Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
Άνοιξη 2008

ΗΥ463 - Συστήματα Ανάκτησης Πληροφοριών
Information Retrieval (IR) Systems

Ομαδοποίηση Εγγράφων
(Document Clustering)

Γιάννης Τζίτζικας

Διάλεξη :13
Ημερομηνία :

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.

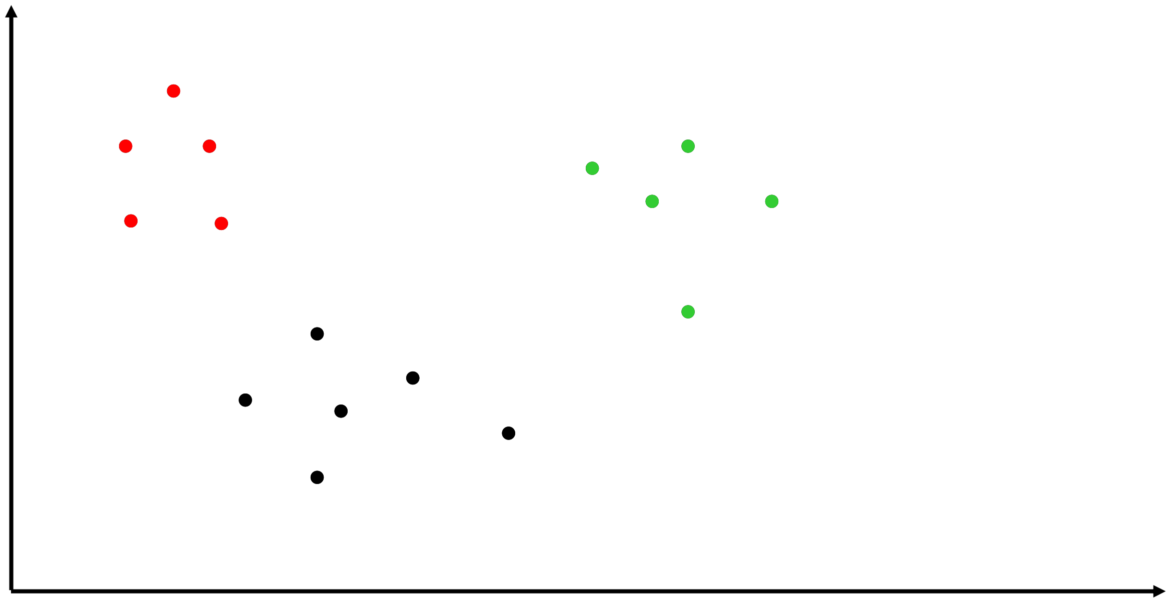
Clustering

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:

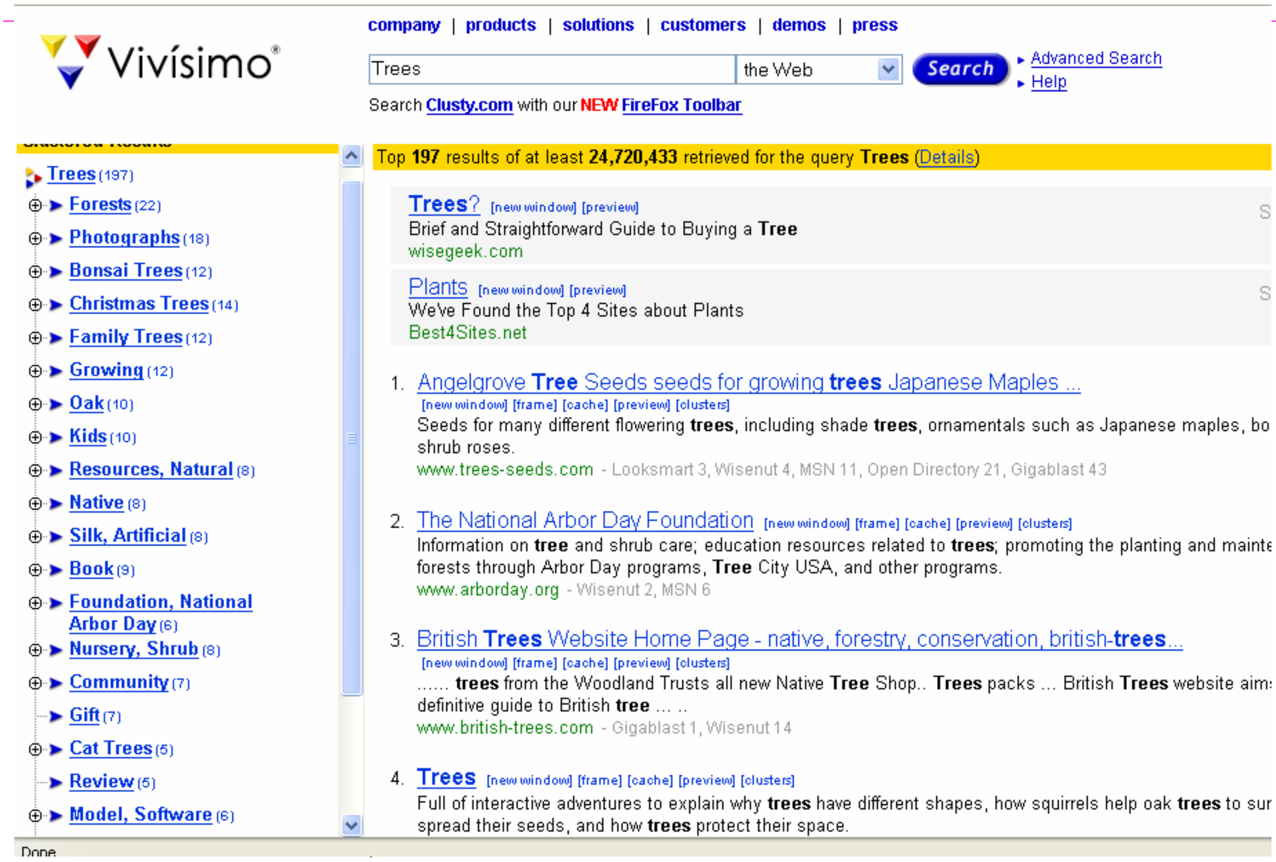
$C_1 \cup \ldots \cup C_k = N$ and $C_i \cap C_j = \varnothing$, for all $i \neq j$.

- According to the above definition, clusters are disjoint ($C_i \cap C_j = \varnothing$, for all $i \neq j$.) However there are clustering approaches that yield overlapping clusters (it may be $C_i \cap C_j \neq \varnothing$)
- 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



Search Clusty.com with our NEW FireFox Toolbar

- 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)

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

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

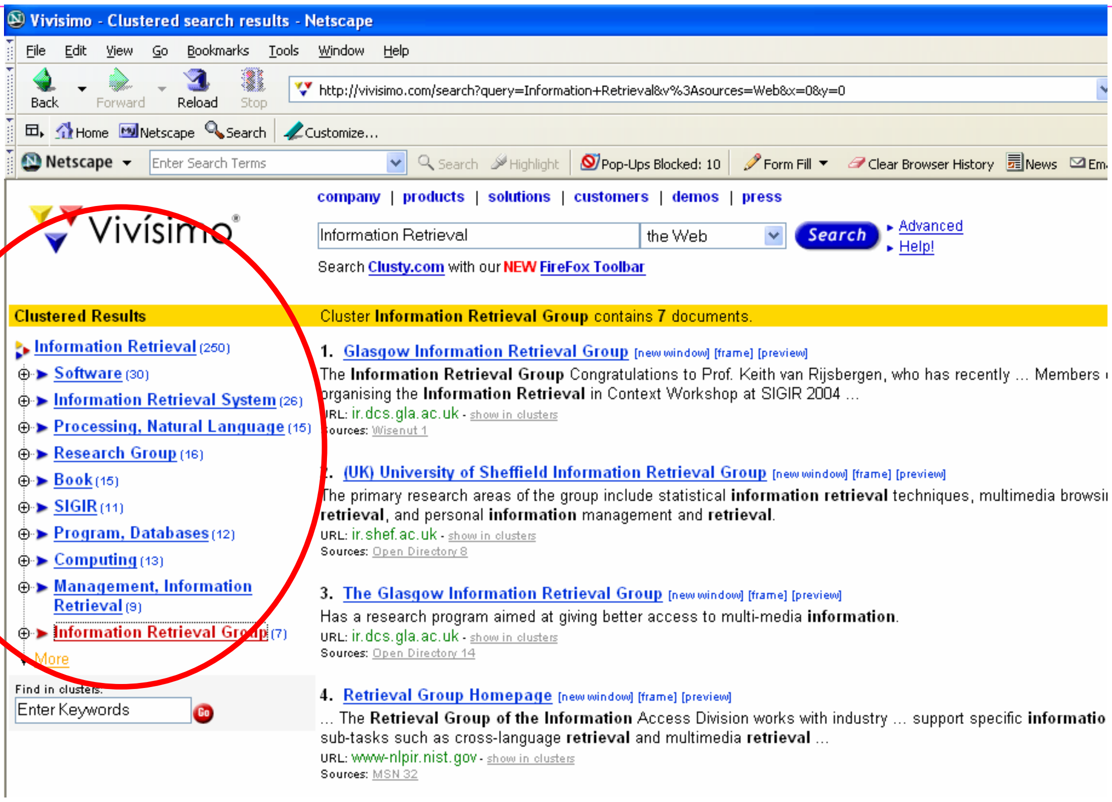
1. Angelgrove **Tree** Seeds seeds for growing **trees** Japanese Maples ... [new window] [frame] [cache] [preview] [clusters]
Seeds for many different flowering **trees**, including shade **trees**, ornamentals such as Japanese maples, bo shrub roses.
www.trees-seeds.com - Looksmart 3, Wisenut 4, MSN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

Παράδειγμα ομαδοποίησης αποτελεσμάτων www.vivisimo.com


Vivisimo - Clustered search results - Netscape
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://vivisimo.com/search?query=Information+Retrieval&v%3Asources=Web&x=0&y=0
Home Netscape Search Customize...
Netscape Enter Search Terms Search Highlight Pop-Ups Blocked: 10 Form Fill Clear Browser History News Em
Vivísimo®
company | products | solutions | customers | demos | press
Information Retrieval
the Web
Search
Advanced
Help!
Go

Search Clusty.com with our NEW FireFox Toolbar

Clustered Results

- Information Retrieval (250)
- Software (30)
- Information Retrieval System (26)
- Processing, Natural Language (15)
- Research Group (16)
- Book (15)
- SIGIR (11)
- Program, Databases (12)
- Computing (13)
- Management, Information Retrieval (9)
- Information Retrieval Group (7)
- More

Find in clusters:
Enter Keywords

Cluster **Information Retrieval Group** contains **7** documents.

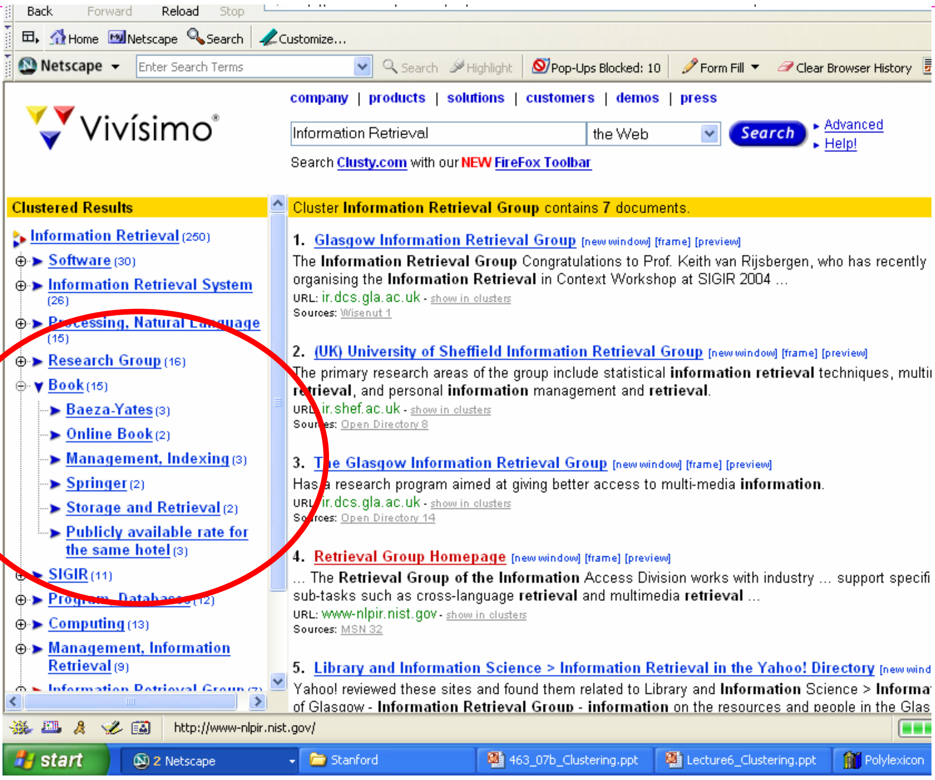
1. Glasgow Information Retrieval Group [new window] [frame] [preview]
The **Information Retrieval Group** Congratulations to Prof. Keith van Rijsbergen, who has recently ... Members organising the **Information Retrieval** in Context Workshop at SIGIR 2004 ...
URL: ir.dcs.gla.ac.uk - show in clusters
Sources: 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
Sources: 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
Sources: 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
Sources: MSN 32

Παράδειγμα ομαδοποίησης αποτελεσμάτων



Search Clusty.com with our NEW FireFox Toolbar

Clustered Results

- Information Retrieval (250)
  - Software (30)
  - Information Retrieval System (26)
  - Processing, Natural Language (15)
  - 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)
  - Program, Databases (12)
  - Computing (13)
  - Management, Information Retrieval (9)
  - Information Retrieval Group (7)

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. (UK) 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 8

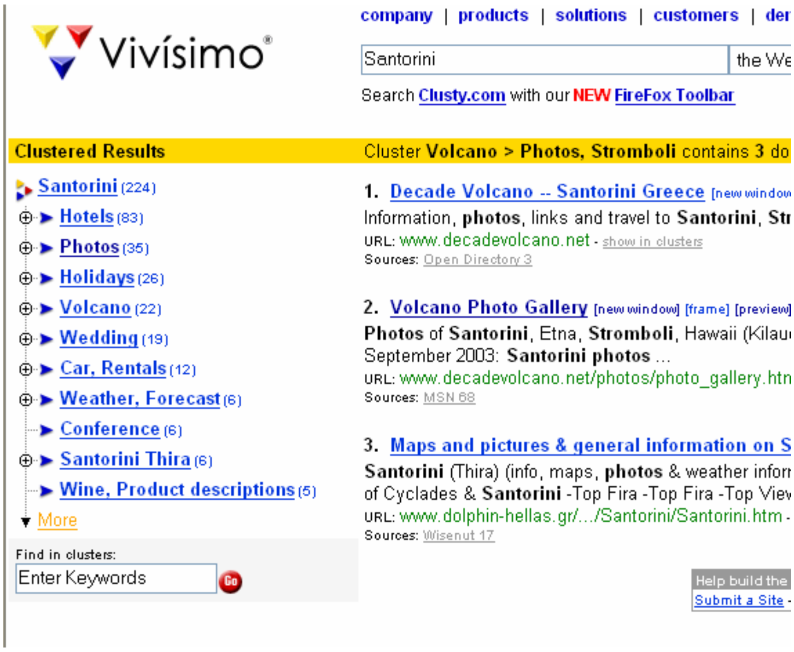
3. The Glasgow Information Retrieval Group [new window] [frame] [preview]
Has a research program aimed at giving better access to multi-media **information**.
URL: ir.dcs.gla.ac.uk - show in clusters
Sources: Open Directory 14

4. Retrieval Group Homepage [new window] [frame] [preview]
... The **Retrieval Group of the Information** Access Division works with industry ... support specifi 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 wind
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.htn
Sources: MSN 68

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

Submit a Site -

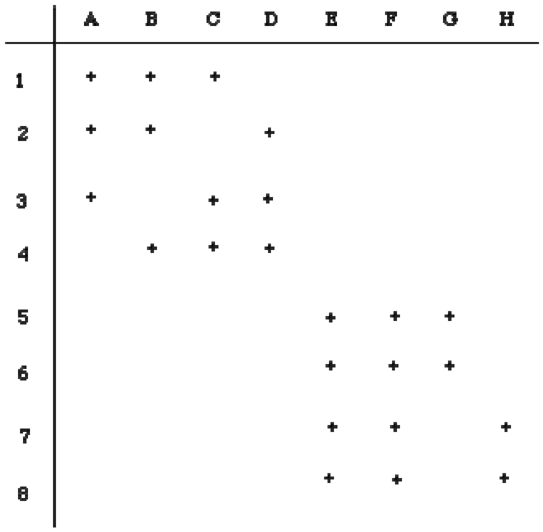
Τύποι Ομαδοποίησης

- 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):
  - 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 $|X \cap Y|$ 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)

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:

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

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

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

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)

Ένας τρόπος επιτάχυνσης της αποτίμησης των επερωτήσεων θα μπορούσε να είναι ο εξής

Method:

1/ Cluster all documents of the collection

- **We have to do it only once**

2/ Represent clusters by mean or average document

- **We have to do it only once**

3/ compare each received query to the cluster representatives

- **It is like ranking the cluster representatives (as if they were document vectors)**

4/ Return the documents of the most similar(s) 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

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

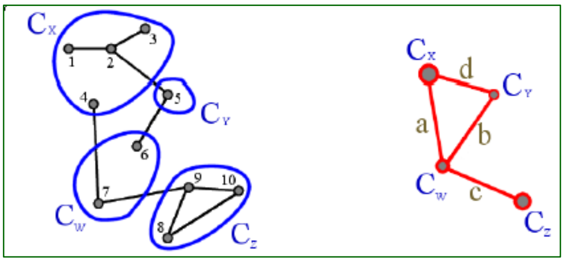
Criteria

- Stability under growth
  - The method produces a clustering which is unlikely to be altered drastically when further objects are incorporated (stable under growth)
- Stability
  - The method is stable in the sense that small errors in the description of objects lead to small changes in the clustering
- Order Independence
  - 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
  - the <u>nodes</u> of the graph are the objects to be clustered
  - an <u>edge</u> between two nodes of the graph exist if the similarity of the nodes is greater than some threshold
  - we can view the clustering process as a process that groups similar nodes into a set of subgraphs

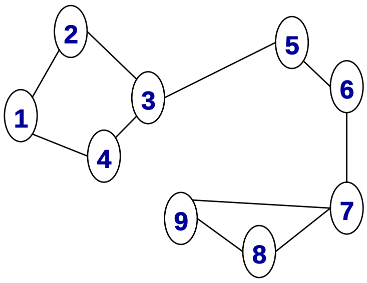


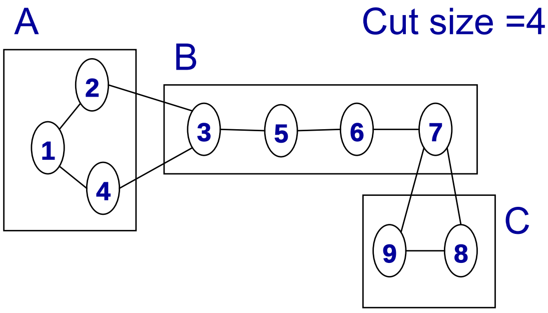
Quality criteria for graph clustering methods

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

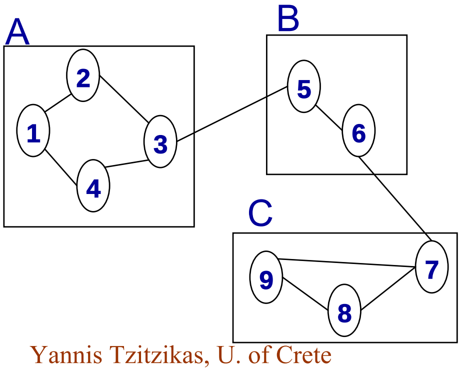
- high cohesion:
  - there should be many internal edges
- low “cut size”:
  - The cut size (else called *external cost)* of a clustering measures how many edges are external to all sub-graphs, that is, how many edges cross cluster boundaries.
- Uniformity of cluster size is also often desirable.
  -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

Example





Cut size =4



Cut size =2

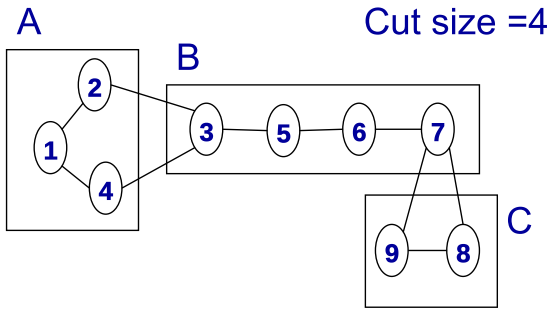
Quality Measures for Graph Clustering

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

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

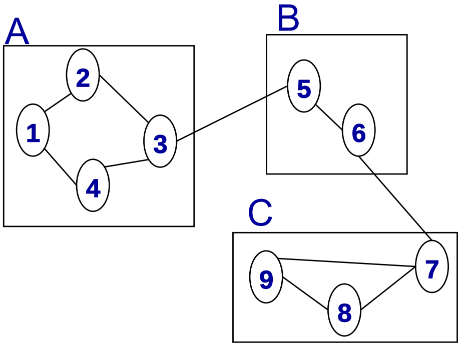
- E^{in}: the “internal” edges: those that connect nodes of the same cluster
- E^{ex}: the “external” edges: those that cross cluster boundaries
- maximum value of CC: 1
  - when all edges are internal
- minimum value of CC: -1
  - when all edges are external

Example



Cut size =4

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

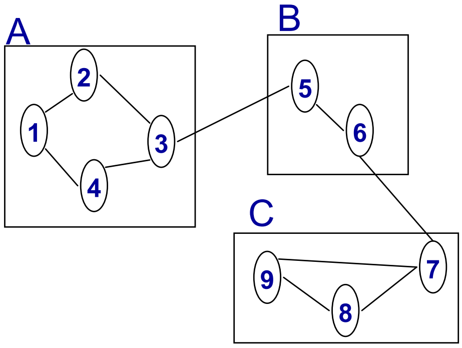


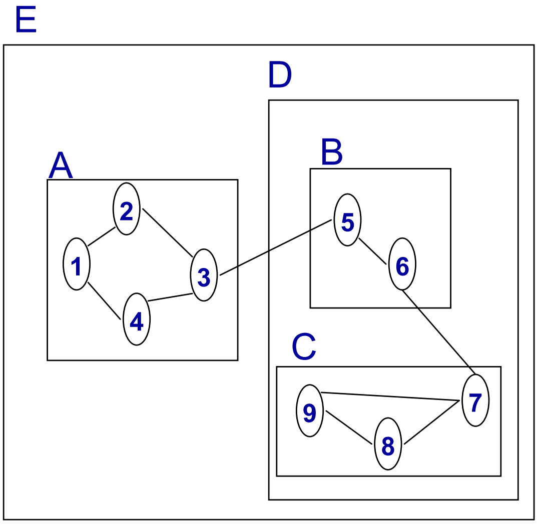
Cut size =2

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

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.





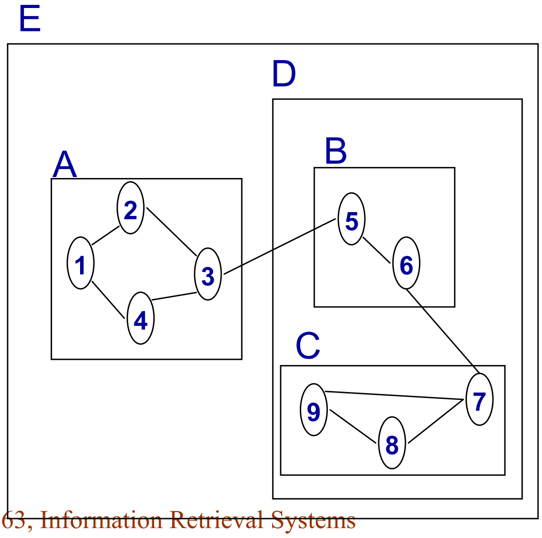
Hierarchical Clustered Graph

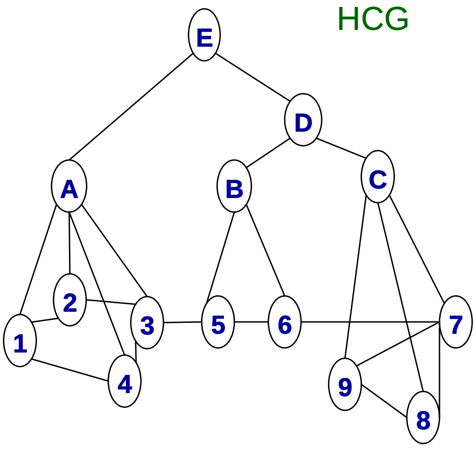
A Hierarchical Clustered Graph (HCG) is a pair (G,T) where

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)



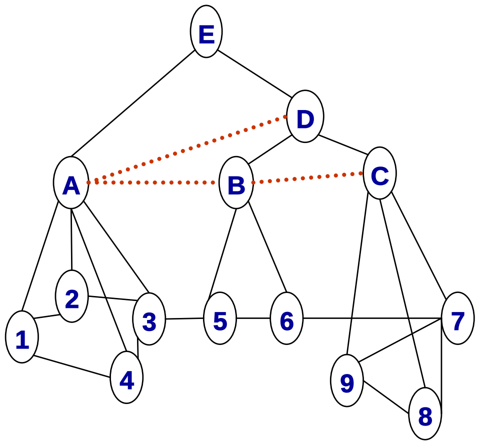


Implied Edges

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



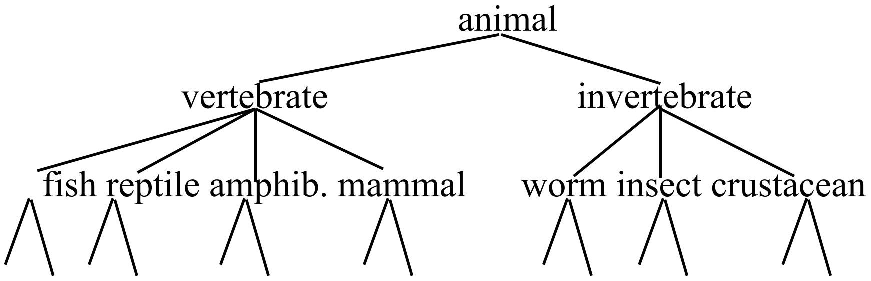
A Hierarchical Compound Graph is a triad (G,T, I) where (G,T) is a hierarchical clustered graph (HCG), and I the set of implied edges set.

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*).
- 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/ Βάλε 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’}

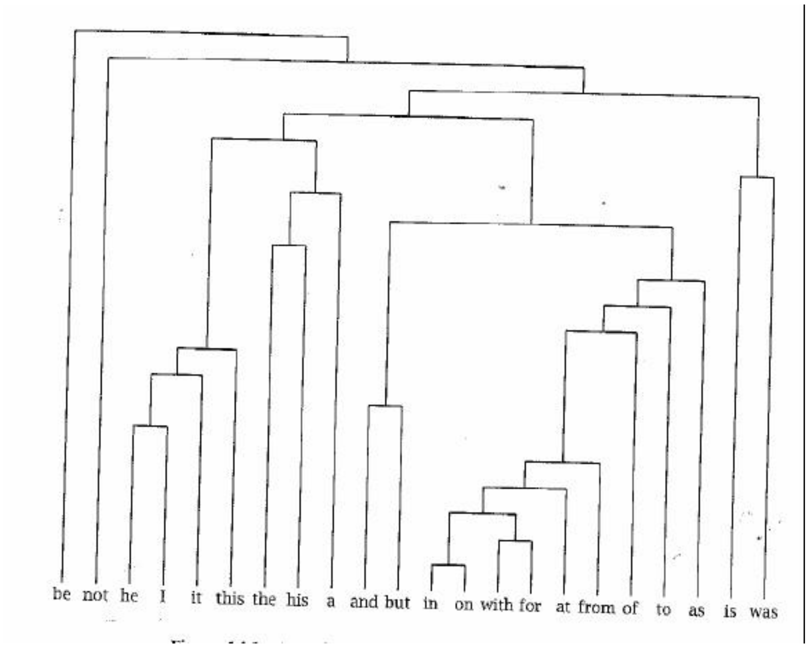
3. Βρες το ζεύγος {Cu,Cv} με την υψηλότερη (inter-cluster) ομοιότητα

4. Συγχώνευσε τα clusters Cu, Cv

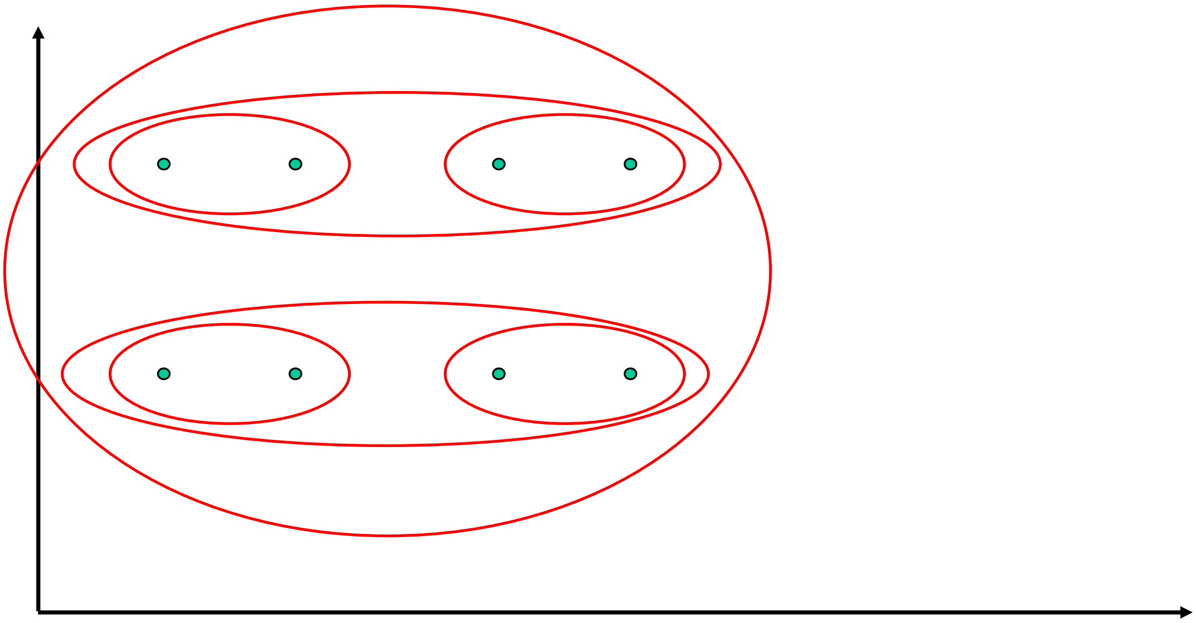
5. Επανέλαβε (από το βήμα 2) έως ότου να καταλήξουμε να έχουμε 1 μόνο cluster

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

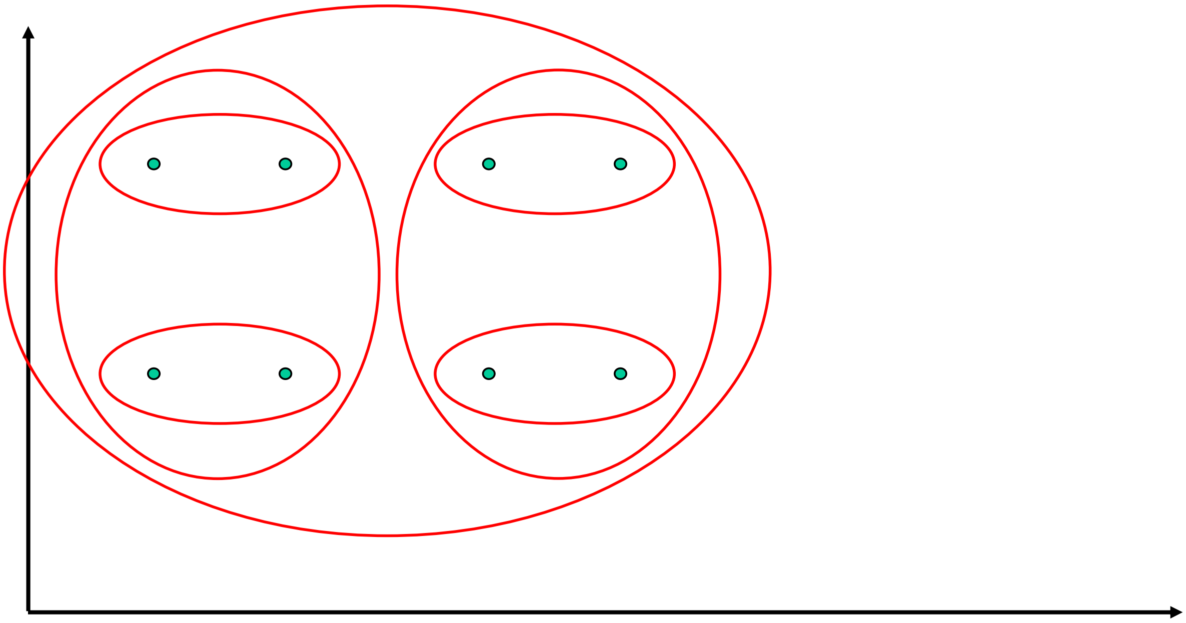
Dendogram or Cluster Hierarchy



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

$$\begin{array}{c|ccccc}
d_1 & & & & & \\
d_2 & s_{21} & & & & \\
d_3 : & s_{31} & s_{32} & & & \\
: & : & : & & : & \\
d_n & s_{n1} & s_{n2} & \ldots & s_{n,n-1} & \\
\hline
 & d_1 & d_2 & \ldots. & d_{n-1} & d_n
\end{array}$$

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

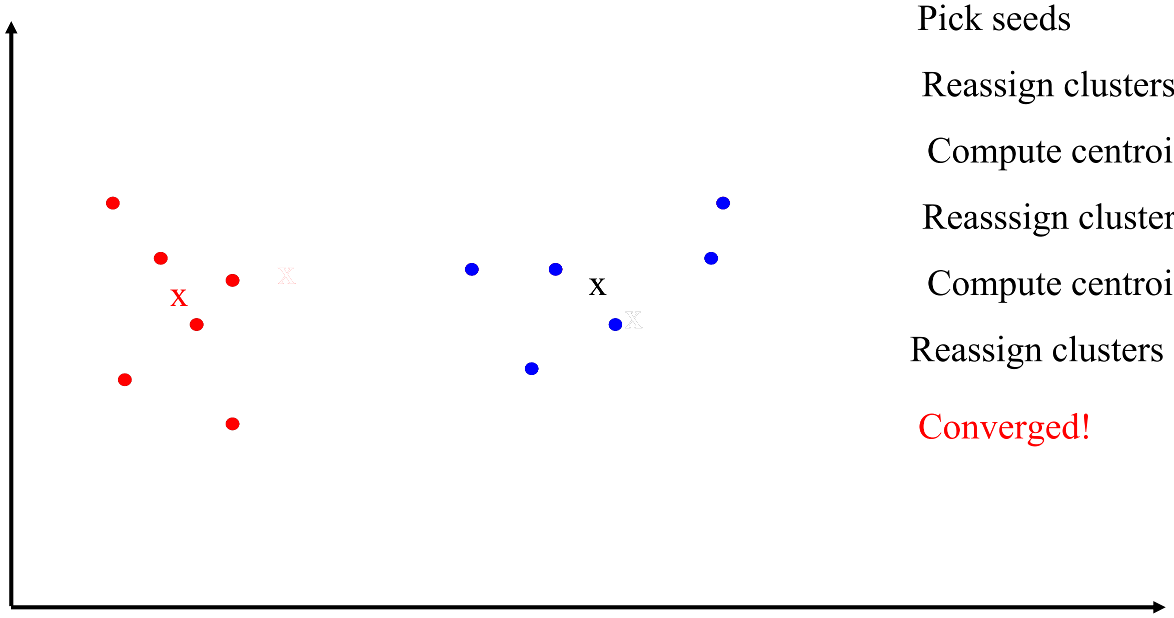
K-Means

- 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

Some notes

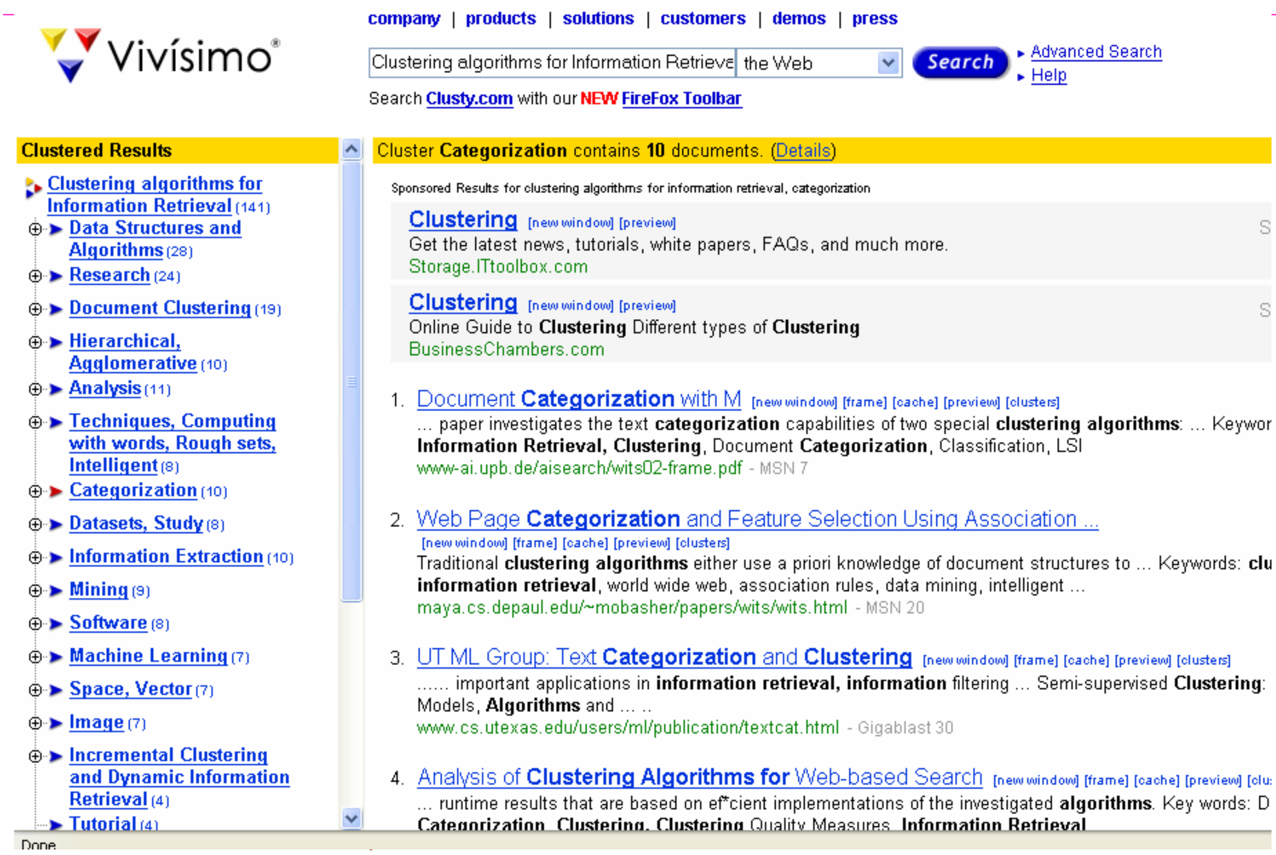
- HAC and K-Means have been applied to text in a straightforward way.
- Typically use ***normalized***, TF/IDF-weighted vectors and cosine similarity.
- Optimize computations for sparse vectors.
- Applications:
  - During retrieval, **add other documents** in the same cluster as the initial retrieved 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

Παραδείγματα Ομαδοποίησης Αποτελεσμάτων Αναζήτησης (Results Clustering)

Vivisimo.com



Vivísimo
company | products | solutions | customers | demos | press
Clustering algorithms for Information Retrieva
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 (28)
  - 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)

Done

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

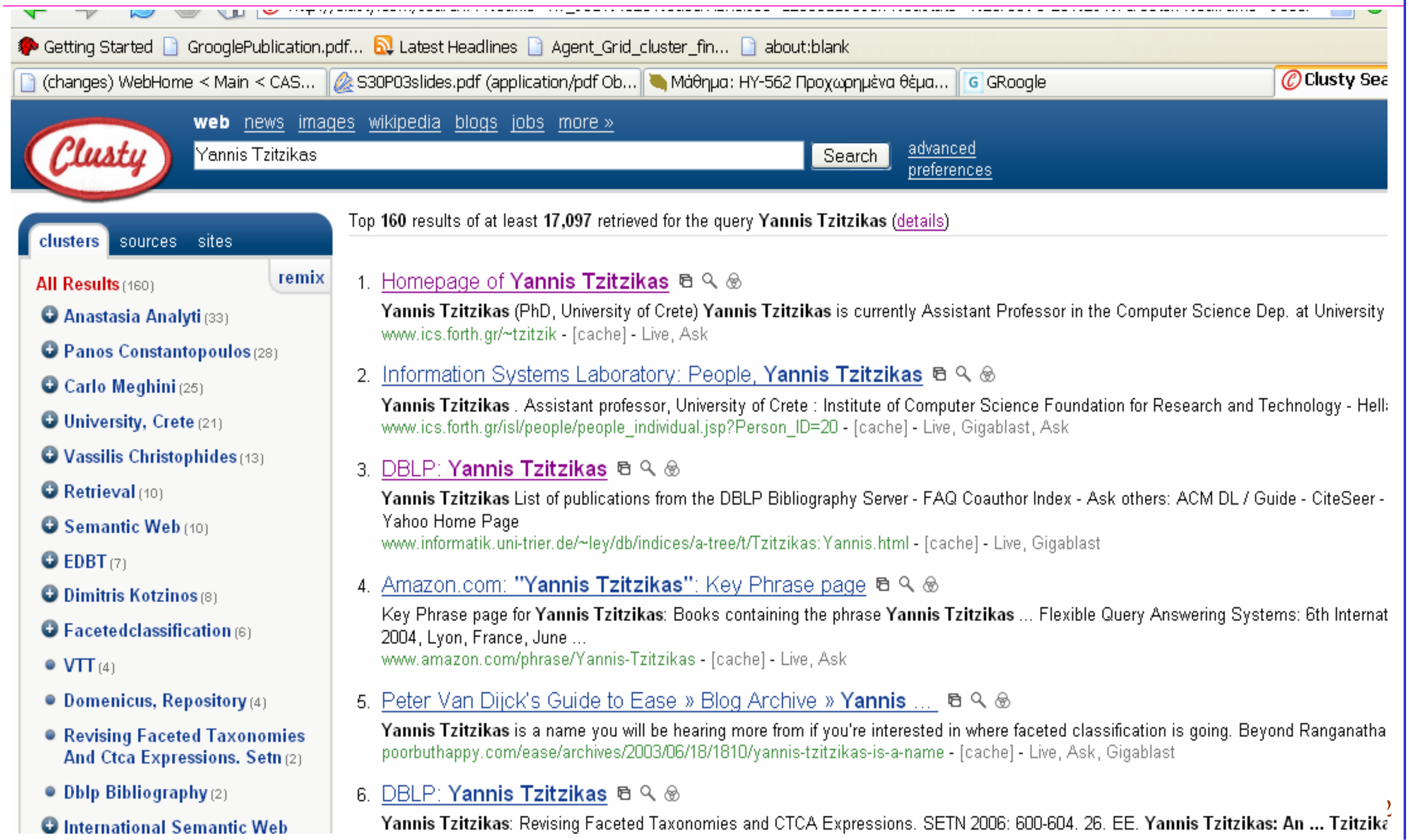
1. Document **Categorization** with M [new window] [frame] [cache] [preview] [clusters]
... paper investigates the text **categorization** capabilities of two special **clustering algorithms**: ... Keywor
Information Retrieval, Clustering, Document **Categorization**, Classification, LSI
www-ai.upb.de/aisearch/wits02-frame.pdf - 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 ef*cient implementations of the investigated **algorithms**. Key words: D
Categorization, Clustering, Clustering Quality Measures, **Information Retrieval**

Clusty.com



Getting Started
GrooglePublication.pdf...
Latest Headlines
Agent_Grid_cluster_fin...
about:blank
(changes) WebHome < Main < CAS...
S30P03slides.pdf (application/pdf Ob...
Μάθημα: HY-562 Προχωρημένα θέμα...
GRoogle
Clusty Sea
Clusty
web news images wikipedia blogs jobs more »
Yannis Tzitzikas
Search
advanced
preferences
clusters sources sites
remix

- All Results (160)
  - Anastasia Analyti (33)
  - Panos Constantopoulos (28)
  - Carlo Meghini (25)
  - University, Crete (21)
  - Vassilis Christophides (13)
  - Retrieval (10)
  - Semantic Web (10)
  - EDBT (7)
  - Dimitris Kotzinos (8)
  - Facetedclassification (6)
  - VTT (4)
  - Domenicus, Repository (4)
  - Revising Faceted Taxonomies And Ctca Expressions. Setn (2)
  - Dblp Bibliography (2)
  - International Semantic Web

Top **160** results of at least **17,097** retrieved for the query **Yannis Tzitzikas** (details)

1. Homepage of **Yannis Tzitzikas**
Yannis Tzitzikas (PhD, University of Crete) **Yannis Tzitzikas** is currently Assistant Professor in the Computer Science Dep. at University
www.ics.forth.gr/~tzitzik - [cache] - Live, Ask

2. Information Systems Laboratory: People, **Yannis Tzitzikas**
Yannis Tzitzikas . Assistant professor, University of Crete : Institute of Computer Science Foundation for Research and Technology - Hell
www.ics.forth.gr/isl/people/people_individual.jsp?Person_ID=20 - [cache] - Live, Gigablast, Ask

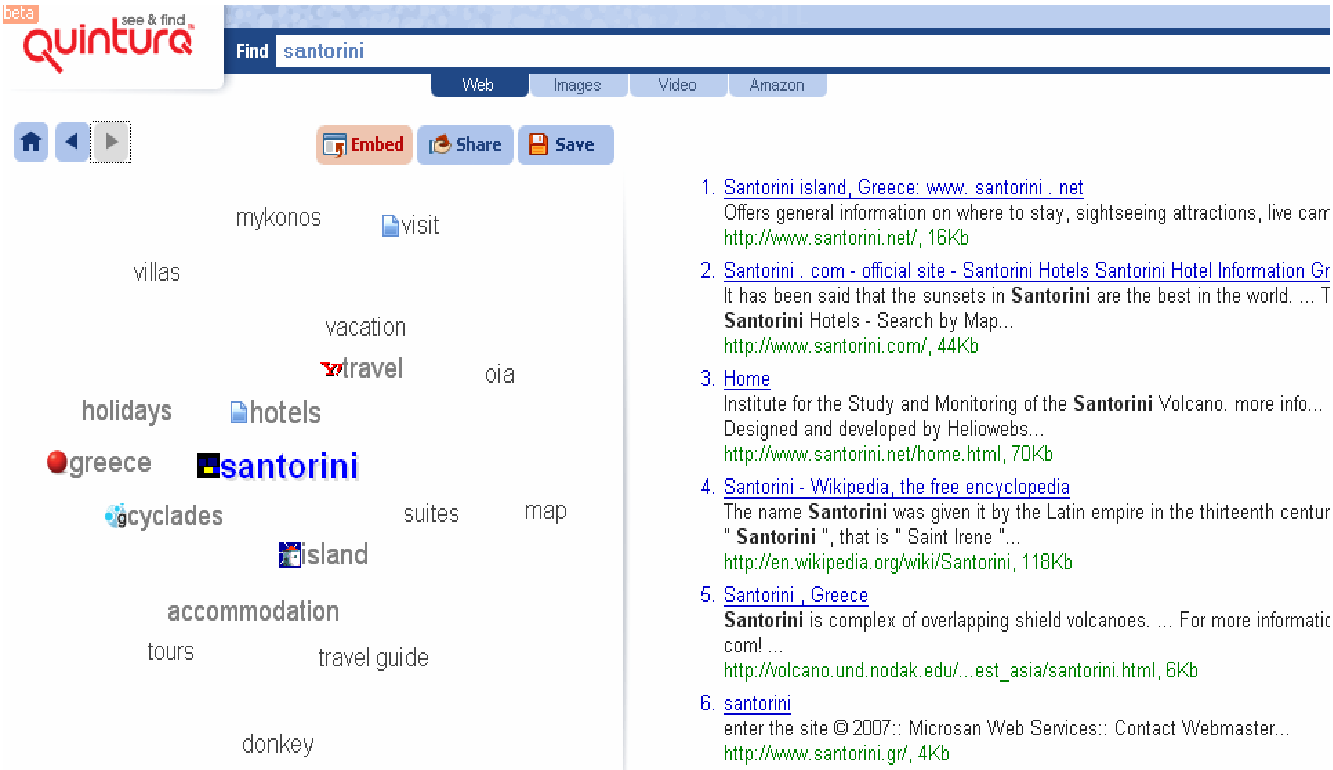
3. DBLP: **Yannis Tzitzikas**
Yannis Tzitzikas List of publications from the DBLP Bibliography Server - FAQ Coauthor Index - Ask others: ACM DL / Guide - CiteSeer - Yahoo Home Page
www.informatik.uni-trier.de/~ley/db/indices/a-tree/t/Tzitzikas:Yannis.html - [cache] - Live, Gigablast

4. Amazon.com: "**Yannis Tzitzikas**": Key Phrase page
Key Phrase page for **Yannis Tzitzikas**: Books containing the phrase **Yannis Tzitzikas** ... Flexible Query Answering Systems: 6th Internat 2004, Lyon, France, June ...
www.amazon.com/phrase/Yannis-Tzitzikas - [cache] - Live, Ask

5. Peter Van Dijck's Guide to Ease » Blog Archive » **Yannis** ...
Yannis Tzitzikas is a name you will be hearing more from if you're interested in where faceted classification is going. Beyond Ranganatha
poorbuthappy.com/ease/archives/2003/06/18/1810/yannis-tzitzikas-is-a-name - [cache] - Live, Ask, Gigablast

6. DBLP: **Yannis Tzitzikas**
Yannis Tzitzikas: Revising Faceted Taxonomies and CTCA Expressions. SETN 2006: 600-604. 26. EE. **Yannis Tzitzikas: An ... Tzitzika**

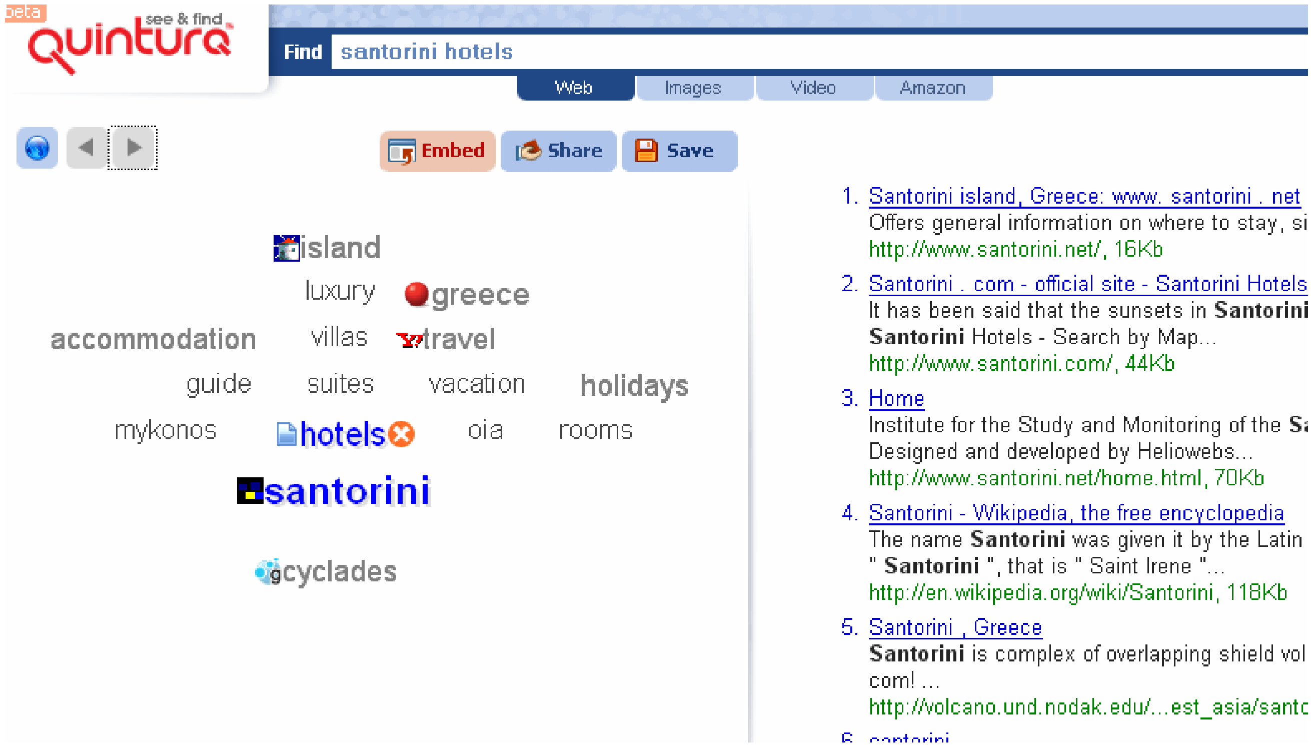
http://www.quintura.com/



1. Santorini island, Greece: www. santorini . net
Offers general information on where to stay, sightseeing attractions, live carr
http://www.santorini.net/, 16Kb
2. Santorini . com - official site - Santorini Hotels Santorini Hotel Information Gr
It has been said that the sunsets in **Santorini** are the best in the world. ... T
Santorini Hotels - Search by Map...
http://www.santorini.com/, 44Kb
3. Home
Institute for the Study and Monitoring of the **Santorini** Volcano. more info...
Designed and developed by Heliowebs...
http://www.santorini.net/home.html, 70Kb
4. Santorini - Wikipedia, the free encyclopedia
The name **Santorini** was given it by the Latin empire in the thirteenth centur
" **Santorini** ", that is " Saint Irene "...
http://en.wikipedia.org/wiki/Santorini, 118Kb
5. Santorini , Greece
Santorini is complex of overlapping shield volcanoes. ... For more informatic
com! ...
http://volcano.und.nodak.edu/...est_asia/santorini.html, 6Kb
6. santorini
enter the site © 2007:: Microsan Web Services:: Contact Webmaster...
http://www.santorini.gr/, 4Kb

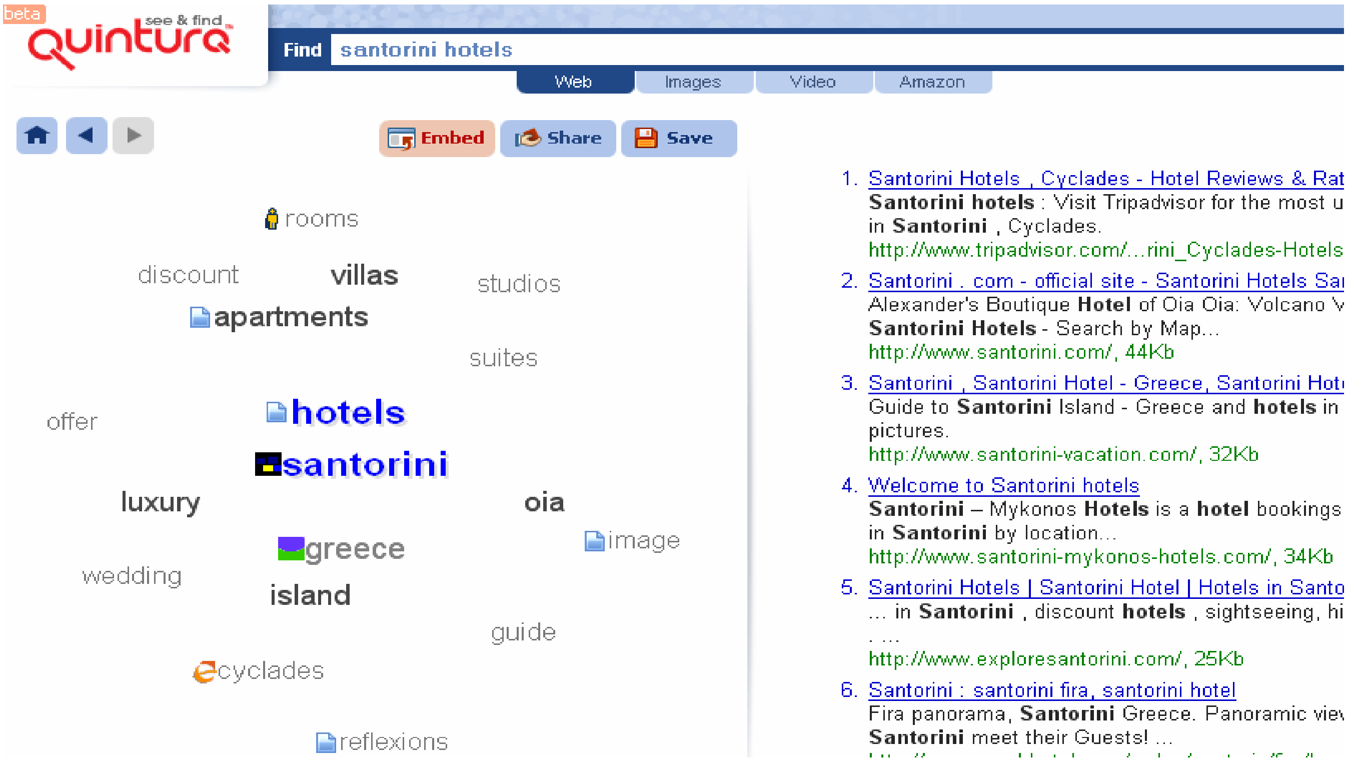
http://www.quintura.com/

- On mouse over "hotels"



1. Santorini island, Greece: www. santorini . net
Offers general information on where to stay, si
http://www.santorini.net/, 16Kb
2. Santorini . com - official site - Santorini Hotels
It has been said that the sunsets in **Santorini**
Santorini Hotels - Search by Map...
http://www.santorini.com/, 44Kb
3. Home
Institute for the Study and Monitoring of the **Sa**
Designed and developed by Heliowebs...
http://www.santorini.net/home.html, 70Kb
4. Santorini - Wikipedia, the free encyclopedia
The name **Santorini** was given it by the Latin
" **Santorini** ", that is " Saint Irene "...
http://en.wikipedia.org/wiki/Santorini, 118Kb
5. Santorini , Greece
Santorini is complex of overlapping shield vol
com! ...
http://volcano.und.nodak.edu/...est_asia/santc
6. santorini

http://www.quintura.com/

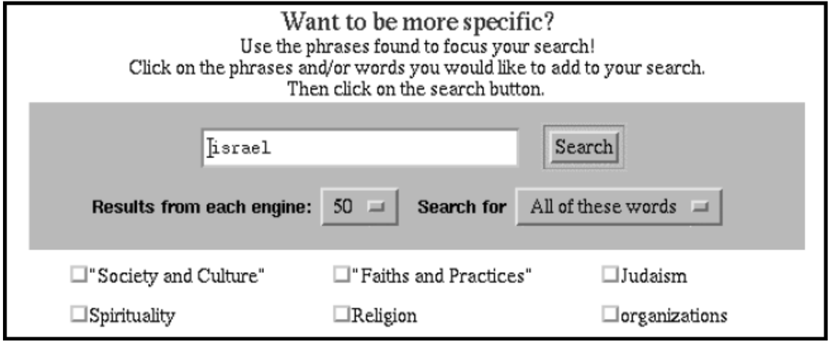


santorini hotels

1. Santorini Hotels , Cyclades - Hotel Reviews & Rat
Santorini hotels : Visit Tripadvisor for the most u
in Santorini , Cyclades.
http://www.tripadvisor.com/...rini_Cyclades-Hotels
2. Santorini . com - official site - Santorini Hotels Sa
Alexander's Boutique Hotel of Oia Oia: Volcano V
Santorini Hotels - Search by Map...
http://www.santorini.com/, 44Kb
3. Santorini , Santorini Hotel - Greece, Santorini Hot
Guide to Santorini Island - Greece and hotels in
pictures.
http://www.santorini-vacation.com/, 32Kb
4. Welcome to Santorini hotels
Santorini – Mykonos Hotels is a hotel bookings
in Santorini by location...
http://www.santorini-mykonos-hotels.com/, 34Kb
5. Santorini Hotels | Santorini Hotel | Hotels in Santo
... in Santorini , discount hotels , sightseeing, hi
. ...
http://www.exploresantorini.com/, 25Kb
6. Santorini : santorini fira, santorini hotel
Fira panorama, Santorini Greece. Panoramic vie
Santorini meet their Guests! ...

Grouper

- It is a Research Web Meta-Search Engine
- Users can specify the number of documents to be retrieved (10-200) from each of the participating search engines. The system queries 10 search engines, so it will retrieve 70-1000 documents.
- Clustering is applied on snippets returned by the search engines.
  - Snippet: a fragment of a web page returned by search engines summarizing the context of search keywords
- Clusters together documents with large common subphrases.
- It uses the Suffix Tree Clustering (STC) algorithm
  - create overlapping clusters because all suffixes of each phrase are generated



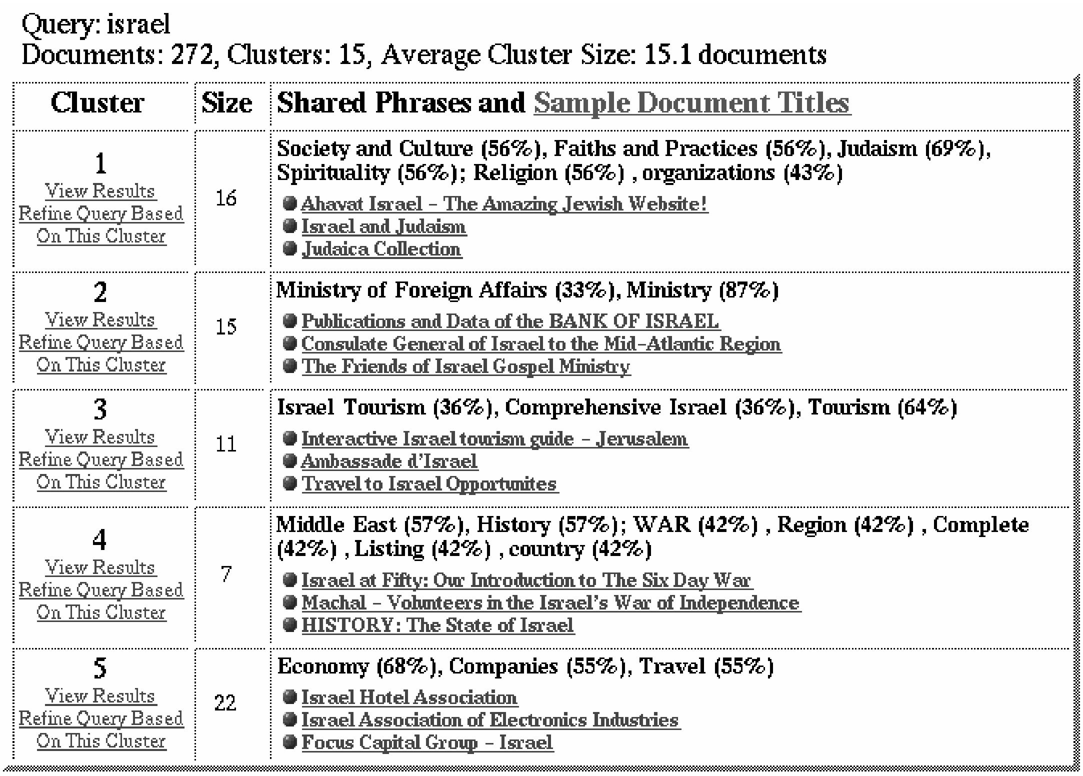
Want to be more specific?
Use the phrases found to focus your search!
Click on the phrases and/or words you would like to add to your search.
Then click on the search button.

israel

Results from each engine: 50 Search for All of these words

□"Society and Culture" □"Faiths and Practices" □Judaism
□Spirituality □Religion □organizations

Grouper



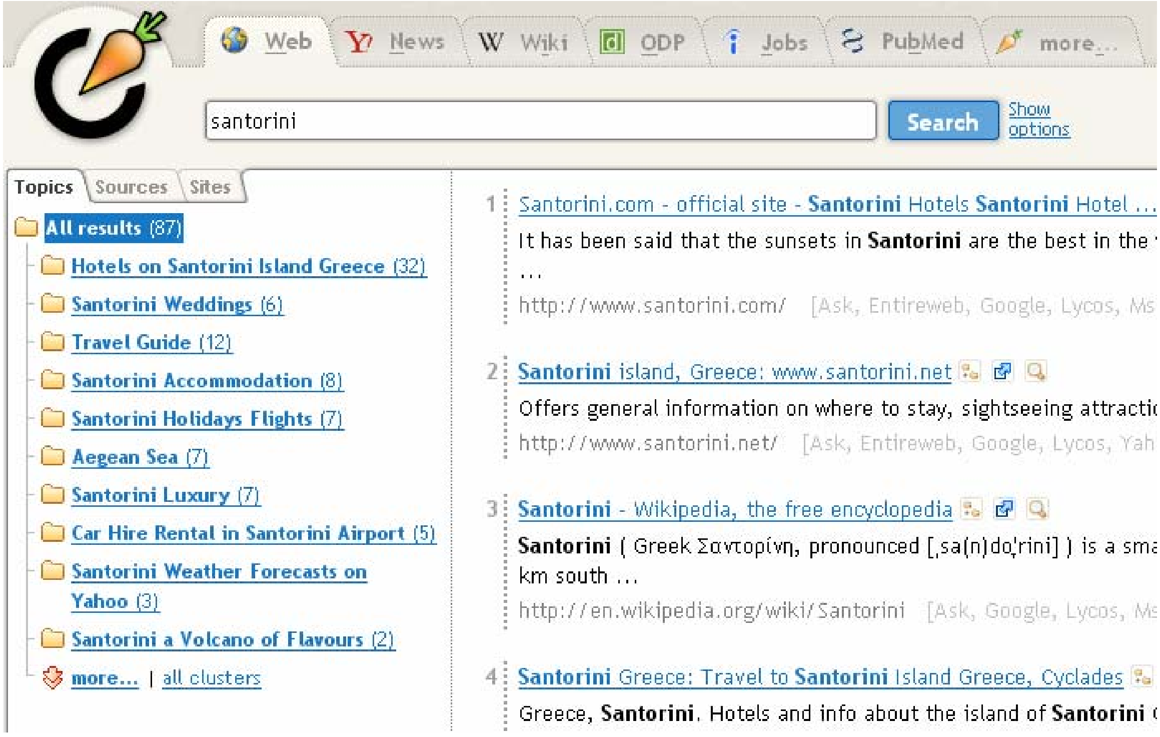
Query: israel
Documents: 272, Clusters: 15, Average Cluster Size: 15.1 documents

Cluster	Size	Shared Phrases and Sample Document Titles
1 View Results Refine Query Based On This Cluster	16	Society and Culture (56%), Faiths and Practices (56%), Judaism (69%), Spirituality (56%); Religion (56%) , organizations (43%) ● Ahavat Israel - The Amazing Jewish Website! ● Israel and Judaism ● Judaica Collection
2 View Results Refine Query Based On This Cluster	15	Ministry of Foreign Affairs (33%), Ministry (87%) ● Publications and Data of the BANK OF ISRAEL ● Consulate General of Israel to the Mid-Atlantic Region ● The Friends of Israel Gospel Ministry
3 View Results Refine Query Based On This Cluster	11	Israel Tourism (36%), Comprehensive Israel (36%), Tourism (64%) ● Interactive Israel tourism guide - Jerusalem ● Ambassade d'Israel ● Travel to Israel Opportunites
4 View Results Refine Query Based On This Cluster	7	Middle East (57%), History (57%); WAR (42%) , Region (42%) , Complete (42%) , Listing (42%) , country (42%) ● Israel at Fifty: Our Introduction to The Six Day War ● Machal - Volunteers in the Israel's War of Independence ● HISTORY: The State of Israel
5 View Results Refine Query Based On This Cluster	22	Economy (68%), Companies (55%), Travel (55%) ● Israel Hotel Association ● Israel Association of Electronics Industries ● Focus Capital Group - Israel

Carrot^2 (www.carrot2.org)

- Open-source Web Meta-Search Engine
- Acquire search results from various sources (YahooAPI, GoogleAPI, MSN Search API, eTools Meta Search, Alexa Web Search, PubMed, OpenSearch, Lucene index, SOLR)
- 5 clustering algorithms are available that are suitable for different kinds of document clustering tasks
  - STC
  - FussyAnts
  - Lingo
  - HAOG-STC
  - Rough k-means
- Open-source implementation of Grouper
- Lingo is the default clustering algorithm used in the Carrot2 live demos.

Carrot^2 (www.carrot2.org)



All results (87)
- Hotels on Santorini Island Greece (32)
- Santorini Weddings (6)
- Travel Guide (12)
- Santorini Accommodation (8)
- Santorini Holidays Flights (7)
- Aegean Sea (7)
- Santorini Luxury (7)
- Car Hire Rental in Santorini Airport (5)
- Santorini Weather Forecasts on Yahoo (3)
- Santorini a Volcano of Flavours (2)
- more... | all clusters

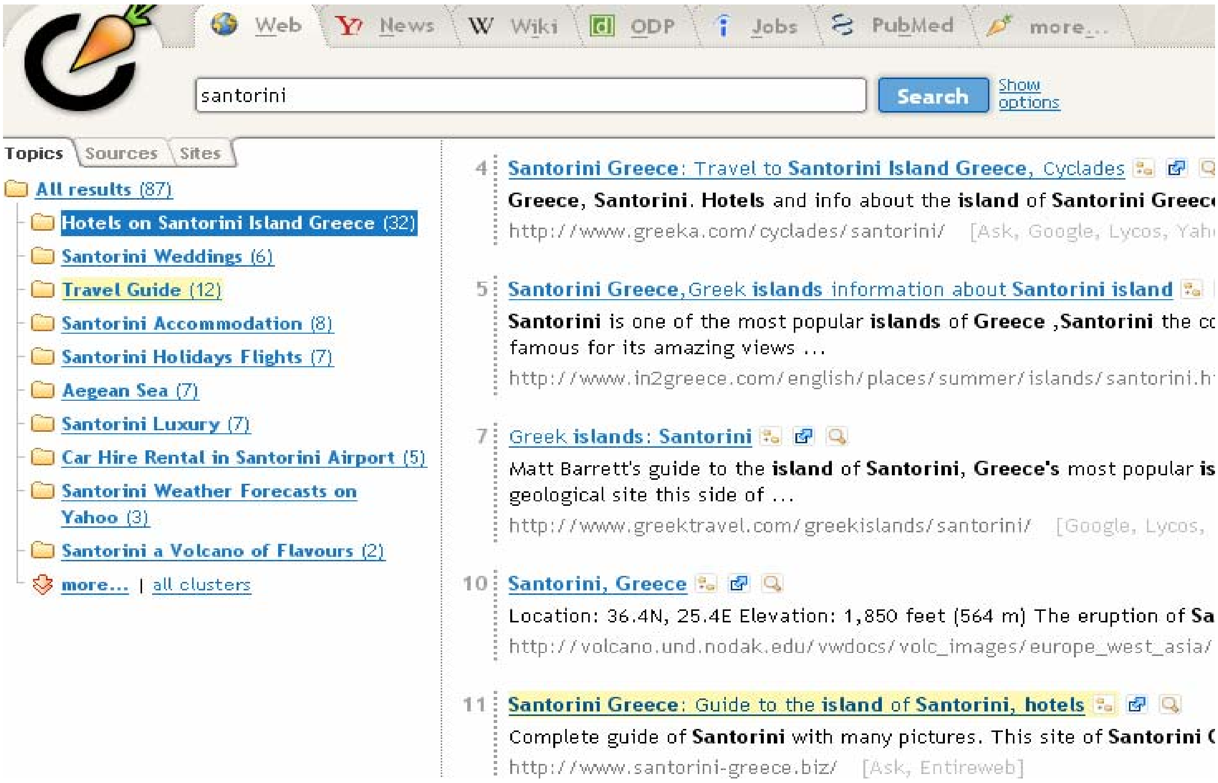
1 Santorini.com - official site - **Santorini** Hotels **Santorini** Hotel ...
It has been said that the sunsets in **Santorini** are the best in the
...
http://www.santorini.com/ [Ask, Entireweb, Google, Lycos, Ms

2 **Santorini** island, Greece: www.santorini.net
Offers general information on where to stay, sightseeing attracti
http://www.santorini.net/ [Ask, Entireweb, Google, Lycos, Yah

3 **Santorini** - Wikipedia, the free encyclopedia
Santorini (Greek Σαντορίνη, pronounced [ˌsa(n)doˈrini]) is a sma
km south ...
http://en.wikipedia.org/wiki/Santorini [Ask, Google, Lycos, Ms

4 **Santorini** Greece: Travel to **Santorini** Island Greece, Cyclades
Greece, **Santorini**. Hotels and info about the island of **Santorini**

Carrot^2 (www.carrot2.org)



All results (87)
- Hotels on Santorini Island Greece (32)
- Santorini Weddings (6)
- Travel Guide (12)
- Santorini Accommodation (8)
- Santorini Holidays Flights (7)
- Aegean Sea (7)
- Santorini Luxury (7)
- Car Hire Rental in Santorini Airport (5)
- Santorini Weather Forecasts on Yahoo (3)
- Santorini a Volcano of Flavours (2)
- more... | all clusters

4 **Santorini Greece**: Travel to **Santorini Island Greece**, Cyclades
Greece, **Santorini**. **Hotels** and info about the **island** of **Santorini Greec**
http://www.greeka.com/cyclades/santorini/ [Ask, Google, Lycos, Yah

5 **Santorini Greece**, Greek **islands** information about **Santorini island**
Santorini is one of the most popular **islands** of **Greece** ,**Santorini** the c
famous for its amazing views ...
http://www.in2greece.com/english/places/summer/islands/santorini.h

7 Greek **islands**: **Santorini**
Matt Barrett's guide to the **island** of **Santorini**, **Greece's** most popular **is**
geological site this side of ...
http://www.greektravel.com/greekislands/santorini/ [Google, Lycos,

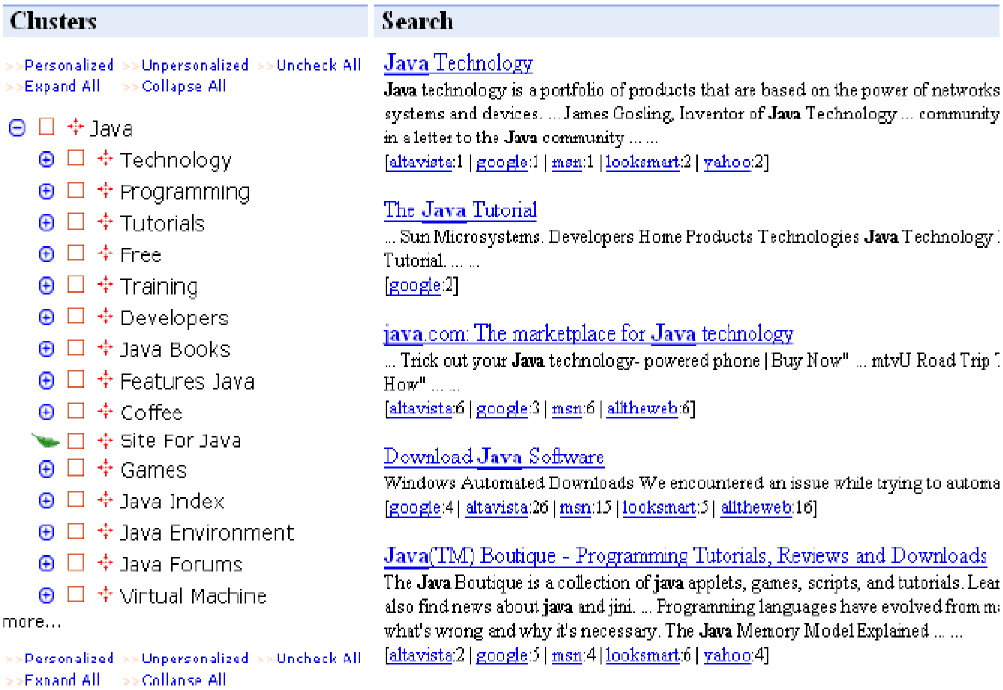
10 **Santorini**, **Greece**
Location: 36.4N, 25.4E Elevation: 1,850 feet (564 m) The eruption of **Sa**
http://volcano.und.nodak.edu/vwdocs/volc_images/europe_west_asia/

11 **Santorini Greece**: Guide to the **island** of **Santorini**, **hotels**
Complete guide of **Santorini** with many pictures. This site of **Santorini** (
http://www.santorini-greece.biz/ [Ask, Entireweb]

SnakeT (http://snaket.di.unipi.it)

- Open-source Web Meta-Search Engine
- Draws about 200 snippets from 16 search engines about Web, Blog, News and Books domain
- Offers both hierarchical clustering and folder labeling with variable-length sentences drawn on-the-fly from snippets
- Use *gapped sentences* as labels, namely sequences of terms occurring not-contiguously into the snippets

SnakeT (http://snaket.di.unipi.it)



Clusters

>>Personalized >>Unpersonalized >>Uncheck All
>>Expand All >>Collapse All

⊖ □ Java
- ⊕ □ Technology
- ⊕ □ Programming
- ⊕ □ Tutorials
- ⊕ □ Free
- ⊕ □ Training
- ⊕ □ Developers
- ⊕ □ Java Books
- ⊕ □ Features Java
- ⊕ □ Coffee
- □ Site For Java
- ⊕ □ Games
- ⊕ □ Java Index
- ⊕ □ Java Environment
- ⊕ □ Java Forums
- ⊕ □ Virtual Machine

more...

>>Personalized >>Unpersonalized >>Uncheck All
>>Expand All >>Collapse All

Search

Java Technology
Java technology is a portfolio of products that are based on the power of networks systems and devices. ... James Gosling, Inventor of **Java** Technology ... community in a letter to the **Java** community ... ...
[altavista:1 | google:1 | msn:1 | looksmart:2 | yahoo:2]

The Java Tutorial
... Sun Microsystems. Developers Home Products Technologies **Java** Technology Tutorial. ... ...
[google:2]

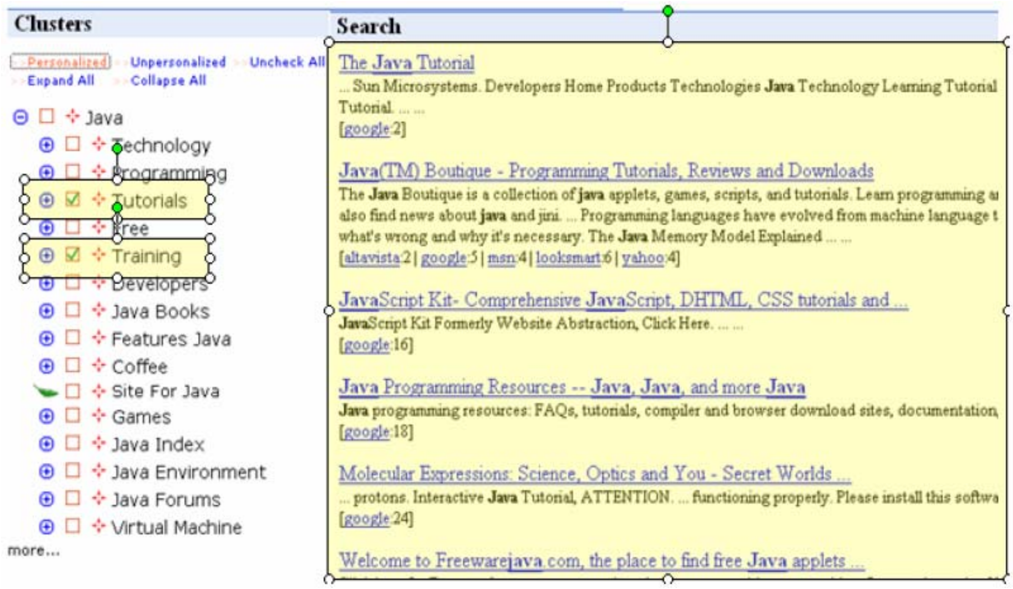
java.com: The marketplace for Java technology
... Trick out your **Java** technology- powered phone | Buy Now" ... mtvU Road Trip How" ... ...
[altavista:6 | google:3 | msn:6 | alltheweb:6]

Download Java Software
Windows Automated Downloads We encountered an issue while trying to automa
[google:4 | altavista:26 | msn:15 | looksmart:5 | alltheweb:16]

Java(TM) Boutique - Programming Tutorials, Reviews and Downloads
The **Java** Boutique is a collection of **java** applets, games, scripts, and tutorials. Lear also find news about **java** and jini. ... Programming languages have evolved from m what's wrong and why it's necessary. The **Java** Memory Model Explained ... ...
[altavista:2 | google:5 | msn:4 | looksmart:6 | yahoo:4]

SNAKET (http://snaket.di.unipi.it)

- Personalized ranking
  - User selects the two labels “Tutorial” and “Training” and gets its personalized ranked list



The Java Tutorial
... Sun Microsystems. Developers Home Products Technologies **Java** Technology Learning Tutorial Tutorial. ... ...
[google:2]

Java(TM) Boutique - Programming Tutorials, Reviews and Downloads
The **Java** Boutique is a collection of **java** applets, games, scripts, and tutorials. Learn programming a also find news about **java** and jini. ... Programming languages have evolved from machine language t what's wrong and why it's necessary. The **Java** Memory Model Explained ... ...
[altavista:2 | google:5 | msn:4 | looksmart:6 | yahoo:4]

JavaScript Kit- Comprehensive JavaScript, DHTML, CSS tutorials and ...
JavaScript Kit Formerly Website Abstraction, Click Here. ... ...
[google:16]

Java Programming Resources -- Java, Java, and more Java
Java programming resources: FAQs, tutorials, compiler and browser download sites, documentation,
[google:18]

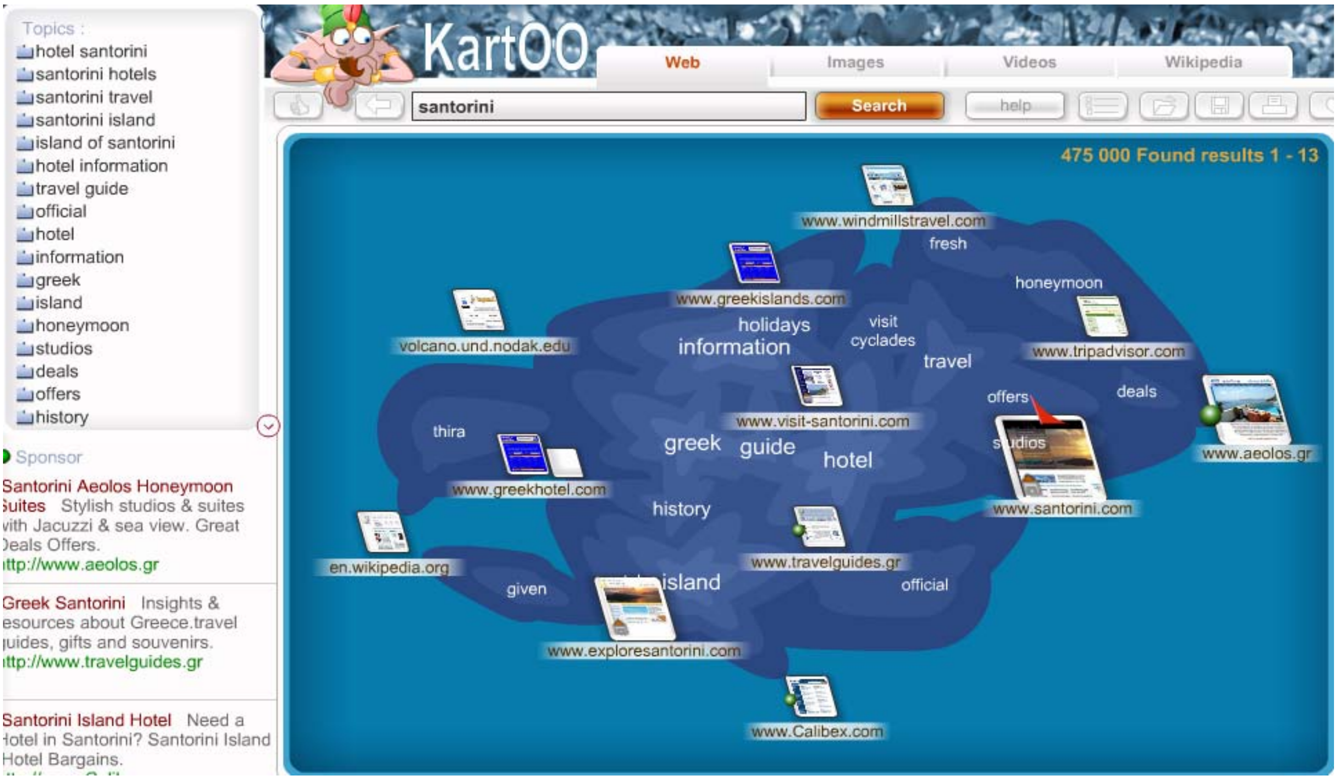
Molecular Expressions: Science, Optics and You - Secret Worlds ...
... protons. Interactive **Java** Tutorial, ATTENTION. ... functioning properly. Please install this softwa
[google:24]

Welcome to Freewarejava.com, the place to find free Java applets ...

Next-generation meta search engines

- Display a visual interface
- Goal:
  - Make easier the visualization of internet & intranet information
  - Help user apprehend huge quantities of information
- KartOO (www.kartoo.com/)
  - Display search results as two-dimensional, interactive maps
  - Map
    - Sites are represented by more or less important size pages, depending on their relevance
    - On mouse over these pages, the concerned keywords are illuminated and a brief description of the site appears on the left side of the screen
  - Queries
    - can follow specific syntax
      - E.g. TEXT : Search on the text of the page as a priority
        LINK : Search a word on the hypertext link
    - follow natural language

KartOO (www.kartoo.com/)



santorini

475 000 Found results 1 - 13

Topics :

- hotel santorini
- santorini hotels
- santorini travel
- santorini island
- island of santorini
- hotel information
- travel guide
- official
- hotel
- information
- greek
- island
- honeymoon
- studios
- deals
- offers
- history

Sponsor

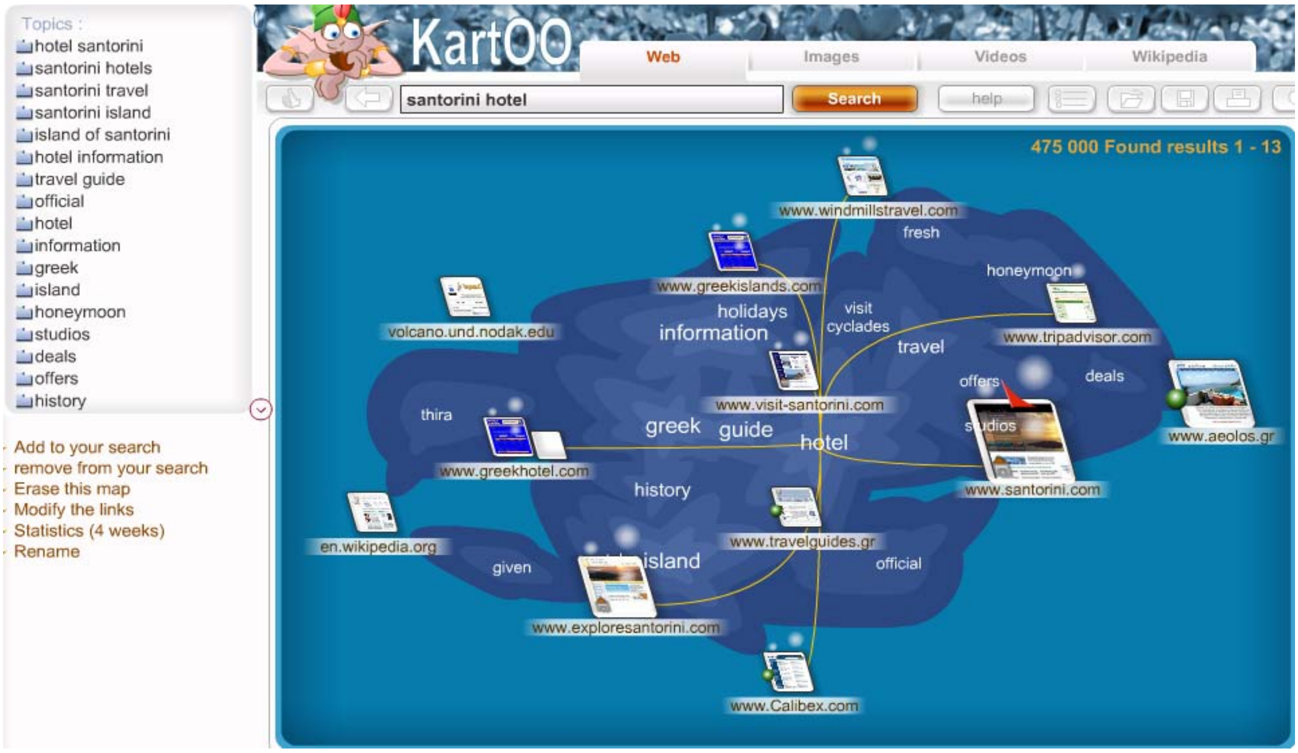
Santorini Aeolos Honeymoon Suites Stylish studios & suites vith Jacuzzi & sea view. Great Deals Offers. ıttp://www.aeolos.gr

Greek Santorini Insights & esources about Greece.travel juides, gifts and souvenirs. ıttp://www.travelguides.gr

Santorini Island Hotel Need a Hotel in Santorini? Santorini Island Hotel Bargains.

KartOO (www.kartoo.com/)

- On mouse over “hotel”



santorini hotel

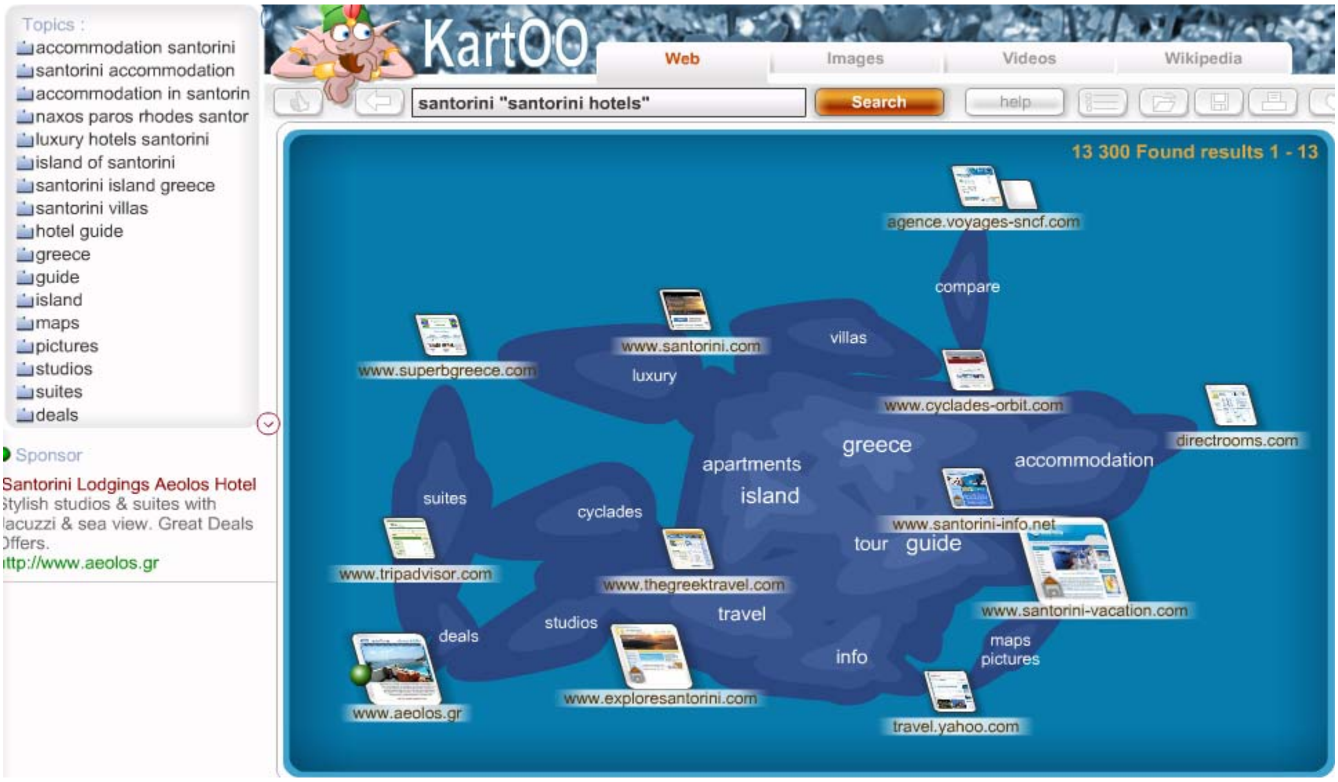
475 000 Found results 1 - 13

Topics :

- hotel santorini
- santorini hotels
- santorini travel
- santorini island
- island of santorini
- hotel information
- travel guide
- official
- hotel
- information
- greek
- island
- honeymoon
- studios
- deals
- offers
- history

- Add to your search
- remove from your search
- Erase this map
- Modify the links
- Statistics (4 weeks)
- Rename

KartOO
(www.kartoo.com/)



santorini "santorini hotels"

Topics :
- accommodation santorini
- santorini accommodation
- accommodation in santorin
- naxos paros rhodes santor
- luxury hotels santorini
- island of santorini
- santorini island greece
- santorini villas
- hotel guide
- greece
- guide
- island
- maps
- pictures
- studios
- suites
- deals

Sponsor

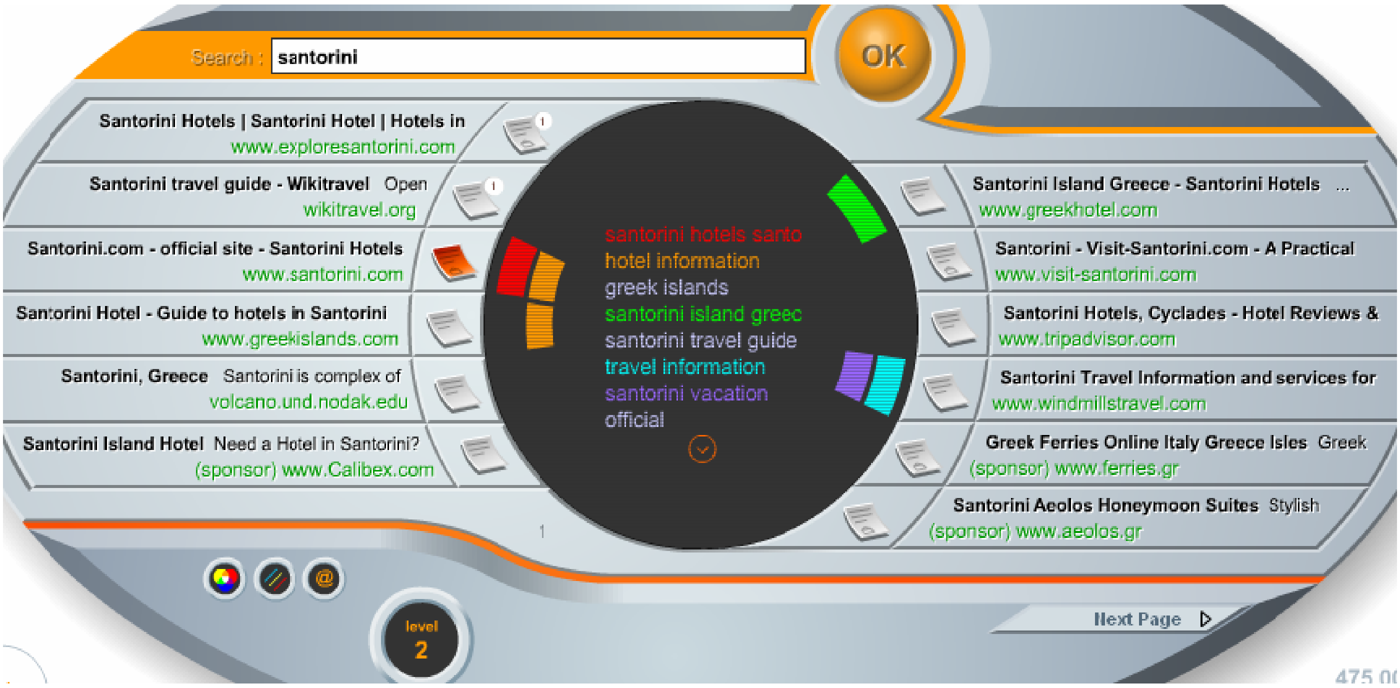
Santorini Lodgings Aeolos Hotel
Stylish studios & suites with Jacuzzi & sea view. Great Deals Offers.
http://www.aeolos.gr

13 300 Found results 1 - 13

UJIKO
(www.ujiko.com/)

- New version of KartOO
- Uses the brand new Yahoo (c) search technology, which indexes more than 5 billion pages.
- Customizable
  - User is free to decide if the website is relevant whether or not by a button “heart” and “trashcan”.
  - Clicking on a site, make it go up automatically in the results list with all the associated sites, which share common topics (wide personalization).
- Separation into levels
  - Each time user visit a new site, he gains one point of expertise.
  - With every 10 points, user move to the next level.
  - New buttons appear giving you access to advanced features (search video, images, news, encyclopedia, advanced filters, animated skins, web archive, traffic details…)
- From level 2, in the center of screen are set of themes are displayed
- Some of these topics are coloured and linked to small bricks with the same color: these indicate which sites are associated with a specific theme.

UJIKO
(www.ujiko.com/)



Search : santorini

Santorini Hotels | Santorini Hotel | Hotels in
www.exploresantorini.com

Santorini travel guide - Wikitravel Open
wikitravel.org

Santorini.com - official site - Santorini Hotels
www.santorini.com

Santorini Hotel - Guide to hotels in Santorini
www.greekislands.com

Santorini, Greece Santorini is complex of
volcano.und.nodak.edu

Santorini Island Hotel Need a Hotel in Santorini?
(sponsor) www.Calibex.com

Santorini Island Greece - Santorini Hotels ...
www.greekhotel.com

Santorini - Visit-Santorini.com - A Practical
www.visit-santorini.com

Santorini Hotels, Cyclades - Hotel Reviews &
www.tripadvisor.com

Santorini Travel Information and services for
www.windmillstravel.com

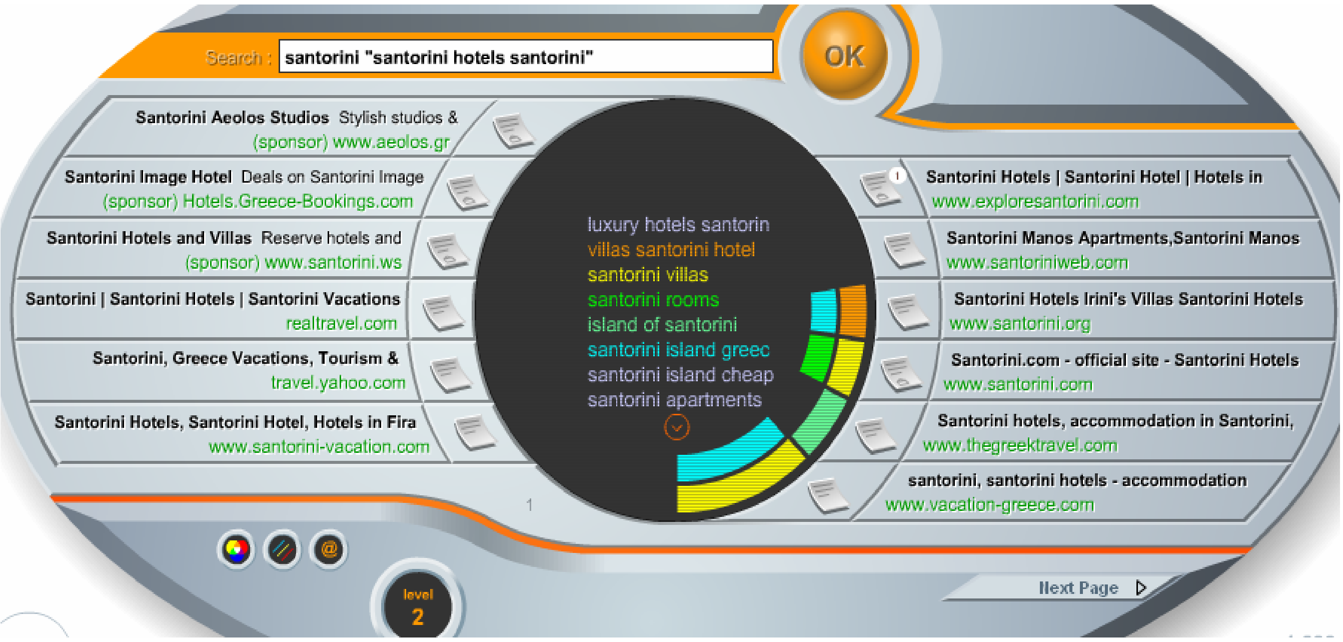
Greek Ferries Online Italy Greece Isles Greek
(sponsor) www.ferries.gr

Santorini Aeolos Honeymoon Suites Stylish
(sponsor) www.aeolos.gr

santorini hotels santo
hotel information
greek islands
santorini island greec
santorini travel guide
travel information
santorini vacation
official

UJIKO
(www.ujiko.com/)

- Refined query



Search : santorini "santorini hotels santorini"

Santorini Aeolos Studios Stylish studios &
(sponsor) www.aeolos.gr

Santorini Image Hotel Deals on Santorini Image
(sponsor) Hotels.Greece-Bookings.com

Santorini Hotels and Villas Reserve hotels and
(sponsor) www.santorini.ws

Santorini | Santorini Hotels | Santorini Vacations
realtravel.com

Santorini, Greece Vacations, Tourism &
travel.yahoo.com

Santorini Hotels, Santorini Hotel, Hotels in Fira
www.santorini-vacation.com

Santorini Hotels | Santorini Hotel | Hotels in
www.exploresantorini.com

Santorini Manos Apartments,Santorini Manos
www.santoriniweb.com

Santorini Hotels Irini's Villas Santorini Hotels
www.santorini.org

Santorini.com - official site - Santorini Hotels
www.santorini.com

Santorini hotels, accommodation in Santorini,
www.thegreektravel.com

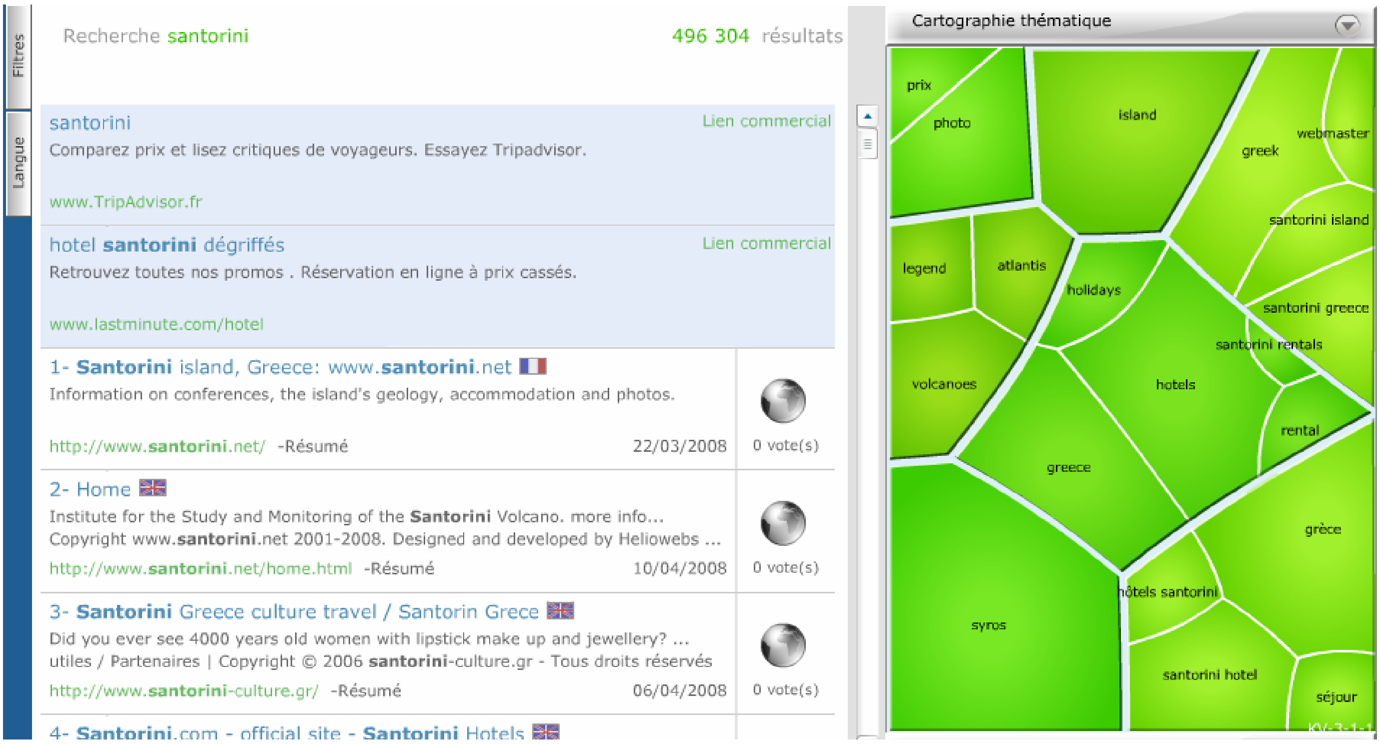
santorini, santorini hotels - accommodation
www.vacation-greece.com

luxury hotels santorin
villas santorini hotel
santorini villas
santorini rooms
island of santorini
santorini island greec
santorini island cheap
santorini apartments

KartOOvisu
(www.kvisu.com/)

- **Display results with a thematic map**
- **Functionalities added**
  - Cartography of topics to filter results
  - Contextual summary generated when click on a topic
  - An integrated history of previous searches

KartOOvisu
(www.kvisu.com/)



Recherche santorini 496 304 résultats

santorini Lien commercial
Comparez prix et lisez critiques de voyageurs. Essayez Tripadvisor.
www.TripAdvisor.fr

hotel santorini dégriffés Lien commercial
Retrouvez toutes nos promos . Réservation en ligne à prix cassés.
www.lastminute.com/hotel

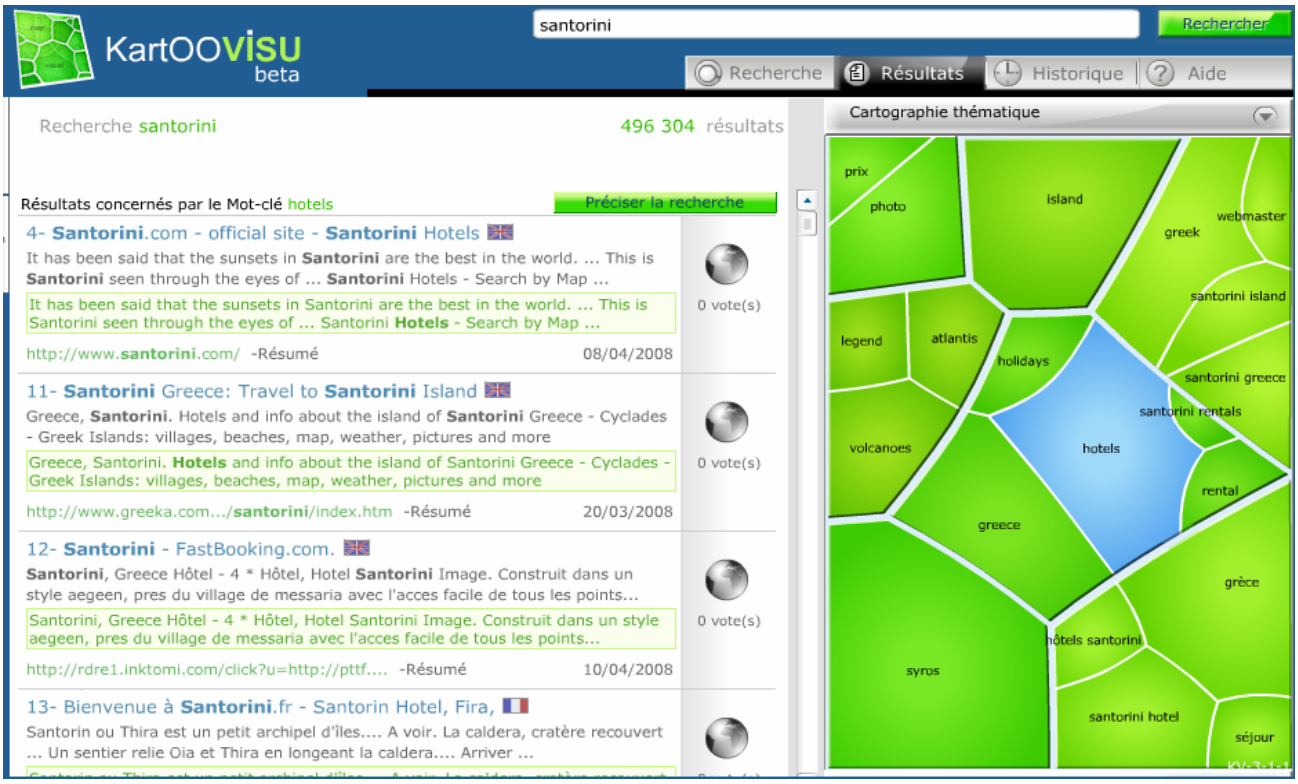
1- Santorini island, Greece: www.santorini.net
Information on conferences, the island's geology, accommodation and photos.
http://www.santorini.net/ -Résumé 22/03/2008 0 vote(s)

2- Home
Institute for the Study and Monitoring of the Santorini Volcano. more info...
Copyright www.santorini.net 2001-2008. Designed and developed by Heliowebs ...
http://www.santorini.net/home.html -Résumé 10/04/2008 0 vote(s)

3- Santorini Greece culture travel / Santorin Grece
Did you ever see 4000 years old women with lipstick make up and jewellery? ...
utiles / Partenaires | Copyright © 2006 santorini-culture.gr - Tous droits réservés
http://www.santorini-culture.gr/ -Résumé 06/04/2008 0 vote(s)

4- Santorini.com - official site - Santorini Hotels

KartOOvisu (www.kvisu.com/)



Recherche santorini 496 304 résultats

Résultats concernés par le Mot-clé hotels

Préciser la recherche

4- **Santorini**.com - official site - **Santorini** Hotels

It has been said that the sunsets in **Santorini** are the best in the world. ... This is **Santorini** seen through the eyes of ... **Santorini** Hotels - Search by Map ...

It has been said that the sunsets in Santorini are the best in the world. ... This is Santorini seen through the eyes of ... Santorini **Hotels** - Search by Map ...

http://www.**santorini**.com/ -Résumé 08/04/2008

0 vote(s)

11- **Santorini** Greece: Travel to **Santorini** Island

Greece, **Santorini**. Hotels and info about the island of **Santorini** Greece - Cyclades - Greek Islands: villages, beaches, map, weather, pictures and more

Greece, Santorini. **Hotels** and info about the island of Santorini Greece - Cyclades - Greek Islands: villages, beaches, map, weather, pictures and more

http://www.greeka.com.../**santorini**/index.htm -Résumé 20/03/2008

0 vote(s)

12- **Santorini** - FastBooking.com.

Santorini, Greece Hôtel - 4 * Hôtel, Hotel **Santorini** Image. Construit dans un style aegeen, pres du village de messaria avec l'acces facile de tous les points...

Santorini, Greece Hôtel - 4 * Hôtel, Hotel Santorini Image. Construit dans un style aegeen, pres du village de messaria avec l'acces facile de tous les points...

http://rdre1.inktomi.com/click?u=http://pttf.... -Résumé 10/04/2008

0 vote(s)

13- Bienvenue à **Santorini**.fr - Santorin Hotel, Fira,

Santorin ou Thira est un petit archipel d'îles.... A voir. La caldera, cratère recouvert ... Un sentier relie Oia et Thira en longeant la caldera.... Arriver ...

Result Clustering of grOOGLE’2007

- Ομαδοποίηση των εγγράφων που εμπεριέχουν τη φράση αναζήτησης με τους αλγορίθμους
  - Kmeans
  - Hierarchical (agglomerative)
- Ορθή συλλογή (φιλτράρισμα) και ονομασία των πιο πάνω αποτελεσμάτων
  - Kmeans έχει επεκταθεί
    - με ένα επιπλέον βήμα, το οποίο δίνει ένα όνομα σε κάθε cluster
    - και με μεθόδους που δημιουργούν ιεραρχίες πάνω σ’ αυτά τα ονόματα
      - Bottom-up Intersection (BU-i)
      - Bottom-up Weighted (BU-w)
      - Top-Down (TD)

Result Clustering of grOOGLE’2007

- Kmeans
  1. Δημιουργείται ένα Label για κάθε Cluster στο οποίο δίνετε ως όνομα αντίστοιχο αριθμό καταχώρισης (1,2,3….K).
  2. Τα N έγγραφα, top-L έγγραφα της τρέχουσας απάντησης, τοποθετούνται τυχαία στα παραπάνω clusters, βάζοντας τα πρώτα N/K στο πρώτο , τα επόμενα τα N/K στο επόμενο κ.ο.
     - στην περίπτωση που υπάρχει modulo (υπόλοιπο) μοιράζονται αντίστοιχα στα παραπάνω Clusters
     - L : παράμετρος με default τιμή 100
  3. Για κάθε ένα i=1,…,K από τα Clusters υπολογίζονται τα Centroids. Μπορεί να είναι:
     - Ο μέσος όρος των βαρών των αντίστοιχων εγγράφων (Centroids).
     - Το έγγραφο με την πλησιέστερη τιμή στον μέσο όρο των βαρών των αντίστοιχων εγγράφων (Memoids).
  4. Για κάθε ένα από τα έγγραφα (documents), αναζητάτε το πιο κοντινό (στην τιμή του βάρους) Centroid (δημιουργία του centroid vector του κάθε cluster)
  5. Τα βήματα 3,4 επαναλαμβάνονται μέχρι όλα τα labels να είναι διαφορετικά μεταξύ τους σε κάθε γύρο.
     - Ο αριθμός των επαναλήψεων δίνεται σαν παράμετρος
     - Ωστόσο, η διαδικασία του αλγορίθμου σταματάει σε περίπτωση που δεν έχουμε μετακίνηση- αλλαγή του αντίστοιχου label
  6. Υπολογίζεται το όνομα και η ιεραρχικότητα (που τυχόν δημιουργείτε) για τα αποτελέσματα.

Result Clustering of grOOGLE’2007

- Μέθοδοι που χρησιμοποιούνται για τη δημιουργία ιεραρχιών
  - Bottom-up Intersection (BU-i)
    - Βασίζεται στην ομοιότητα των όρων μεταξύ των original clusters
    - Αρχικά, οι κόμβοι με τα ονόματα με τη μεγαλύτερη (σε μέγεθος) τομή ομαδοποιούνται δημιουργώντας ένα νέο κόμβο με παιδιά αυτούς τους κόμβους
    - Το όνομα ενός νέου κόμβου είναι η τομή των ονομάτων των παιδιών του
    - Η διαδικασία συνεχίζεται μέχρι να φτάσουμε σε έναν κόμβο
      - Οι κόμβοι που ήδη έχουν γονείς αγνοούνται
  - Bottom-up Weighted (BU-w)
    - Βασίζεται στα βάρη των centroid vectors
    - Αρχικά, γίνεται ταξινόμηση των λέξεων του ονόματος κάθε cluster με βάση το βάρος τους σε φθίνουσα σειρά
    - Στη συνέχεια, τα ονόματα των clusters ταξινομούνται αλφαβητικά
    - Έτσι, τα clusters που έχουν τους ίδιους πιο βεβαρημένους όρους θα τοποθετηθούν διαδοχικά
    - Γίνεται ομαδοποίηση δύο ή περισσότερων clusters κάτω από τον ίδιο κόμβο εάν τα ονόματα τους έχουν κάποιο κοινό πρόθεμα
  - Top-Down (TD)
    - Τα original K clusters θεωρούνται παιδιά του κόμβου root
    - Εφαρμόζεται ξανά ο K-means στα περιεχόμενα του κάθε cluster
    - Η διαδικασία γίνεται αναδρομικά έως το δέντρο να έχει βάθος maxDepth ή το μέγεθος του cluster να είναι μικρότερο από ένα όριο(sz_{mn})

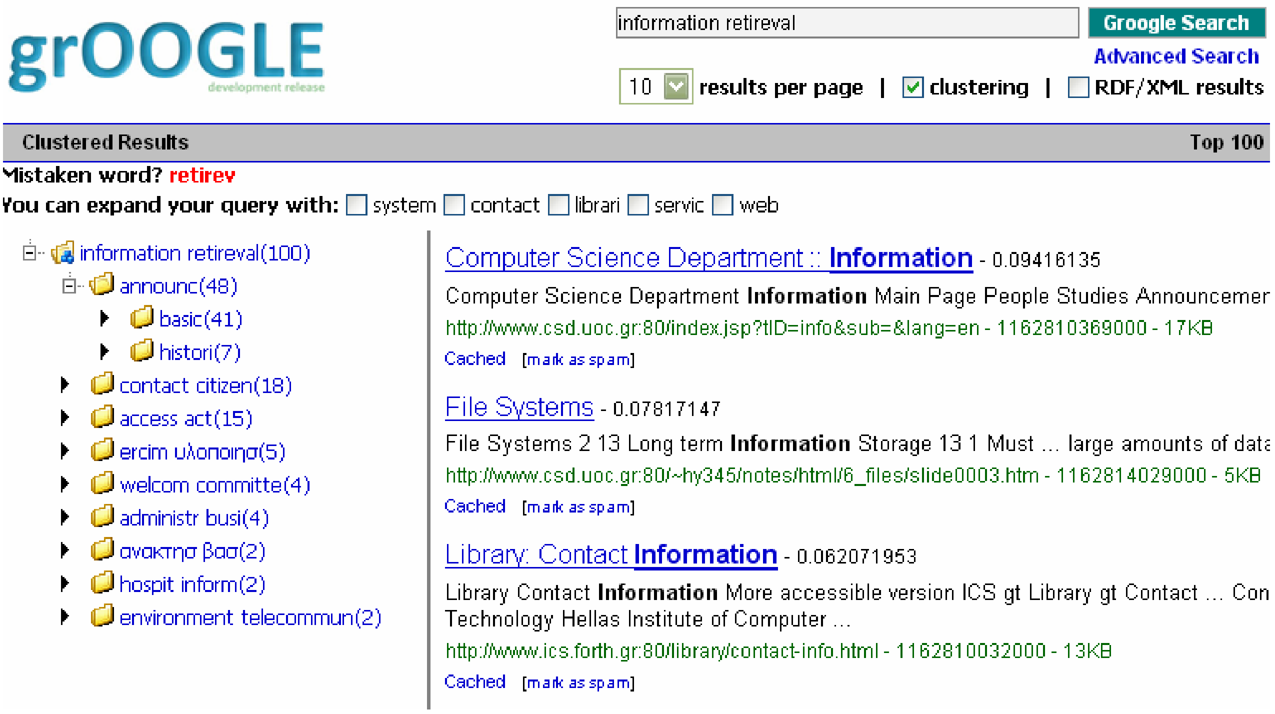
Result Clustering of grOOGLE’2007

- Παραμετροποίηση
  - **K:** αριθμός των αρχικών κέντρων κατανομής του αλγορίθμου, αλλά και εν τέλει το ελάχιστο πλήθος ομάδων που μπορεί να προκύψουν
  - **Max number of docs:** μέγιστο μέγεθος των εγγράφων που μπορεί να γίνει clustered (default 100)
  - **Minimum title:** ελάχιστο μήκος του ονόματος για ένα cluster. Επίσης μπορεί να αλλάξει αρκετά τις τιμές ονοματοδοσίας, αν χρησιμοποιηθεί στην bottom up (InterSection approach)
  - **Max Depth:** μέγιστο βάθος που μπορεί να έχει το δέντρο. Στην περίπτωση του top Down Hierarchy η πραγματική τιμή είναι maxDepth +1 μιας και εφαρμόζουμε τουλ. σε ένα πιο κάτω επίπεδο τον συγκεκριμένο αλγόριθμο
  - **Name hierarchy: οι** μέθοδοι δημιουργίας ιεραρχικότητας των clusters
  - **Max number of words:** ο αριθμός των πιο βεβαρημένων όρων που χρησιμοποιούνται από κάθε έγγραφο
  - **Max Title Length:** μέγιστο πλήθος λέξεων που μπορεί να απαρτίζεται ένα cluster
  - **Min docs in cluster**(sz_{mn})**:** ελάχιστος αριθμός από έγγραφα που θα υπάρχουν σε ένα cluster

- Προβλήματα
  - Στο ευρετήριο του grOOGLE’2007 αποθηκεύονται μόνο οι ρίζες των λέξεων, με αποτέλεσμα τη μείωση της αναγνωσιμότητας των ονομάτων των cluster που δημιουργούνται

Result Clustering of grOOGLE’2007

- BU-w



grOOGLE development release

information retireval | Groogle Search

Advanced Search

10 results per page | ☑ clustering | □ RDF/XML results

Clustered Results Top 100

Mistaken word? retirev

You can expand your query with: □ system □ contact □ librari □ servic □ web

- information retireval(100)
  - announc(48)
    - basic(41)
    - histori(7)
  - contact citizen(18)
  - access act(15)
  - ercim υλοποιησ(5)
  - welcom committe(4)
  - administr busi(4)
  - ανακτησ βασ(2)
  - hospit inform(2)
  - environment telecommun(2)

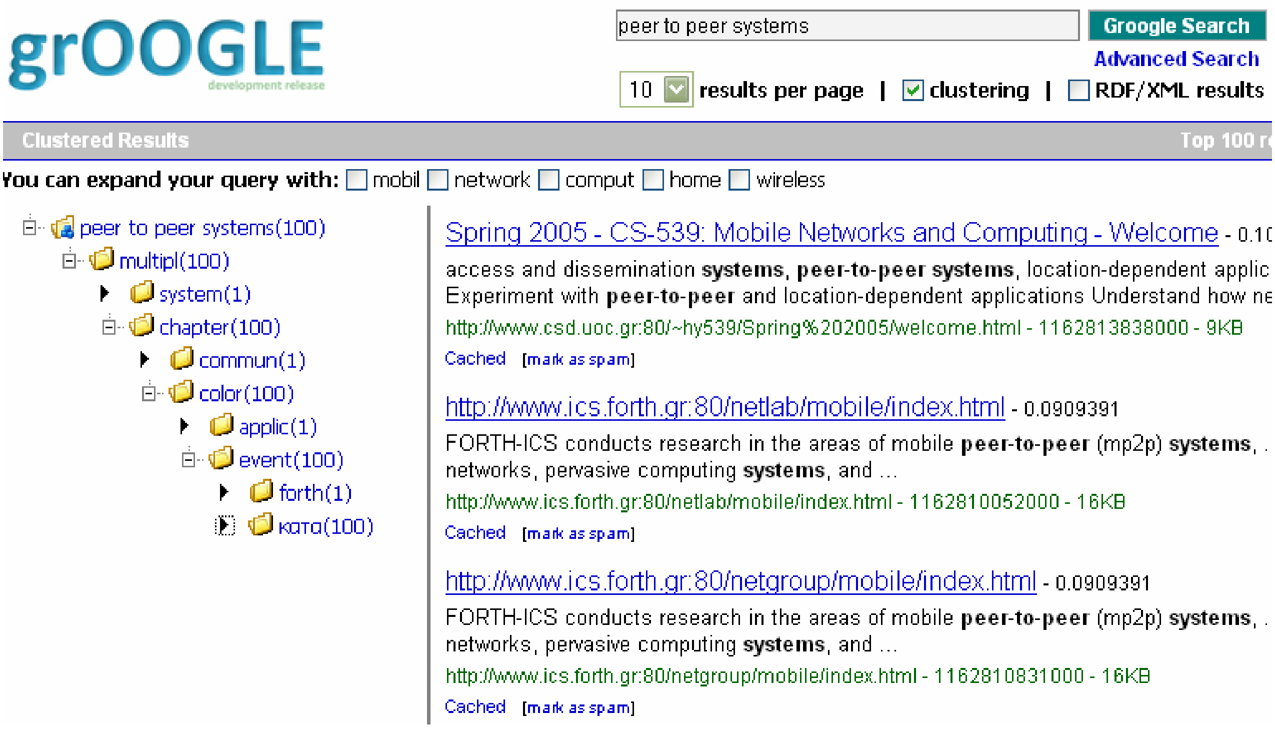
Computer Science Department :: **Information** - 0.09416135
Computer Science Department **Information** Main Page People Studies Announcemer
http://www.csd.uoc.gr:80/index.jsp?tID=info&sub=&lang=en - 1162810369000 - 17KB
Cached [mark as spam]

File Systems - 0.07817147
File Systems 2 13 Long term **Information** Storage 13 1 Must ... large amounts of data
http://www.csd.uoc.gr:80/~hy345/notes/html/6_files/slide0003.htm - 1162814029000 - 5KB
Cached [mark as spam]

Library: Contact **Information** - 0.062071953
Library Contact **Information** More accessible version ICS gt Library gt Contact ... Con
Technology Hellas Institute of Computer ...
http://www.ics.forth.gr:80/library/contact-info.html - 1162810032000 - 13KB
Cached [mark as spam]

Result Clustering of grOOGLE’2007

- Hierarchical



peer to peer systems | Groogle Search

Advanced Search

10 results per page | ☑ clustering | □ RDF/XML results

Clustered Results Top 100 r

You can expand your query with: □ mobil □ network □ comput □ home □ wireless

- peer to peer systems(100)
  - multipl(100)
    - system(1)
    - chapter(100)
      - commun(1)
      - color(100)
        - applic(1)
        - event(100)
          - forth(1)
          - κατα(100)

Spring 2005 - CS-539: Mobile Networks and Computing - Welcome - 0.1(
access and dissemination **systems**, **peer-to-peer systems**, location-dependent applic
Experiment with **peer-to-peer** and location-dependent applications Understand how ne
http://www.csd.uoc.gr:80/~hy539/Spring%202005/welcome.html - 1162813838000 - 9KB
Cached [mark as spam]

http://www.ics.forth.gr:80/netlab/mobile/index.html - 0.0909391
FORTH-ICS conducts research in the areas of mobile **peer-to-peer** (mp2p) **systems**, .
networks, pervasive computing **systems**, and ...
http://www.ics.forth.gr:80/netlab/mobile/index.html - 1162810052000 - 16KB
Cached [mark as spam]

http://www.ics.forth.gr:80/netgroup/mobile/index.html - 0.0909391
FORTH-ICS conducts research in the areas of mobile **peer-to-peer** (mp2p) **systems**, .
networks, pervasive computing **systems**, and ...
http://www.ics.forth.gr:80/netgroup/mobile/index.html - 1162810831000 - 16KB
Cached [mark as spam]

Term Ranking

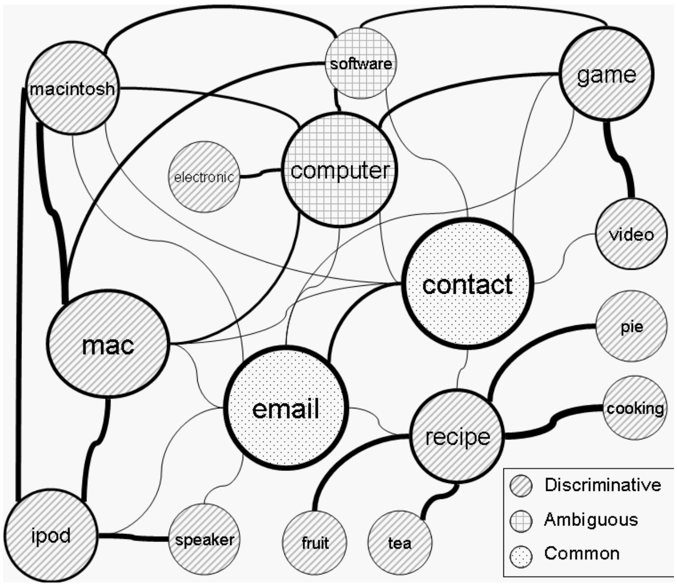
- TermRank [6]
  - variation of PageRank algorithm
  - based on a relational graph representation of the content of web document collections
  - achieves desirable ranking of discriminative terms higher than ambiguous terms, and ranking ambiguous terms higher than common terms
  - Is shown to perform substantially better than frequency based classical methods

Term Rank algorithm

- Not only term frequency based such as TF and TF/IDF, but also considers term-term associations.
- Only the blocks in which the search keyword appear in each Web page are retrieved.
  - Block refers to the text fragments delimited by a set of pre-determined tags such as '<div>','<span>','<table>','<p>','<ul>' and '<ol>.
- Terms are separated into three categories:
  - Discriminative:
    - belong to a specific context are strongly related with a distinct sense of the keyword search term
    - E.g. 'Mac', 'ipod' and 'recipe' – examples from apple data.
  - Ambiguous:
    - have many senses
    - E.g. 'software' and 'computer' appear in both *Computers* and V*ideo games* categories of the 'apple' data.
  - Common:
    - appear in many distinct contexts of a keyword search term
    - E.g. 'email', 'contact', and 'search'.

Term Rank algorithm

- Relational Graph (from Apple data)



Given a relation graph G, TermRank is calculated by:

$$TR(i) = \sum_{j \in \mathcal{N}(i)} \frac{TR(j).w_{ij}}{\sum_{k \in \mathcal{N}(j)} w_{jk}}$$

Iteration 0:

$$TR^{(0)}(i) = \frac{w_i}{\sum_{j \in \mathcal{V}(\mathcal{G})} w_j} = TF(i)$$

Iteration (t+1):

$$TR^{(t+1)}(i) = \sum_{j \in \mathcal{N}(i)} \frac{TR^{(t)}(j).w_{ij}}{\sum_{k \in \mathcal{N}(j)} w_{jk}}$$

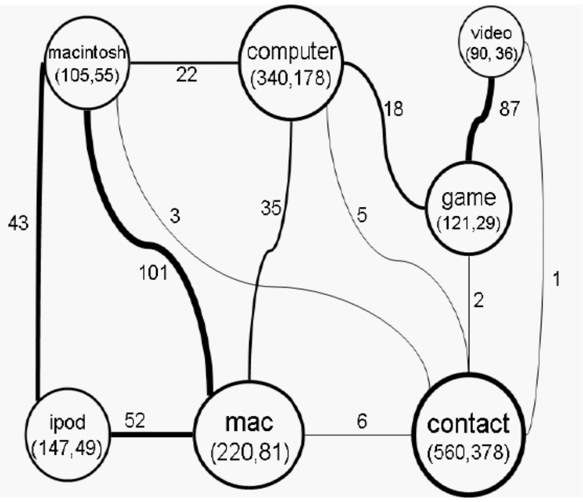
wij: number of times the edge (i,j) appears in the entire data

N(x): set of neighbors of the node x

TermRank runs until the difference between iterations is less than δ which is a small value.

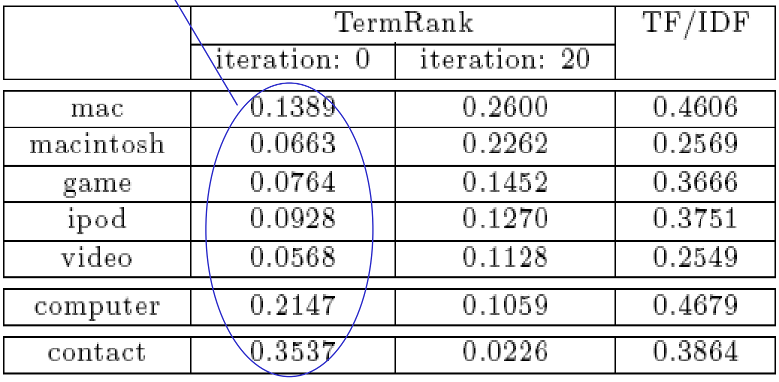
Term Rank algorithm

- Sample run



TF/IDF ranks
computer
mac
contact
ipod
game
macintosh
video

TF



	TermRank		TF/IDF
	iteration: 0	iteration: 20	
mac	0.1389	0.2600	0.4606
macintosh	0.0663	0.2262	0.2569
game	0.0764	0.1452	0.3666
ipod	0.0928	0.1270	0.3751
video	0.0568	0.1128	0.2549
computer	0.2147	0.1059	0.4679
contact	0.3537	0.0226	0.3864

Discriminative →

Ambiguous →

Common →

More clustering algorithms

- Sentences and flat clustering
  - STC [1]
  - Salient phrases extraction [2]
- Single words and flat clustering
  - Sactter/Gather (Buckshot and Fractation algorithms) [3,4]
- Single words and hierarchical clustering
  - Frequent Itemset Hierarchical Clustering (FIHC) [5]

Suffix Tree Clustering (STC)

- Build a suffix tree
  - Incremental
  - Linear time in document collection size
- Treat document as string
  - Use proximity information
  - Vector space model: document is a set of words
- Number of clusters can vary
  - so only the top few clusters are reported
    - typically 10 clusters
- Generate overlapping clusters

Suffix Tree Clustering (STC)

- Step 1 – Document Cleaning
  - Perform stemming
    - delete word prefixes and suffixes
    - reduce plural to singular
  - mark sentence boundaries
    - E.g. Punctuation and HTML tags
  - strip non-word tokens
    - numbers, HTML tags, most punctuation
  - original document strings are kept as pointers from the beginning of each word in the transformed string to its position in the original string

Suffix Tree Clustering (STC)

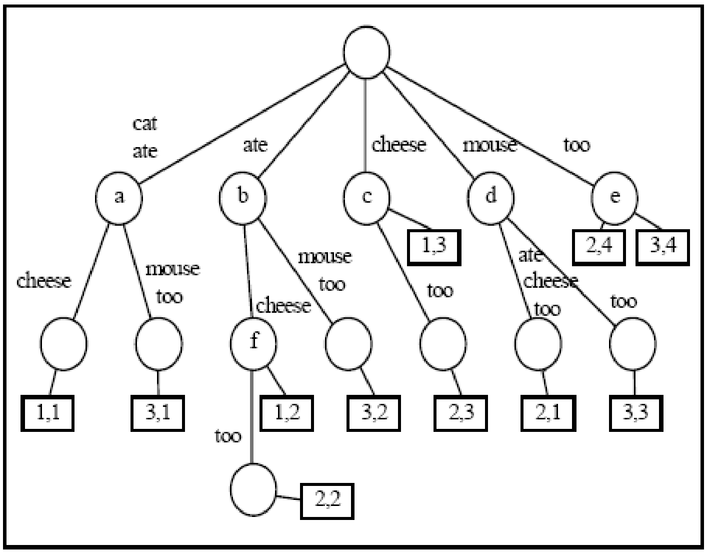
- Step 2 – Identifying base clusters

  Base cluster: a set of documents that share a common phrase.

  - Create a suffix tree in time linear with the size of the collection.
  - Suffix tree of is a compact trie containing all the suffixes of all strings.
  - Documents are treated as strings of words, not characters.
  - Suffixes contain one or more whole words.

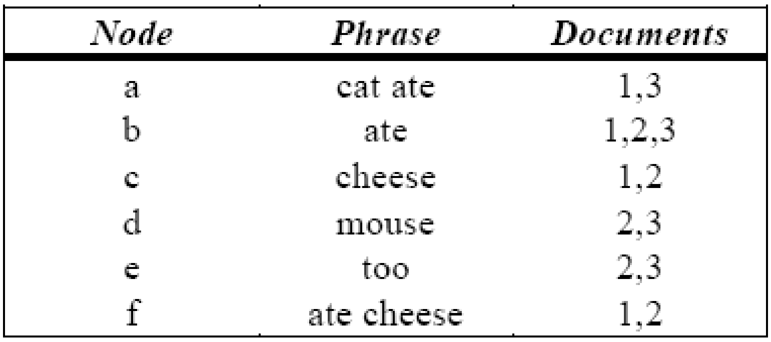
Suffix Tree Clustering (STC)

- Example: The suffix tree of the strings:
  - "cat ate cheese"
  - "mouse ate cheese too"
  - "cat ate mouse too“



Suffix Tree Clustering (STC)

- Each node represents a group of documents and their common phrase
- Six nodes from the example and their corresponding base clusters:



Node	*Phrase*	*Documents*
a	cat ate	1,3
b	ate	1,2,3
c	cheese	1,2
d	mouse	2,3
e	too	2,3
f	ate cheese	1,2

- Each base cluster is assigned a score S(B) = |B| f(|P|)
  - |B|: number of documents in base cluster B
  - |P|: number of words in phrase P
- Zero score is assigned to words appearing in the stop list or in too few (3 or less) or too many (more than 40% of the collection) documents

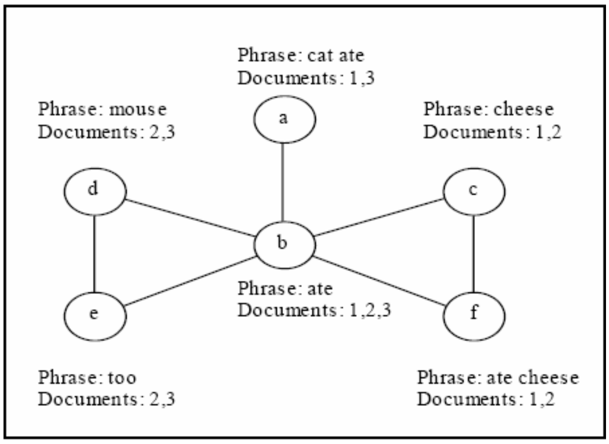
Suffix Tree Clustering (STC)

- Step 3 – Combining Base Clusters
  - Base clusters with a high overlap in their document set are merged.
  - Overlap is identified with a binary similarity measure. Given two base clusters Bm and Bn, similarity of Bm and Bn is 1 iff:

    |Bm∩Bn|/|Bm| > 0.5 and

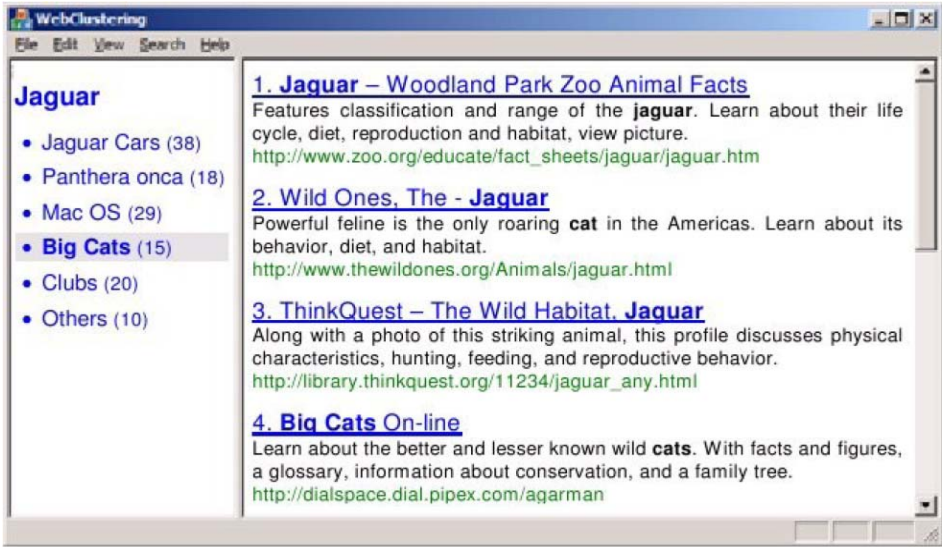
    |Bm∩Bn|/|Bn| > 0.5

    Otherwise, similarity is 0.
  - Base cluster graph of the example. There is one connected component, therefore one cluster.
  - Merge clusters using a single-link clustering algorithm.
    - minimal similarity between base clusters serves as the halting criterion



Salient phrases extraction

- First extracts and ranks salient phrases as candidate cluster names, based on a regression model learned from human labeled training data.
- Documents are assigned to relevant salient phrases to form candidate clusters.
- Final clusters are generated by merging these candidate clusters.



1. Jaguar – Woodland Park Zoo Animal Facts
Features classification and range of the **jaguar**. Learn about their life cycle, diet, reproduction and habitat, view picture.
http://www.zoo.org/educate/fact_sheets/jaguar/jaguar.htm

2. Wild Ones, The - Jaguar
Powerful feline is the only roaring **cat** in the Americas. Learn about its behavior, diet, and habitat.
http://www.thewildones.org/Animals/jaguar.html

3. ThinkQuest – The Wild Habitat, Jaguar
Along with a photo of this striking animal, this profile discusses physical characteristics, hunting, feeding, and reproductive behavior.
http://library.thinkquest.org/11234/jaguar_any.html

4. Big Cats On-line
Learn about the better and lesser known wild **cats**. With facts and figures, a glossary, information about conservation, and a family tree.
http://dialspace.dial.pipex.com/agarman

Salient phrases extraction

- The algorithm is composed of four steps:
  1. Search result fetching
  2. Document parsing and phrase property
  3. Salient phrase ranking
  4. Post-processing

- Search result fetching
  – Get the webpage of search results returned by a certain Web search engine.
  – These web pages are analyzed by an HTML parser and result items are extracted.
  – Each extracted phrase is in fact the name of a candidate cluster.
  – Several properties for each distinct phrase are calculated during parsing.

Salient phrases extraction

- Document parsing and phrase property calculation
  - Titles and snippets can be weighted differently
    - There is a higher probability that salient phrases occur in titles
  - Apply stemming to each word using Porter's algorithm
  - The stop words are included in n-gram generation, so that they could be shown when they are adjacent to meaningful keywords in clusters names.
  - Utilize a regression model to combine these properties into a single salience score.
- Salient phrase ranking
  - The salience phrases are then ranked by the score in descending order.
  - After salient phrases are ranked, the corresponding document lists constitute the candidate clusters, with the salient phrases being cluster names.

Salient phrases extraction

- Post-processing
  - The phrases that contain only stop words or the query words are filtered out.
  - Then, merge the clusters and phrases to reduce duplicated clusters.
  - Specifically, if the overlapped part of two clusters exceeds a certain threshold (75% in experiments), they are merged into one cluster.
  - Cluster names are adjusted according to the new generated cluster.
  - Finally, the top most clusters are shown to user.

Salient phrases extraction

- Salient phrases extraction
  - Denote the current phrase (an n-gram) as w, and the set of documents that contains w as D(w).
  - Five properties which are calculated during the documents parsing.
    - Phrase Frequency / Inverse Document Frequency

$$TFIDF = f(w) \cdot \log \frac{N}{|D(w)|}$$

    - Phrase Length: LEN = n
      - A longer name is preferred for user's browsing
    - Intra-Cluster Similarity
      - First, convert documents into vectors $\mathbf{d}_i = (x_{i1}, x_{i2}, \ldots)$
      - For each candidate cluster, we then calculate its centroid as: $\mathbf{o} = \frac{1}{|D(w)|} \sum_{d_i \in D(w)} \mathbf{d}_i$
      - ICS is calculated as the average cosine similarity between the documents and the centroid

$$ICS = \frac{1}{|D(w)|} \sum_{d_i \in D(w)} \cos(\mathbf{d}_i, \mathbf{o})$$

Salient phrases extraction

- Cluster Entropy:
  - represent the distinctness of a phrase

$$CE = -\sum_t \frac{|D(w) \cap D(t)|}{|D(w)|} \log \frac{|D(w) \cap D(t)|}{|D(w)|}$$

- Phrase Independence
  - a phrase is independent when the entropy of its context is high (i.e., the left and right contexts are random enough).

$$IND_l = -\sum_{t=l(W)} \frac{f(t)}{TF} \log \frac{f(t)}{TF}$$

$$IND = \frac{IND_l + IND_r}{2}$$

- Regression is a classic statistical problem which tries to determine the relationship between two random variables **x** = (*x*1, *x*2, ..., *xp*) and *y*
- Independent variable **x** can be just the vector of the five properties described by **x** = (*TFIDF*, *LEN*, *ICS*, *CE*, *IND*), and dependent *y* can be any real-valued score.

Scatter/Gather

- Scatter/Gather
  - Allows the user to find a set of documents of interest through browsing
  - It iterates:
    - Scatter
      - Take the collection and scatter it into n clusters.
    - Gather
      - Pick the clusters of interest and merge them.
  - Uses non-hierarchical partitioning algorithms:
    - Fractation
      - Create an initial partitioning
    - Buckshot
      - Do on-the-fly clustering to tailor the results from Fractation
- Buckshot
  - fast for online clustering
- Fractionation
  - accurate for offline initial clustering of the entire set

Buckshot and Fractation Algorithm

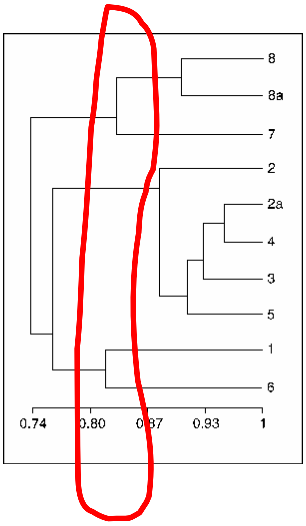
- Seed-based partitional clustering algorithms have three steps:
  1. Find *k* cluster centers.
  2. Assign each document in the collection to the nearest center.
  3. Refine the partitioning.
- Buckshot and Fractation:
  - Different strategies for generating the initial *k* cluster centers from *n* documents
  - Idea:
    - cluster a sample (with slow but high-quality techniques), then assign the entire set

Buckshot and Fractation Algorithm

- Buckshot
  - combines HAC and K-Means clustering.
  - First, randomly take a sample of instances of size $\sqrt{kn}$
  - Run group-average HAC on this sample, which takes only O(*n*) time.
  - Use the results of HAC as initial seeds for K-means.
  - Overall algorithm is O(*kn*)

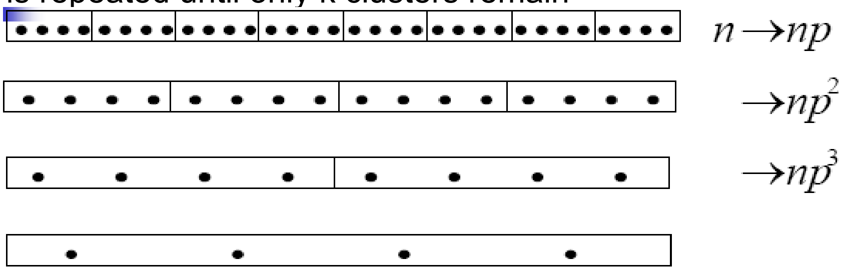
    and avoids problems of bad seed selection.

Cut where You have k clusters



Buckshot and Fractation Algorithm

- Fractation
  - Splits document collections into m buckets (m>k)
  - Clusters each bucket, applying GAC algorithm to each bucket, reducing m to pm where p is the reduction factor(0<p<1)
  - These clusters are treated as the individuals
  - Process is repeated until only k clusters remain



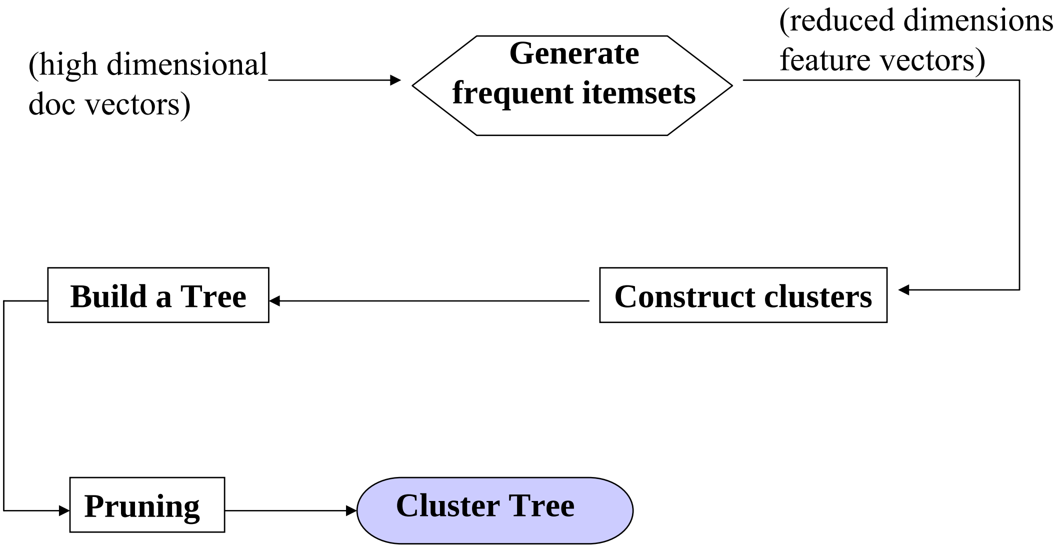
#of buckets (assuming n/m is an integer for simplicity) :

$$\frac{n}{m}+\frac{np}{m}+\frac{np^2}{m}+\cdots=\frac{n}{m}(1+p+p^2+\cdots p^h)=\frac{n}{m}\cdot\frac{1-p^{h+1}}{1-p}=O\left(\frac{n}{m}\right)$$

time complexity : $O(m^2)\times O\left(\frac{n}{m}\right)=O(mn)$ *where* $m>1$ and $0<p<1$

Frequent Itemset Hierarchical Clustering (FIHC)

- overview



(high dimensional doc vectors)

Generate frequent itemsets

(reduced dimensions feature vectors)

Construct clusters

Build a Tree

Pruning

Cluster Tree

Frequent Itemset Hierarchical Clustering (FIHC)

- Definition: Global Frequent Itemset
  - A *global frequent itemset* refers to a set of items (words) that appear together in more than a user-specified fraction of the document set.
  - The *global support* of an itemset is the percentage of documents containing the itemset.

    e.g. 7% of the documents contain both words.

    {apple, window} has global support 7%.
  - A *global frequent item* refers to an item that belongs to some global frequent itemset, e.g., “apple”.
  - A global frequent item is *cluster frequent* in a cluster C_i if the item is contained in some minimum fraction of documents in C_i.

Frequent Itemset Hierarchical Clustering (FIHC)

- Preprocessing
  - Remove stop words
  - Stemming
  - Construct vector model

    doc_i = (item $frequency_1$, if_2, if_3, …, if_m)

    e.g.

    (apple, boy, cat, window)

    doc_1 = (5, 2, 1, 1) ← ***document vector***

    doc_2 = (4, 0, 0, 3)

    doc_3 = (0, 3, 1, 5)

    doc_4 = (8, 0, 2, 0)

    doc_5 = (5, 0, 0, 3)
  - Suppose we set the minimum support to 60%. The global frequent itemsets are:

    {apple}, {cat}, {window}, {apple, window}
  - Store the frequencies only for glob ← ***feature vector*** rder to reduce dimensions.

    (apple, cat, window)

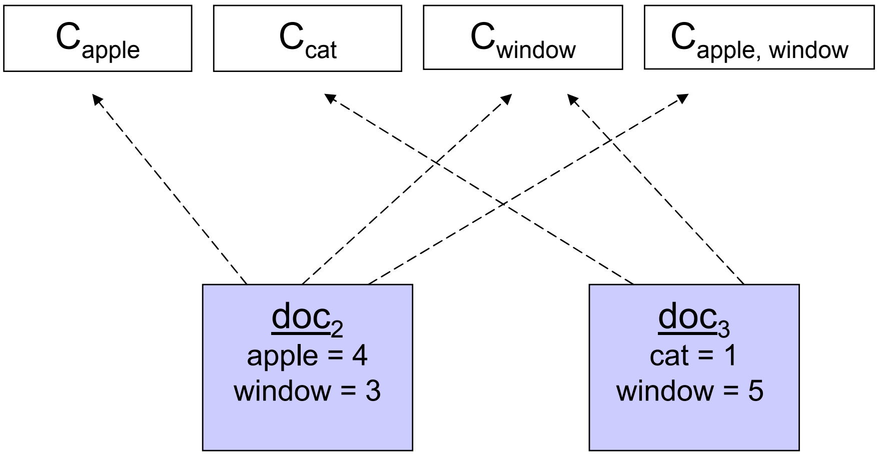
    doc_1 = (5, 1, 1)

    doc_2 = (4, 0, 3)

Frequent Itemset Hierarchical Clustering (FIHC)

- Stage 1 – Construct Clusters
  - Step 1 - Construct Initial Clusters
    - Construct a cluster for each global frequent itemset.

      Global frequent itemsets = {apple}, {cat}, {window}, {apple, window}
    - All documents containing this itemset are included in the same cluster.
    - Linear with respect to the number of documents



C_{apple} C_{cat} C_{window} $C_{apple, window}$

doc_2
apple = 4
window = 3

doc_3
cat = 1
window = 5

Frequent Itemset Hierarchical Clustering (FIHC)

- Step 2 - Making Clusters Disjoint (no-overlapping)
  - Remove docj from all the initial clusters Ci that contain docj but one for which Score(Ci<-docj) is maximazied, and is called “best” initial cluster
  - If there are more than one Ci, choose the one that has the most number of items in the cluster label
  - Intuitively, a cluster C_i is good for a document doc_j if there are many global frequent items in doc_j that appear in many documents in C_i.
    - Goodness of an initial cluster Ci for a document docj is measured by Score(Ci <- docj)

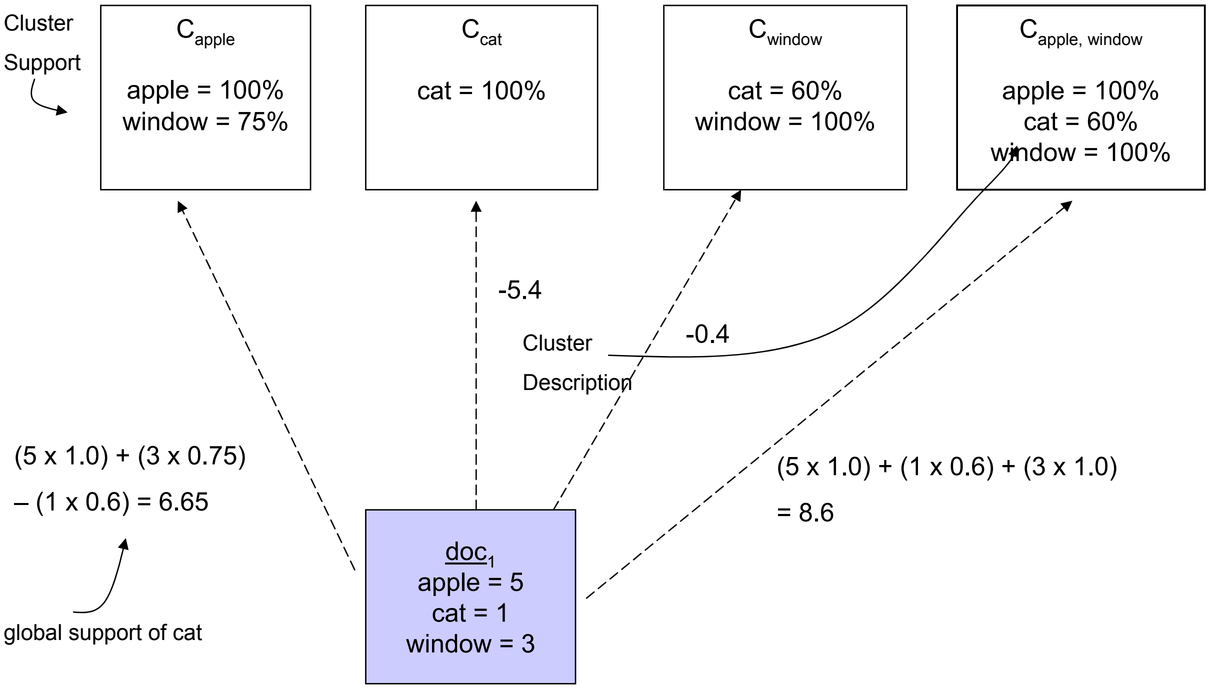
- Score Function
  - Assign each doc_i to the initial cluster C_i that has the highest $score_i$:

$$Score(C_i \leftarrow doc_j) = [\sum_x n(x) * cluster\_support(x)] - [\sum_{x'} n(x') * global\_support(x')]$$

  - x represents a global frequent item in docj and the item is also cluster frequent in Ci
  - x’ represents a global frequent item in docj but the item is not cluster frequent in Ci
  - n(x) is the frequency of x in the feature vector of docj
  - n(x’) is the frequency of x’ in the feature vector of docj

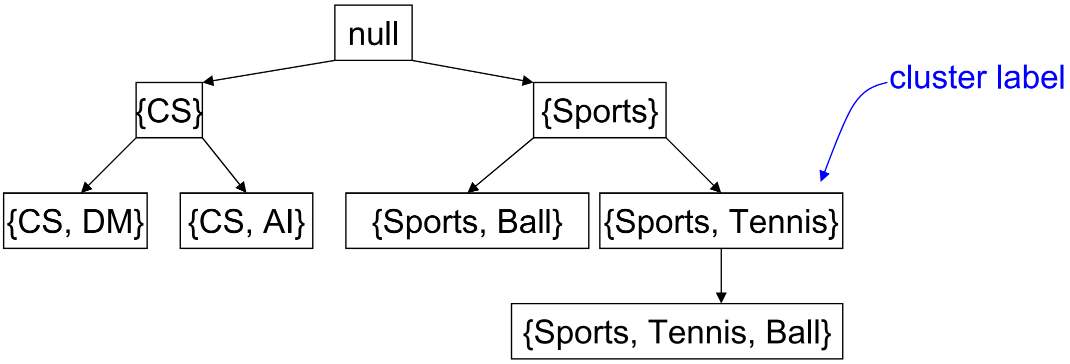
Frequent Itemset Hierarchical Clustering (FIHC)

- Score Function (Example)



Frequent Itemset Hierarchical Clustering (FIHC)

- Stage 2 – Build cluster tree
  - Step 1 - Tree Construction
    - Put the more specific clusters at the bottom of the tree.
    - Put the more general clusters at the top of the tree.
    - Build a tree from bottom-up by choosing a parent for each cluster (start from the cluster with the largest number of items in its cluster label).
    - Depth of the tree is the maximum size of global frequent itemsets.



Frequent Itemset Hierarchical Clustering (FIHC)

- Step 2 - Prune Cluster Tree
  - Merge similar clusters
    - Based on Inter-Cluster Similarity
  - Documents of the same class (topic) are likely to be distributed over different subtrees, which would lead to poor clustering quality.

- Inter-Cluster Similarity
  - *Inter_Sim* of C_a and C_b:

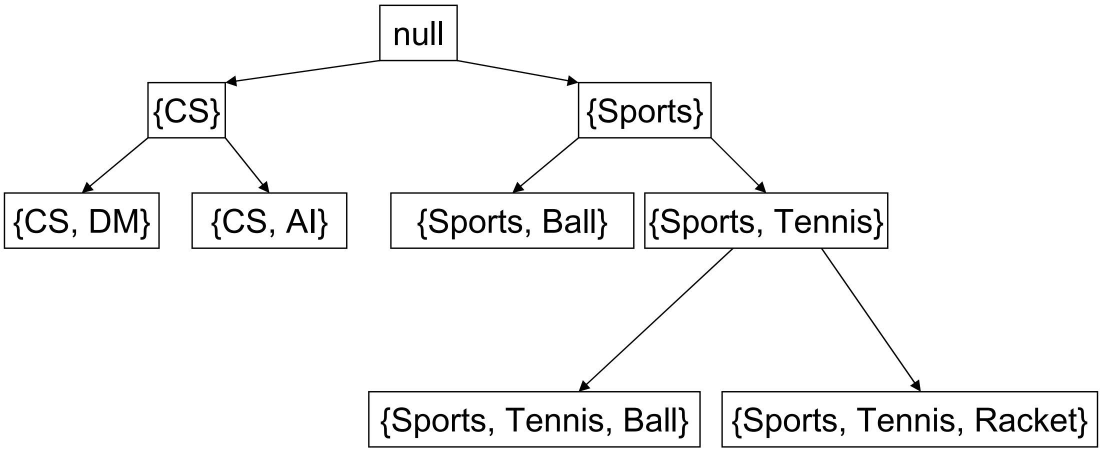
$$Inter\_Sim(C_a \leftrightarrow C_b) = [Sim(C_a \leftarrow C_b) * Sim(C_b \leftarrow C_a)]^{\frac{1}{2}}$$

  - Reuse the score function to calculate Sim(C_i ← C_j).

$$Sim(C_i \leftarrow C_j) = \frac{Score(C_i \leftarrow doc(C_j))}{\sum_x n(x) + \sum_{x'} n(x')} + 1$$

Frequent Itemset Hierarchical Clustering (FIHC)

- Step 3 - Child Pruning
  - Efficiently shorten a tree by replacing child clusters by their parent.
  - A child is pruned only if it is similar to its parent.
  - Prune if Inter_Sim > 1
  - Is applied to level 2 and below, except leaf nodes



null

{CS}

{Sports}

{CS, DM}

{CS, AI}

{Sports, Ball}

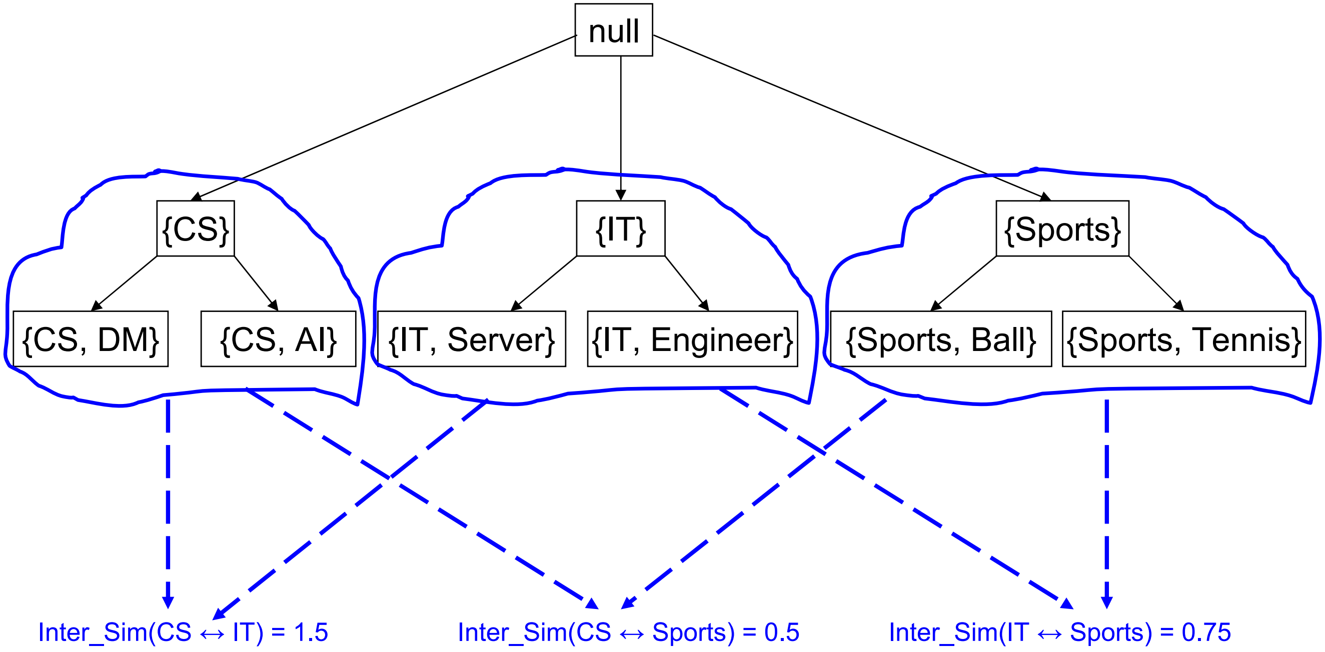
{Sports, Tennis}

{Sports, Tennis, Ball}

{Sports, Tennis, Racket}

Frequent Itemset Hierarchical Clustering (FIHC)

- Step 4 - Sibling Merging
  - Narrow a tree by merging similar subtrees at level 1.



null

{CS}

{IT}

{Sports}

{CS, DM}

{CS, AI}

{IT, Server}

{IT, Engineer}

{Sports, Ball}

{Sports, Tennis}

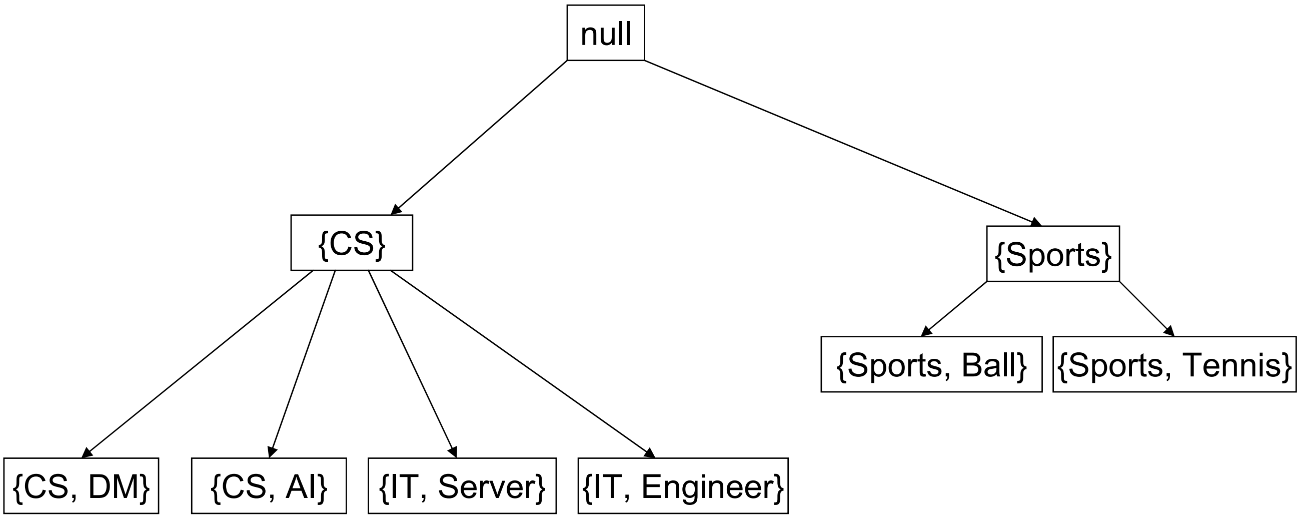
Inter_Sim(CS ↔ IT) = 1.5

Inter_Sim(CS ↔ Sports) = 0.5

Inter_Sim(IT ↔ Sports) = 0.75

Frequent Itemset Hierarchical Clustering (FIHC)

- Step 4 - Sibling Merging



null

{CS}

{Sports}

{Sports, Ball}

{Sports, Tennis}

{CS, DM}

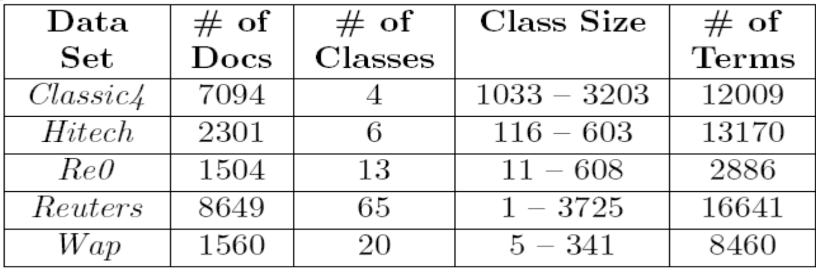
{CS, AI}

{IT, Server}

{IT, Engineer}

Frequent Itemset Hierarchical Clustering (FIHC)

- Data Sets



Data Set	# of Docs	# of Classes	Class Size	# of Terms
Classic4	7094	4	1033 – 3203	12009
Hitech	2301	6	116 – 603	13170
Re0	1504	13	11 – 608	2886
Reuters	8649	65	1 – 3725	16641
Wap	1560	20	5 – 341	8460

  - Each document is pre-classified into a single natural class.

- Evaluation *for natural class Ki and cluster Cj*

$$Recall(K_i, C_j) = \frac{n_{ij}}{|K_i|}$$

$$Precision(K_i, C_j) = \frac{n_{ij}}{|C_j|}$$

•nij : number of members of natural class Ki in cluster Cj

•K : all natural classes

•|D| : total number of documents in the data set

F-measure

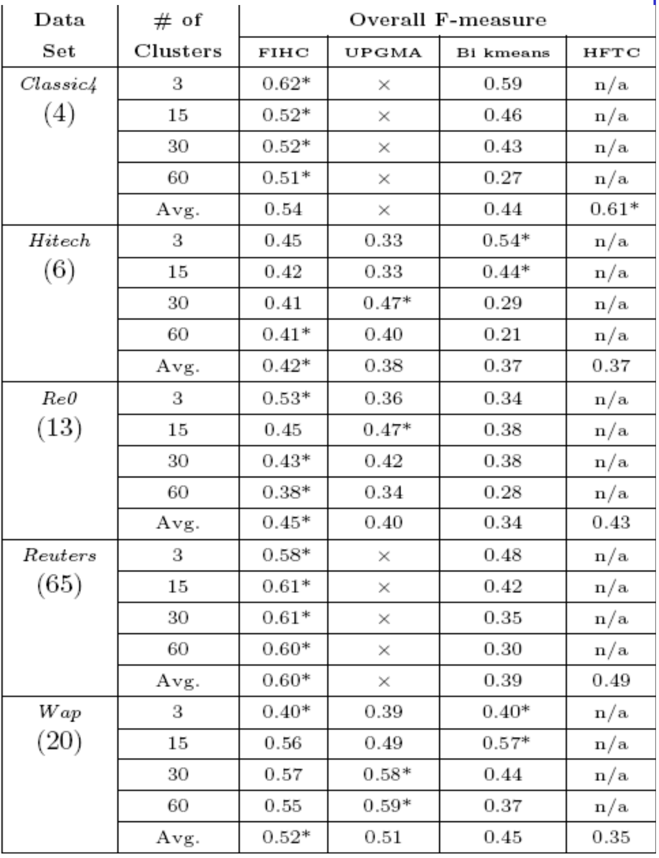
$$F(K_i, C_j) = \frac{2 * Recall(K_i, C_j) * Precision(K_i, C_j)}{Recall(K_i, C_j) + Precision(K_i, C_j)}$$

Overall F-measure

$$F(C) = \sum_{K_i \in K} \frac{|K_i|}{|D|} max_{C_j \in C}\{F(K_i, C_j)\}$$

Frequent Itemset Hierarchical Clustering (FIHC)

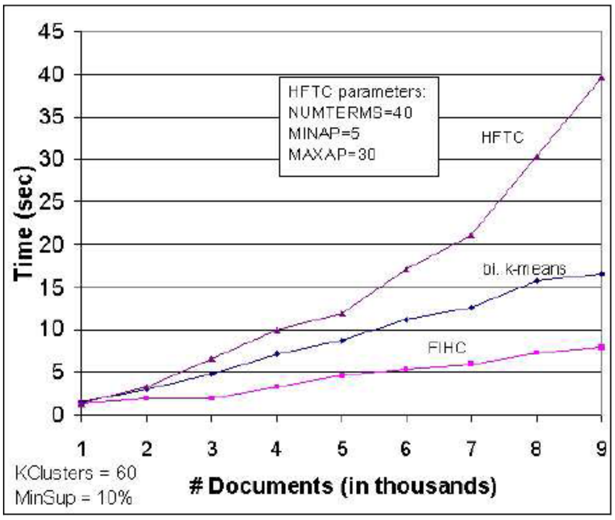
- F-measure comparison of our FIHC method and the other four methods on five data sets
- *x* = not scalable to run
- * = best competitor
- For FIHC and HFTC, we use minimum support from 3% to 6%



Data Set	# of Clusters	Overall F-measure			
		FIHC	UPGMA	Bi kmeans	HFTC
Classic4 (4)	3	0.62*	×	0.59	n/a
	15	0.52*	×	0.46	n/a
	30	0.52*	×	0.43	n/a
	60	0.51*	×	0.27	n/a
	Avg.	0.54	×	0.44	0.61*
Hitech (6)	3	0.45	0.33	0.54*	n/a
	15	0.42	0.33	0.44*	n/a
	30	0.41	0.47*	0.29	n/a
	60	0.41*	0.40	0.21	n/a
	Avg.	0.42*	0.38	0.37	0.37
Re0 (13)	3	0.53*	0.36	0.34	n/a
	15	0.45	0.47*	0.38	n/a
	30	0.43*	0.42	0.38	n/a
	60	0.38*	0.34	0.28	n/a
	Avg.	0.45*	0.40	0.34	0.43
Reuters (65)	3	0.58*	×	0.48	n/a
	15	0.61*	×	0.42	n/a
	30	0.61*	×	0.35	n/a
	60	0.60*	×	0.30	n/a
	Avg.	0.60*	×	0.39	0.49
Wap (20)	3	0.40*	0.39	0.40*	n/a
	15	0.56	0.49	0.57*	n/a
	30	0.57	0.58*	0.44	n/a
	60	0.55	0.59*	0.37	n/a
	Avg.	0.52*	0.51	0.45	0.35

Frequent Itemset Hierarchical Clustering (FIHC)

- Efficiency
  - Comparison on efficiency with the *Reuters* document set
  - UPGMA is excluded from this graph, because it is too slow and not scalable to run.



Frequent Itemset Hierarchical Clustering (FIHC)

- Complexity Analysis
  - Clustering: $\Sigma_{f \in F}$ *global_support(f)*, where f is a global frequent itemset. (two scans on documents)
  - Constructing tree: Removed empty clusters first. $O(n)$, where n is the number of documents.
  - Child pruning: one scan on remaining clusters.
  - Sibling merging: $O(g^2)$, where g is the number of remaining clusters at level 1.

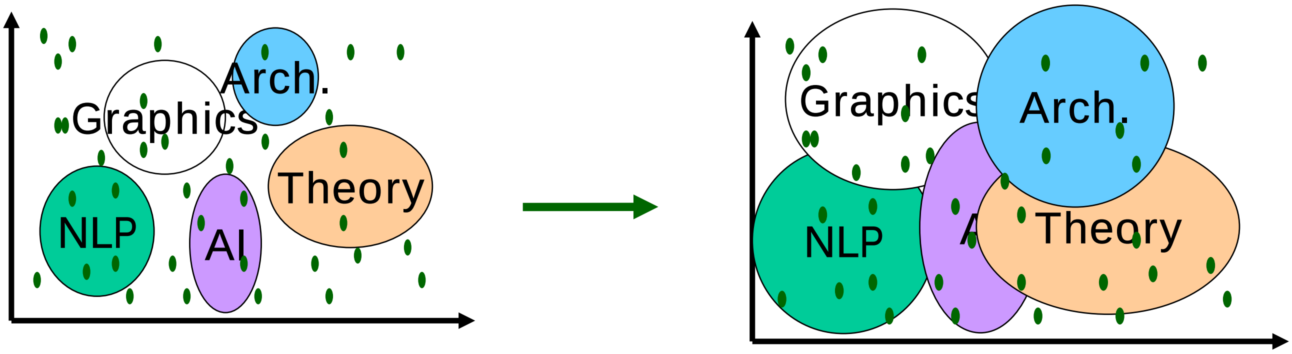
Related Issues

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

References

1. O. Zamir and O. Etzioni, Web document clustering: a feasibility demonstration, in: *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*,1998, pp 46-54.
2. H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster web search results. In ACM SIGIR, pages 210–217, New York, NY, USA, 2004.
3. D. R. Cutting, J. O. Pedersen, D. Karger, and J.W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 318–329, 1992.
4. M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pages 76–84, 1996.
5. Fung, B., Wang, K. & Ester, M. (2003). Hierarchical document clustering using frequent itemsets. In Proceedings of the SIAM International Conference on Data Mining, Cathedral Hill Hotel, San Francisco, CA, May 1-3.
6. F.Gelgi, H.Davulcu and S.Vadrevu, "Term Ranking for Clustering Web Search Results", *10th International Workshop on the Web and Databases (WebDB'07)*, June, 2007, Beijing, China.