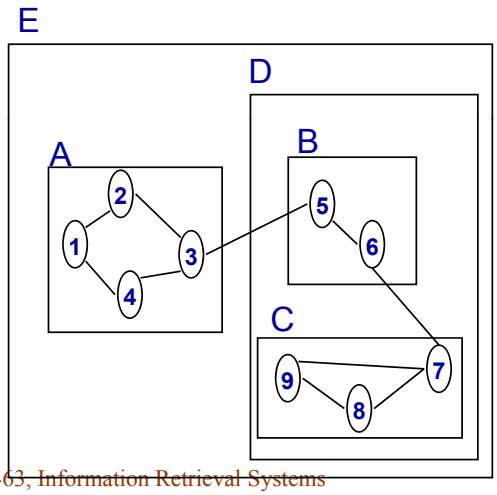
Hierarchical Clustered Graph

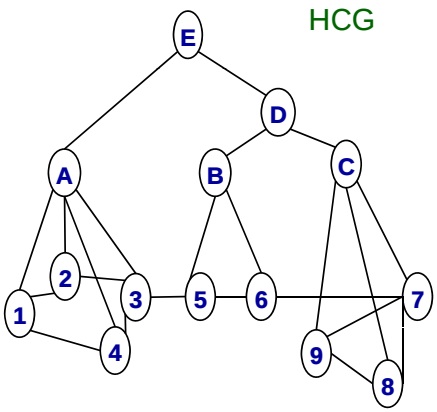
A Hierarchical Clustered Graph (HCG) is a pair (G,T) where

G is the underlying graph, and

T is a rooted tree such that the leaves of T are the nodes of G.

(the tree T represents an inclusion relationship: the leaves of T are nodes of G, the internal nodes of T represent a set of graph nodes, i.e. a cluster)



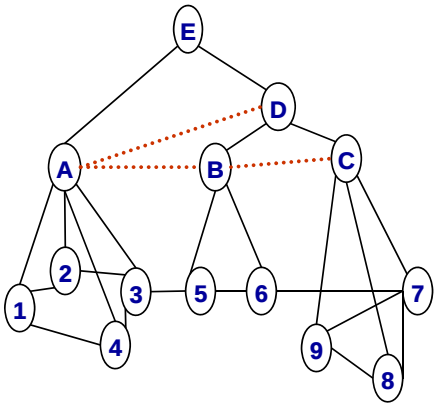


Hierarchical Clustered Graph
Implied Edges

Implied edges: edges between the internal nodes.

Two clusters are connected iff the nodes that they contain are related.

Multiple implied edges (between the same pair of clusters) can be ignored or summed up to form weighted implied edges. Thresholding can applied in order to filter out some implied edges



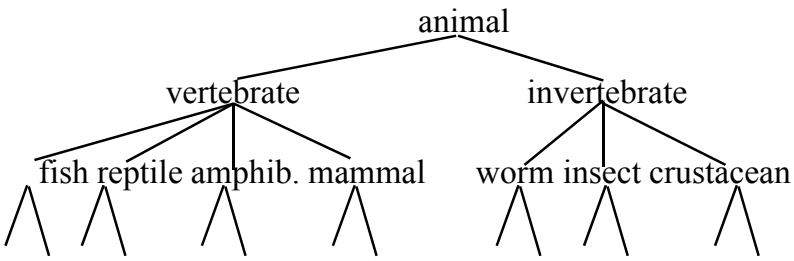
A Hierarchical Compound Graph is a triad (G,T, I) where (G,T) is a hierarchical clustered graph (HCG), and I the set of implied edges set.

Graph Theoretic Clustering **Algorithms**

- Given a graph of objects connected by links that represent similarities greater than some threshold, the following cluster definitions are straightforward:
  - **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
    - Called **single link** clusters
  - **Maximal complete subgraph**: subgraph such that each node is connected to every other node in the subgraph (clique)
    - **Complete link** clusters
- Others are possible and very common:
  - **Average link**: 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

Hierarchical Clustering

- Build a tree-based hierarchical taxonomy (*dendrogram*).
- Recursive application of a standard clustering algorithm can produce a hierarchical clustering.



Hierarchical Clustering Methods

- *Agglomerative (συσσώρευσης)* (*bottom-up*) methods start with each example in its own cluster and iteratively combine them to form larger and larger clusters.
- *Divisive (διαίρεσης)* (*partitional, top-down*) separate all examples immediately into clusters.

An hierarchical (agglomerative) clustering algorithm

1/ Βαλε κάθε έγγραφο σε ένα διαφορετικό cluster

2. Υπολόγισε την ομοιότητα μεταξύ όλων των ζευγαριών cluster

3. Βρες το ζεύγος {Cu,Cv} με την υψηλότερη (inter-cluster) ομοιότητα
4. Συγχώνευσε τα clusters Cu, Cv
5. Επανέλαβε (από το βήμα 2) έως ότου να καταλήξουμε να έχουμε 1 μόνο cluster
6. Επέστρεψε την ιεραρχία των clusters (το ιστορικό των συγχωνεύσεων)

An hierarchical (agglomerative) clustering algorithm

1/ Βαλε κάθε έγγραφο σε ένα διαφορετικό cluster

C:=∅; For i=1 to n C:=C ∪ [di]

2. Υπολόγισε την ομοιότητα μεταξύ όλων των ζευγαριών cluster

Compute **SIM**(c,c') for each c, c' ∈ C

3. Βρες το ζεύγος {Cu,Cv} με την υψηλότερη (inter-cluster) ομοιότητα
4. Συγχώνευσε τα clusters Cu, Cv
5. Επανέλαβε (από το βήμα 2) έως ότου να καταλήξουμε να έχουμε 1 μόνο cluster
6. Επέστρεψε την ιεραρχία των clusters (το ιστορικό των συγχωνεύσεων)

An hierarchical (agglomerative) clustering algorithm

1/ Βαλε κάθε έγγραφο σε ένα διαφορετικό cluster

C:=∅; For i=1 to n C:=C ∪ [di]

2. Υπολόγισε την ομοιότητα μεταξύ όλων των ζευγαριών cluster

Compute **SIM**(c,c') for each c, c' ∈ C

sim(d,d') = CosineSim(d,d') or DiceSim(d,d') or JaccardSim(d,d')

3. Βρες το ζεύγος {Cu,Cv} με την υψηλότερη (inter-cluster) ομοιότητα
4. Συγχώνευσε τα clusters Cu, Cv
5. Επανέλαβε (από το βήμα 2) έως ότου να καταλήξουμε να έχουμε 1 μόνο cluster
6. Επέστρεψε την ιεραρχία των clusters (το ιστορικό των συγχωνεύσεων)

An hierarchical (agglomerative) clustering algorithm

1/ Βαλε κάθε έγγραφο σε ένα διαφορετικό cluster

C:=∅; For i=1 to n C:=C ∪ [di]

2. Υπολόγισε την ομοιότητα μεταξύ όλων των ζευγαριών cluster

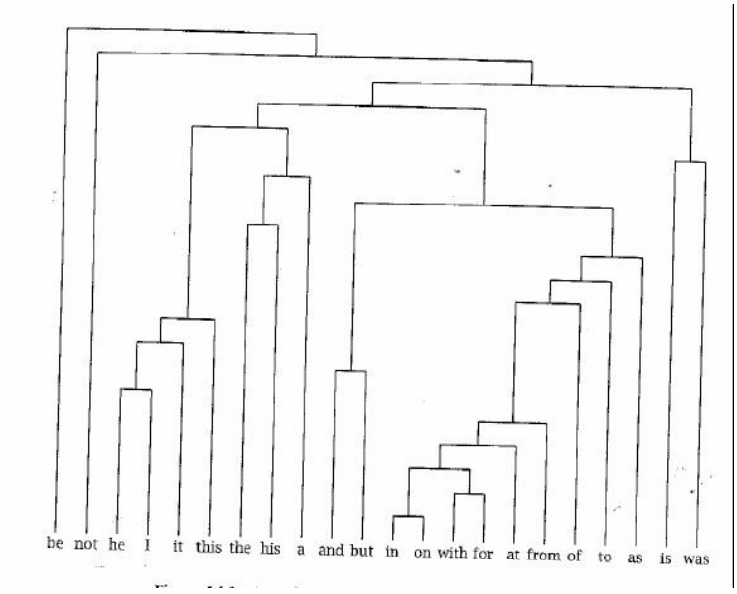
Compute **SIM**(c,c') for each c, c' ∈ C

sim(d,d') = CosineSim(d,d') or DiceSim(d,d') or JaccardSim(d,d')

SIM(c,c')= *single link*: similarity of two most similar. = max{ sim(d,d') |d∈c,d'∈c'}
complete link: similarity of two least similar. = min{ sim(d,d') |d∈c,d'∈c'}
average link: average similarity b. = avg{ sim(d,d') |d∈c,d'∈c'}

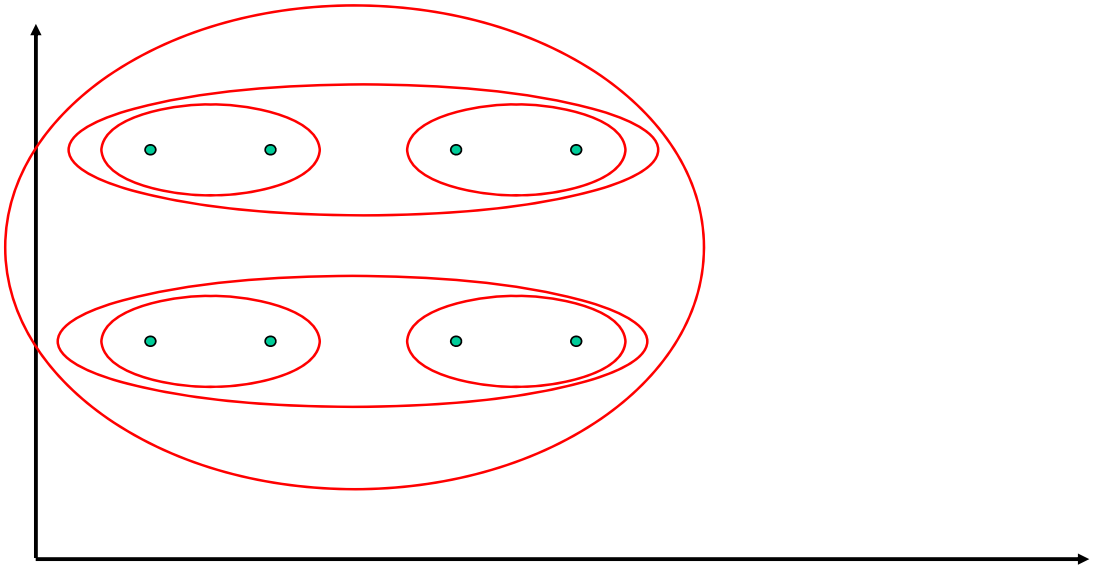
3. Βρες το ζεύγος {Cu,Cv} με την υψηλότερη (inter-cluster) ομοιότητα
4. Συγχώνευσε τα clusters Cu, Cv
5. Επανέλαβε (από το βήμα 2) έως ότου να καταλήξουμε να έχουμε 1 μόνο cluster
6. Επέστρεψε την ιεραρχία των clusters (το ιστορικό των συγχωνεύσεων)

Dendogram or Cluster Hierarchy



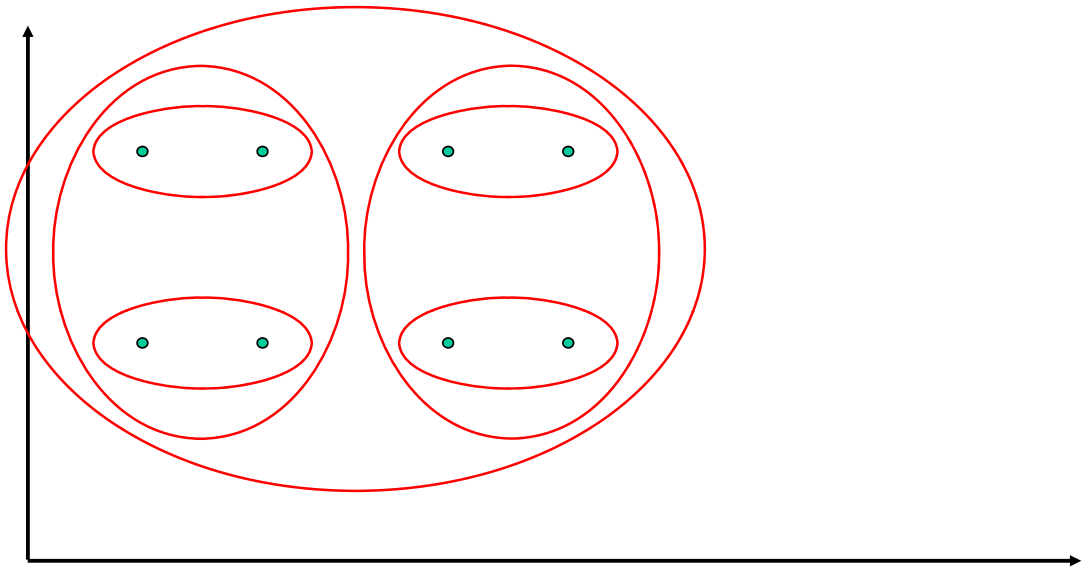
Hierarchical Graph-based Clustering Algorithms

Single-Link Example



Hierarchical Graph-based Clustering Algorithms

Complete-Link Example



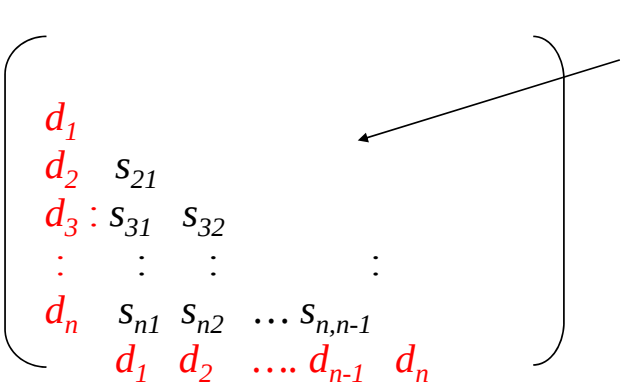
Σύγκριση

- **Single**-link
  - is provably the only method that **satisfies criteria of adequacy**
  - however it produces "**long, straggly (ανάκατα) string**" that are not good clusters
    - Only a single-link required to connect
- **Complete** link
  - produces **good clusters** (more "tight," spherical clusters), but too few of them (**many singletons**)
- **Average**-link
  - For both searching and browsing applications, average-link clustering has been shown to produce **the best overall effectiveness**

Ward's method
(an alternative to single/complete/average link)

- Cluster merging:
  - Merge the pair of clusters whose merger minimizes the increase in the total within-group error sum of squares, based on the Euclidean distance between centroids
- Remarks:
  - this method tends to create symmetric hierarchies

Computing the Document Similarity Matrix



- Optimization: Compute sim(di,dj) only if di and dj have at least one term in common (otherwise it is 0)
  - This is done by exploiting the inverted index

Clustering algorithms based on relationships to cluster representatives or means
(Fast Partition Algorithms)

Fast Partition Methods
Single Pass

It also takes as input a threshold *simThres*

Single Pass

- Assign the document d1 as the representative (**centroid,mean**) for c1
- For each di, calculate the similarity *Sim* with the representative for each existing cluster
- If SimMax is greater than threshold value *simThres*, add the document to the corresponding cluster and recalculate the cluster representative; otherwise use di to initiate a new cluster
- If a document di remains to be clustered, repeat

Fast Partition Methods
K-Means

K-means (or reallocation methods)

- 1/ Select K cluster representatives
- 2/ For i = 1 to N, assign di to the most similar centroid
- 3/ For j = 1 to K, recalculate the cluster centroid cj
- Repeat from step 2 until there is little or no change in cluster membership

- Issues:
  - How should K representatives be chosen?
  - Numerous variations on this basic method
    - cluster splitting and merging strategies
    - criteria for cluster coherence
    - seed selection

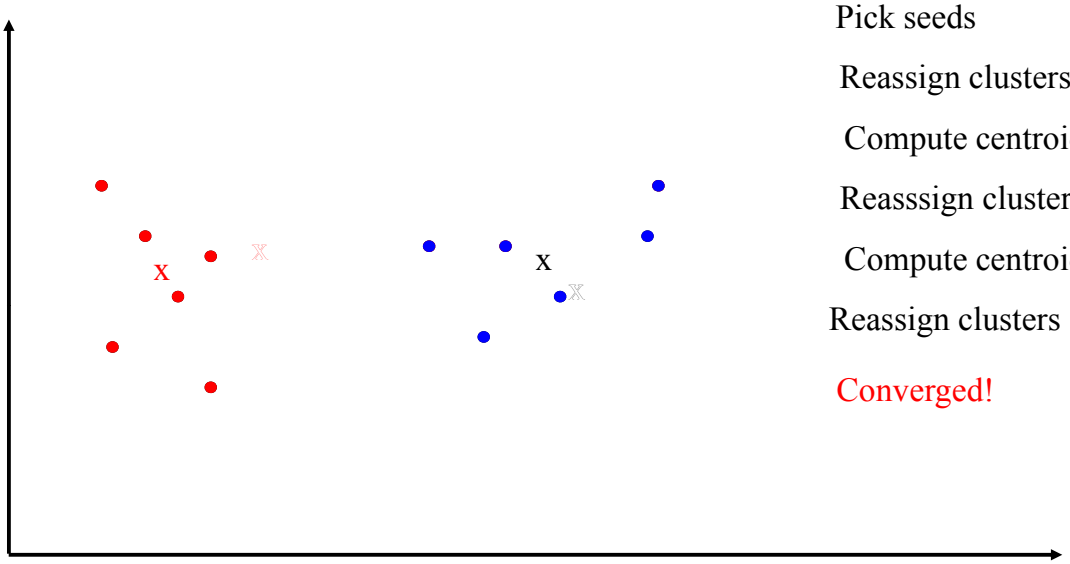
Fast Partition Methods
K-Means (cont)

- Assumes instances are real-valued vectors.
- Clusters based on *centroids*, *center of gravity*, or mean of points in a cluster, *c*:
  - For example, the centroid of (1,2,3), (4,5,6) and (7,2,6) is **(4,3,5).**

$$\vec{\mu}(c) = \frac{1}{|c|}\sum_{\vec{x}\in c}\vec{x}$$

- Reassignment of instances to clusters is based on distance to the current cluster centroids.

K Means Example (K=2)



Pick seeds

Reassign clusters

Compute centroids

Reasssign clusters

Compute centroids

Reassign clusters

Converged!

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

Complexity Remarks

- Computing the matrix with document similarities: O(n^2)
- Simple reallocation clustering method with k clusters O(kn)
  - πιο γρήγορος από τους αλγορίθμους για ιεραρχική ομαδοποίηση
- Agglomerative or Divisive Hierarchical Clustering:
  - απαιτεί n-1 συγχωνεύσεις/διαιρέσεις
  - η πολυπλοκότητα του είναι τουλάχιστον O(n^2)

Cluster Searching
Document Retrieval from a Clustered Data Set

- *Top-down* searching:
  - start at top of cluster hierarchy,choose one of more of the best matching clusters to expand at the next level
    - tends to get lost
- *Bottom-up* searching:
  - create inverted file of "lowestlevel" clusters and rank them
    - more effective
    - indicates that highest similarity clusters (such as nearest neighbor) are the most useful for 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

Some notes

- HAC and K-Means have been applied to text in a straightforward way.
- Typically use ***normalized***, TF/IDF-weighted vectors and cosine similarity.
- Optimize computations for sparse vectors.
- Applications:
  - During retrieval, **add other documents** in the same cluster as the initial retrieved documents to improve recall.
  - **Clustering of results** of retrieval to present more organized results to the user (e.g. vivisimo search engine)
  - **Automated production of hierarchical taxonomies** of documents for browsing purposes (like Yahoo & DMOZ).

Human Clustering (χειρονακτική ομαδοποίηση)

- Questions:
  - 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
- "Human performance on clustering Web pages"
  - Macskassy, Banerjee, Davison, and Hirsh (Rutgers)
  - KDD 1998, and extended technical report
- Αποτελέσματα: Μάλλον δεν υπάρχει μεγάλη συμφωνία
  - γενικά προτίμηση σε μικρά clusters
  - άλλοι χρήστες προτιμούν/δημιουργούν επικαλυπτόμενα, άλλοι αποκλειστικά clusters
  - τα περιεχόμενα των clusters διέφεραν αρκετά
  - γενική ομαδοποίηση (ανεξαρτήτου επερώτησης) δεν φαίνεται να είναι πολύ χρήσιμη