Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
Άνοιξη 2006

ΗΥ463 - Συστήματα Ανάκτησης Πληροφοριών
Information Retrieval (IR) Systems

Ομαδοποίηση Εγγράφων (Document Clustering)

Γιάννης Τζίτζικας

Διάλεξη : 12
Ημερομηνία : 12-4-2006

Clustering

- ***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.

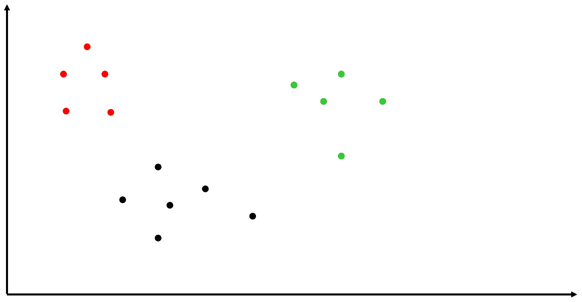
Clustering

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:

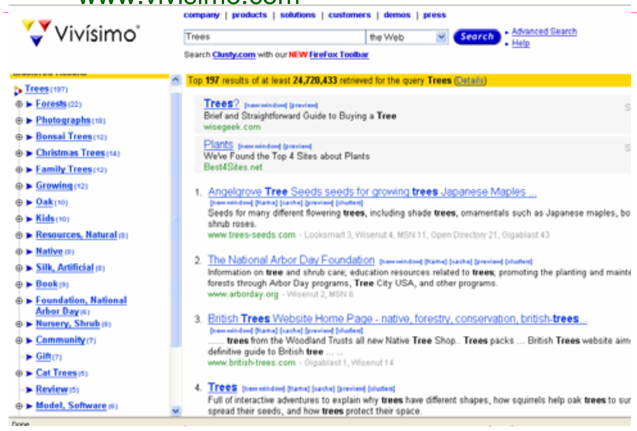
$C_1 \cup \ldots \cup C_k = N$ and $C_i \cap C_j = \emptyset$, for all $i \neq j$.

- 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.

Clustering Example



Παράδειγμα ομαδοποίησης αποτελεσμάτων www.vivisimo.com



company | products | solutions | customers | demos | press

Trees | the Web | Search
- Advanced Search
- Help

Search Clusty.com with our NEW FireFox Toolbar

Top 197 results of at least 24,720,433 retrieved for the query Trees (Details)

- Trees (197)
- Forests (22)
- Photographs (18)
- Bonsai Trees (12)
- Christmas Trees (14)
- Family Trees (12)
- Growing (12)
- Oak (10)
- Kids (10)
- Resources, Natural (8)
- Native (8)
- Silk, Artificial (8)
- Book (9)
- Foundation, National Arbor Day (6)
- Nursery, Shrub (8)
- Community (7)
- Gift (7)
- Cat Trees (5)
- Review (5)
- Model, Software (6)

Trees? [new window] [preview]
Brief and Straightforward Guide to Buying a Tree
wisegeek.com

Plants [new window] [preview]
We've Found the Top 4 Sites about Plants
Best4Sites.net

1. Angelgrove Tree Seeds seeds for growing trees Japanese Maples ...
[new window] [frame] [cache] [preview] [clusters]
Seeds for many different flowering trees, including shade trees, ornamentals such as Japanese maples, bo shrub roses.
www.trees-seeds.com - Looksmart 3, Wisenut 4, MSN 11, Open Directory 21, Gigablast 43

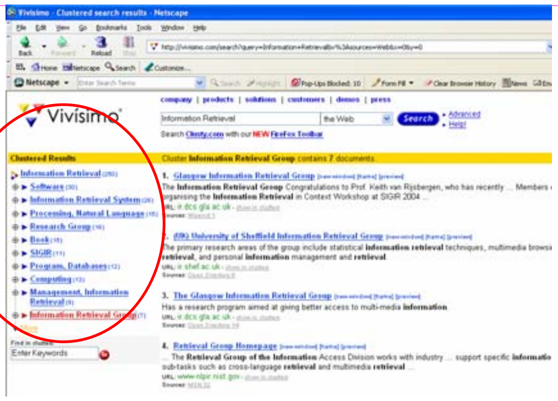
2. The National Arbor Day Foundation [new window] [frame] [cache] [preview] [clusters]
Information on tree and shrub care; education resources related to trees; promoting the planting and mainte forests through Arbor Day programs, Tree City USA, and other programs.
www.arborday.org - Wisenut 2, MSN 6

3. British Trees Website Home Page - native, forestry, conservation, british-trees ...
[new window] [frame] [cache] [preview] [clusters]
...... trees from the Woodland Trusts all new Native Tree Shop.. Trees packs ... British Trees website aim definitive guide to British tree ... ...
www.british-trees.com - Gigablast 1, Wisenut 14

4. Trees [new window] [frame] [cache] [preview] [clusters]
Full of interactive adventures to explain why trees have different shapes, how squirrels help oak trees to sur spread their seeds, and how trees protect their space.

Done

Παράδειγμα ομαδοποίησης αποτελεσμάτων www.vivisimo.com



company | products | solutions | customers | demos | press

Information Retrieval | the Web | Search
- Advanced
- Help!

Search Clusty.com with our NEW FireFox Toolbar

Clustered Results

- Information Retrieval (250)
- Software (30)
- Information Retrieval System (26)
- Processing, Natural Language (16)
- Research Group (16)
- Book (15)
- SIGIR (11)
- Program, Databases (12)
- Computing (12)
- Management, Information Retrieval (9)
- Information Retrieval Group (7)
- More

Find in clusters
Enter Keywords

Cluster Information Retrieval Group contains 7 documents

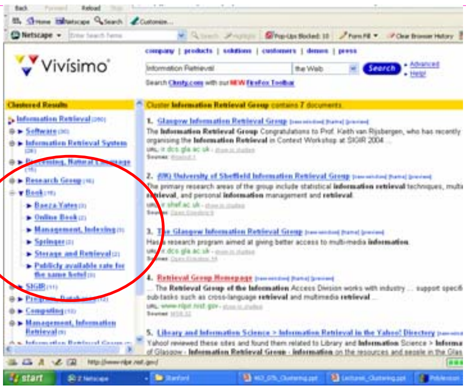
1. Glasgow Information Retrieval Group [new window] [frame] [preview]
The Information Retrieval Group Congratulations to Prof. Keith van Rijsbergen, who has recently ... Members organising the Information Retrieval in Context Workshop at SIGIR 2004 ...
URL: ir.dcs.gla.ac.uk - show in clusters
Source: Wisenut 1

2. (UK) University of Sheffield Information Retrieval Group [new window] [frame] [preview]
The primary research areas of the group include statistical information retrieval techniques, multimedia browsi retrieval, and personal information management and retrieval.
URL: ir.shef.ac.uk - show in clusters
Source: Open Directory 8

3. The Glasgow Information Retrieval Group [new window] [frame] [preview]
Has a research program aimed at giving better access to multi-media information.
URL: ir.dcs.gla.ac.uk - show in clusters
Source: Open Directory 14

4. Retrieval Group Homepage [new window] [frame] [preview]
... The Retrieval Group of the Information Access Division works with industry ... support specific informatio sub-tasks such as cross-language retrieval and multimedia retrieval ...
URL: www-nlpir.nist.gov - show in clusters
Source: MSN 32

Παράδειγμα ομαδοποίησης αποτελεσμάτων



Search Clusty.com with our NEW FireFox Toolbar

Clustered Results

- Information Retrieval (280)
  - Software (30)
  - Information Retrieval System (28)
  - Processing, Natural Language (16)
  - Research Group (16)
  - Book (15)
    - Baeza-Yates (3)
    - Online Book (2)
    - Management, Indexing (3)
    - Springer (2)
    - Storage and Retrieval (2)
    - Publicly available rate for the same hotel (3)
  - SIGIR (11)
  - Programs, Databases (12)
  - Computing (10)
  - Management, Information Retrieval (8)
  - Information Retrieval Group

Cluster **Information Retrieval Group** contains 7 documents

1. Glasgow Information Retrieval Group [new window] [frame] [preview]
The **Information Retrieval Group** Congratulations to Prof. Keith van Rijsbergen, who has recently organising the **Information Retrieval** in Context Workshop at SIGIR 2004 ...
URL: ir.dcs.gla.ac.uk - show in clusters
Sources: Wisenut 1

2. (IR) University of Sheffield Information Retrieval Group [new window] [frame] [preview]
The primary research areas of the group include statistical **information retrieval** techniques, multi ... **retrieval**, and personal **information** management and **retrieval**.
URL: ir.shef.ac.uk - show in clusters
Sources: Open Directory 5

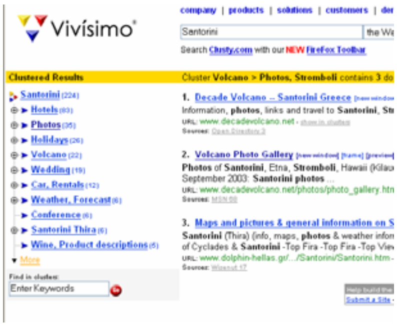
3. The Glasgow Information Retrieval Group [new window] [frame] [preview]
Has a research program aimed at giving better access to multi-media **information**.
URL: ir.dcs.gla.ac.uk - show in clusters
Sources: Open Directory 14

4. Retrieval Group Homepage [new window] [frame] [preview]
... The **Retrieval Group of the Information** Access Division works with industry ... support specif... sub-tasks such as cross-language **retrieval** and multimedia **retrieval** ...
URL: www-nlpir.nist.gov - show in clusters
Sources: MSN 32

5. Library and Information Science > Information Retrieval in the Yahoo! Directory [new window
Yahoo! reviewed these sites and found them related to Library and **Information** Science > **Informa**... of Glasgow - **Information Retrieval Group** - **information** on the resources and people in the Glas...

http://www-nlpir.nist.gov/

q=Santorini



Search Clusty.com with our NEW FireFox Toolbar

Clustered Results

- Santorini (224)
  - Hotels (83)
  - Photos (35)
  - Holidays (26)
  - Volcano (22)
  - Wedding (19)
  - Car, Rentals (12)
  - Weather, Forecast (6)
  - Conference (6)
  - Santorini Thira (6)
  - Wine, Product descriptions (5)
  - More

Find in clusters: Enter Keywords

Cluster **Volcano > Photos, Stromboli** contains 3 do

1. Decade Volcano – Santorini Greece [new window
Information, **photos**, links and travel to **Santorini**, **Str**
URL: www.decadevolcano.net - show in clusters
Sources: Open Directory 3

2. Volcano Photo Gallery [new window] [frame] [preview]
Photos of **Santorini**, Etna, **Stromboli**, Hawaii (Kilau... September 2003: **Santorini photos** ...
URL: www.decadevolcano.net/photos/photo_gallery.htm
Sources: MSN 68

3. Maps and pictures & general information on S
Santorini (Thira) (info, maps, **photos** & weather infor... of Cyclades & **Santorini** -Top Fira -Top Fira -Top View
URL: www.dolphin-hellas.gr/.../Santorini/Santorini.htm -
Sources: Wisenut 17

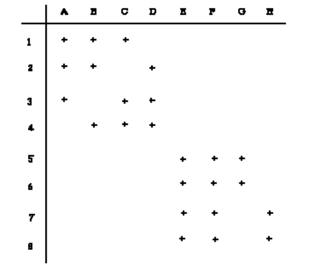
Τύποι Αλγορίθμων Ομαδοποίησης

- 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):
  - Pairwise measure
  - Similarity increases as the number or proportion of shared properties increase
  - 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 ∩Y| ή του εσωτερικού γινομένου (εάν έχουμε βεβαρημένους όρους)
  - **Dice's coefficient** 2 |X ∩Y|/ |X| +|Y|
  - **Jaccard's coefficient** |X ∩Y|/ |X ∪ Y|
  - **Cosine correlation**
- Δεν υπάρχει το «καλύτερο» μέτρο (που να δίνει τα καλύτερα αποτελέσματα σε κάθε περίπτωση)

Παραδείγματα Μέτρων για Έγγραφα

- Dice's coefficient   2 |X ∩Y|/ |X| +|Y|
- Jaccard's coefficient |X ∩Y|/ |X ∪ Y|

Μέτρα για την περίπτωση που τα βάρη δεν είναι δυαδικά:

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

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

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

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)

Method:

- 1/ Cluster documents,
- 2/ Represent clusters by mean or average document,
- 3/ **compare query to cluster representatives**

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

Document Clustering Approaches

- 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) 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)

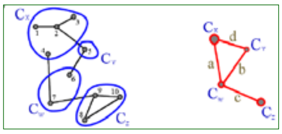
Criteria of Adequacy for Clustering Methods

- The method produces a clustering which is unlikely to be altered drastically when further objects are incorporated (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

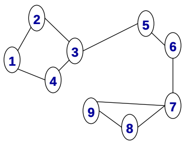


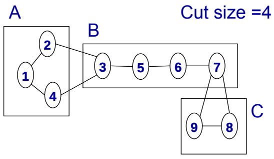
Quality criteria for graph clustering methods

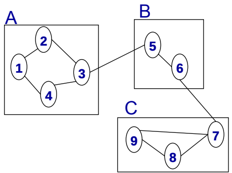
Graph clustering methods should produce clusters with high cohesion and low coupling

- 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







Cut size =2

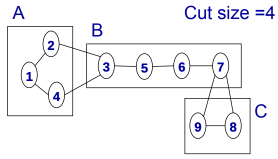
Quality Measures for Graph Clustering

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

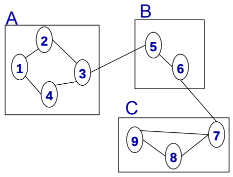
$$CC = \frac{|E^{in}| - |E^{ex}|}{|E|}$$

- 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



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

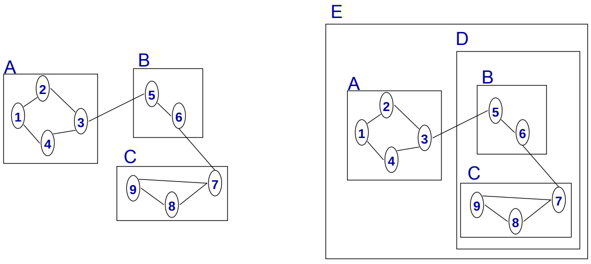


Cut size =2

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

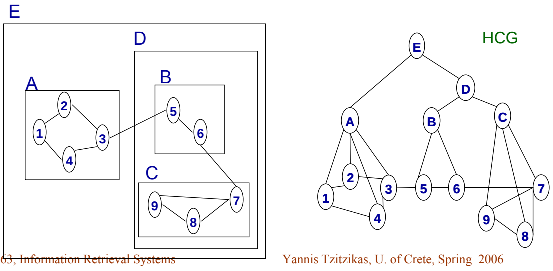
Hierarchical Graph Clustering

- 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)

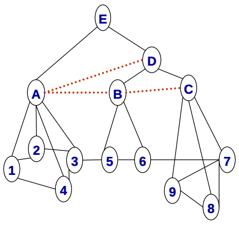


Implied Edges

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



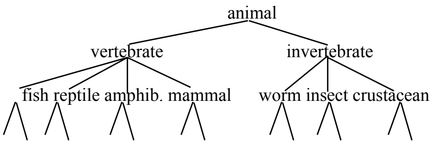
A Hierarchical Compound Graph is a triad (G,T, I) where (G,T) is a hierarchical clustered graph (HCG), and I the set of implied edges set.

Graph Theoretic Clustering Approaches

- 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*) from a set of unlabeled examples.
- 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’)

single link: similarity of two most similar. = max{ sim(d,d’) |d∈c,d’∈c’}

SIM(c,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’}

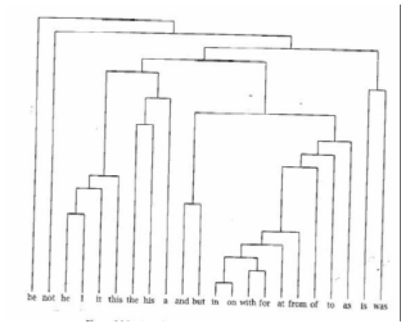
3. Βρες το ζεύγος {Cu,Cv} με την υψηλότερη (inter-cluster) ομοιότητα

4. Συγχώνευσε τα clusters Cu, Cv

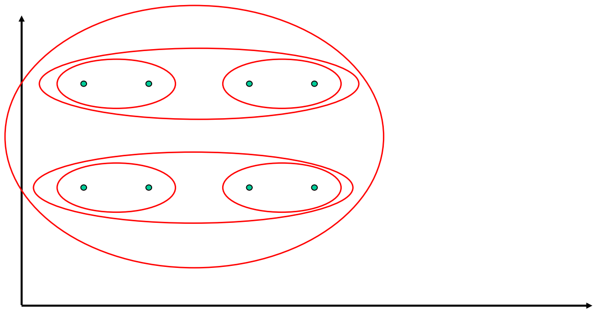
5. Επανέλαβε (από το βήμα 2) έως ότου να καταλήξουμε να έχουμε 1 μόνο cluster

6. Επέστρεψε την ιεραρχία των clusters (το ιστορικό των συγχωνεύσεων)

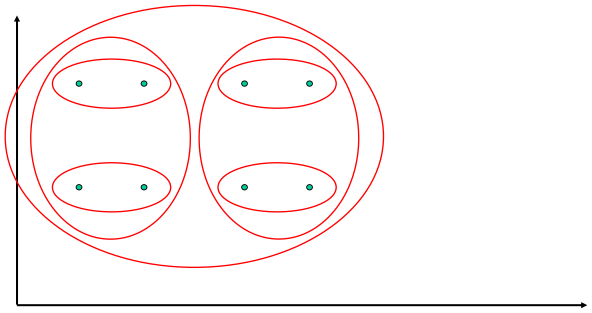
Dendogram or Cluster Hierarchy



Single Link Example



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

$$\left(\begin{array}{lllll} d_1 & & & & \\ d_2 & s_{21} & & & \\ d_3 : & s_{31} & s_{32} & & \\ : & : & : & & : \\ d_n & s_{n1} & s_{n2} & \dots & s_{n,n-1} \\ & d_1 & d_2 & \dots\dots d_{n-1} & d_n \end{array}\right)$$

Empty because sim(X,Y)=sim(Y,X)

- 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

- 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 (or reallocation methods)

- Select K cluster representatives
- For i = 1 to N, assign di to the most similar centroid
- For j = 1 to K, recalculate the cluster centroid cj
- Repeat the above steps 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

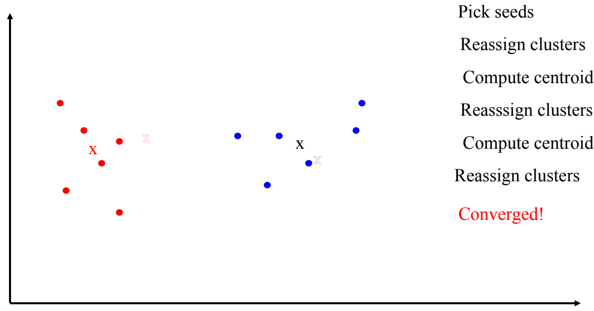
K-Means

- 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

Χειρονακτική Ομαδοποίηση

Human Clustering

- Ερωτήματα
  - 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

Text Clustering

- 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).

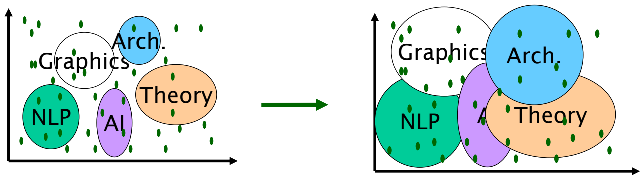
Related Issues

Clustering vs Classification

- Clustering
  - Unsupervised
  - Input
    - Clustering algorithm
    - Similarity measure
    - Number of clusters (e.g. in K Means)
  - No specific information for each document
- Classification (or categorization)
  - Supervised
  - Each document is labeled with a class
  - Build a classifier that assigns documents to one of the classes

Text Classification Example

- Labeled training set
- Classification of all documents



Supervised vs Unsupervised Learning

- This setup is called *supervised learning* in the terminology of Machine Learning
- In the domain of text, various names
  - **Text classification, text categorization**
  - **Document classification/categorization**
  - **“Automatic” categorization**
  - **Routing, filtering …**
- In contrast, the earlier setting of clustering is called *unsupervised learning*
  - Presumes no availability of training samples
  - Clusters output may not be thematically unified.

Text Categorization Examples

Assign labels to each document or web-page:

- Labels are most often **topics** such as Yahoo-categories

  e.g., "finance," "sports," "news>world>asia>business"
- Labels may be **genres**

  e.g., "editorials" "movie-reviews" "news“
- Labels may be **opinion**

  e.g., “like”, “hate”, “neutral”
- Labels may be **domain-specific binary**

  e.g., "interesting-to-me" : "not-interesting-to-me”

  e.g., “spam” : “not-spam”

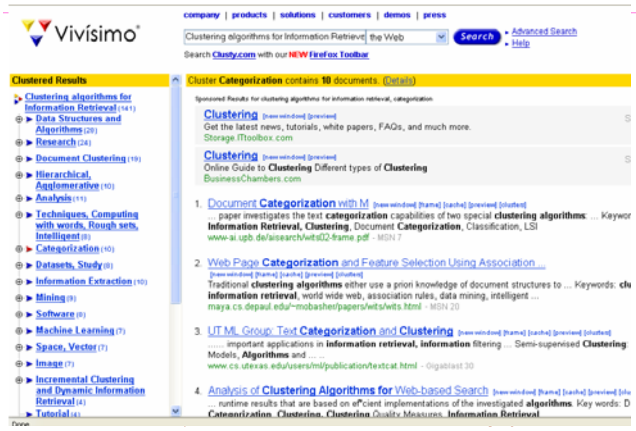
  e.g., “contains adult language” :“doesn’t”

Classification Methods

- Manual classification
  - Used by Yahoo!, Looksmart, about.com, ODP, Medline
  - very accurate when job is done by experts
  - consistent when the problem size and team is small
  - difficult and expensive to scale
- Automatic document classification
  - Hand-coded **rule-based systems**
    - Used by spam filters, Reuters, CIA, Verity, …
      - E.g., assign category if document contains a given boolean combination of words
    - Commercial systems have complex query languages (everything in IR query languages + *accumulators*)
    - Accuracy is often very high if a query has been carefully refined over time by a subject expert
    - Building and maintaining these queries is expensive

Classification Methods (II)

- Supervised learning of document-label assignment function
  - Many new systems rely on machine learning (Autonomy, Kana, MSN, Verity, Enkata, …)
    - k-Nearest Neighbors (simple, powerful)
    - Naive Bayes (simple, common method)
    - Support-vector machines (new, more powerful)
    - … plus many other methods
    - No free lunch: requires hand-classified training data
    - But can be built (and refined) by amateurs



Vivísimo
company | products | solutions | customers | demos | press
Clustering algorithms for Information Retrieve
the Web
Search
Advanced Search
Help

Search Clusty.com with our NEW FireFox Toolbar

Clustered Results

- Clustering algorithms for Information Retrieval (141)
  - Data Structures and Algorithms (20)
  - Research (24)
  - Document Clustering (19)
  - Hierarchical, Agglomerative (10)
  - Analysis (11)
  - Techniques, Computing with words, Rough sets, Intelligent (8)
  - Categorization (10)
  - Datasets, Study (8)
  - Information Extraction (10)
  - Mining (9)
  - Software (8)
  - Machine Learning (7)
  - Space, Vector (7)
  - Image (7)
  - Incremental Clustering and Dynamic Information Retrieval (4)
  - Tutorial (4)

Cluster **Categorization** contains **10** documents. (Details)

Sponsored Results for clustering algorithms for information retrieval, categorization

Clustering [new window] [preview]
Get the latest news, tutorials, white papers, FAQs, and much more.
Storage.ITtoolbox.com

Clustering [new window] [preview]
Online Guide to **Clustering** Different types of **Clustering**
BusinessChambers.com

1. Document **Categorization** with M [new window] [frame] [cache] [preview] [clusters]
   ... paper investigates the text **categorization** capabilities of two special **clustering algorithms**: ... Keywor **Information Retrieval, Clustering**, Document **Categorization**, Classification, LSI
   www-ai.upb.de/aisearch/wits02-frame.pdf - MSN 7
2. Web Page **Categorization** and Feature Selection Using Association ... [new window] [frame] [cache] [preview] [clusters]
   Traditional **clustering algorithms** either use a priori knowledge of document structures to ... Keywords: **clu** **information retrieval**, world wide web, association rules, data mining, intelligent ...
   maya.cs.depaul.edu/~mobasher/papers/wits/wits.html - MSN 20
3. UT ML Group: Text **Categorization** and **Clustering** [new window] [frame] [cache] [preview] [clusters]
   ...... important applications in **information retrieval, information** filtering ... Semi-supervised **Clustering**: Models, **Algorithms** and ... ...
   www.cs.utexas.edu/users/ml/publication/textcat.html - Gigablast 30
4. Analysis of **Clustering Algorithms for** Web-based Search [new window] [frame] [cache] [preview] [clu
   ... runtime results that are based on efﬁcient implementations of the investigated **algorithms**. Key words: D **Categorization, Clustering, Clustering** Quality Measures, **Information Retrieval**

Done