Basic Clustering

Abram Hindle

2012-02-28 Tue

Outline

What is cluster analysis?

- grouping objects by similar features
- often unsupervised analysis of a dataset into "natural" groupings
- explorative
- See http://en.wikipedia.org/wiki/Cluster_analysis

Clustering

- Many different techniques
  - group by
    - connectivity
    - centroids: (the centers of a cluster)
    - distributions
    - densities
    - features

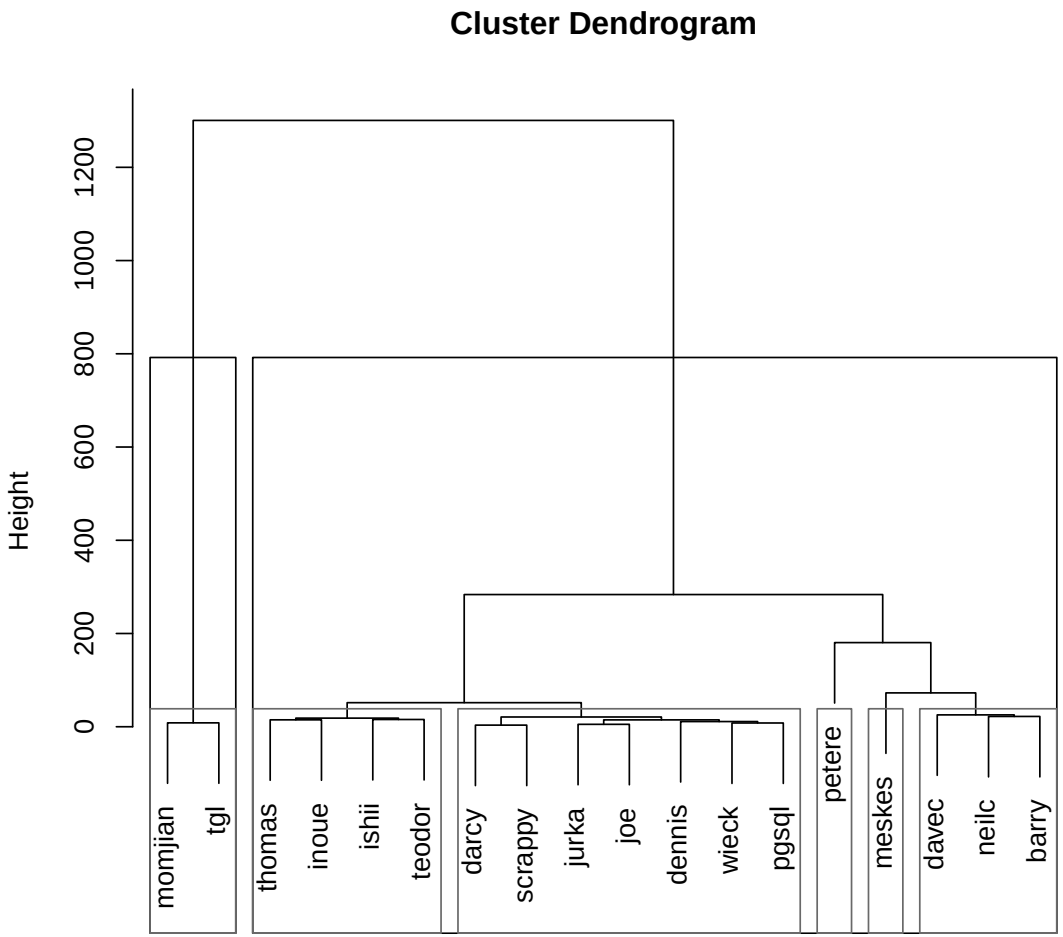
Clustering as labelling?

- Single label
  - Hierarchical clustering
  - K-means
- Multilabel
  - LDA
  - LSI

Hierarchical clustering

- show connectivity and distance
- partner or pair elements and then cluster these pairs
- hierarchical grouping
- the plot is called a Dendrogram
- The distance metric matters
- Distance Metric: Euclidean Distance
  - $\sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$
  - $\sqrt{p} \cdot p$

Hierarchical Clustering Dendrogram:



PostgreSQL Authors
Organized into 2 and 6 clusters

R code

```
v <- read.csv("Author_NFRs.csv"); vv <- v[1:18,]
for (i in 2:17) { vv[,i] <- as.numeric(vv[,i]) }
tv <- t(vv[,2:8])
authors <- matrix(tv,nrow=7,
                  dimnames=list(labels(tv)[[1]],vv[,1]))
pdf("postgresql-author-cluster.pdf")
hc <- hclust(dist(t(authors)),method="ward")
plot(hc,sub="Organized into 2 and 6 clusters",
  xlab="PostgreSQL Authors")
rect.hclust(hc,k=2,border="black")
rect.hclust(hc,k=6,border="dimgrey")
dev.off()
```

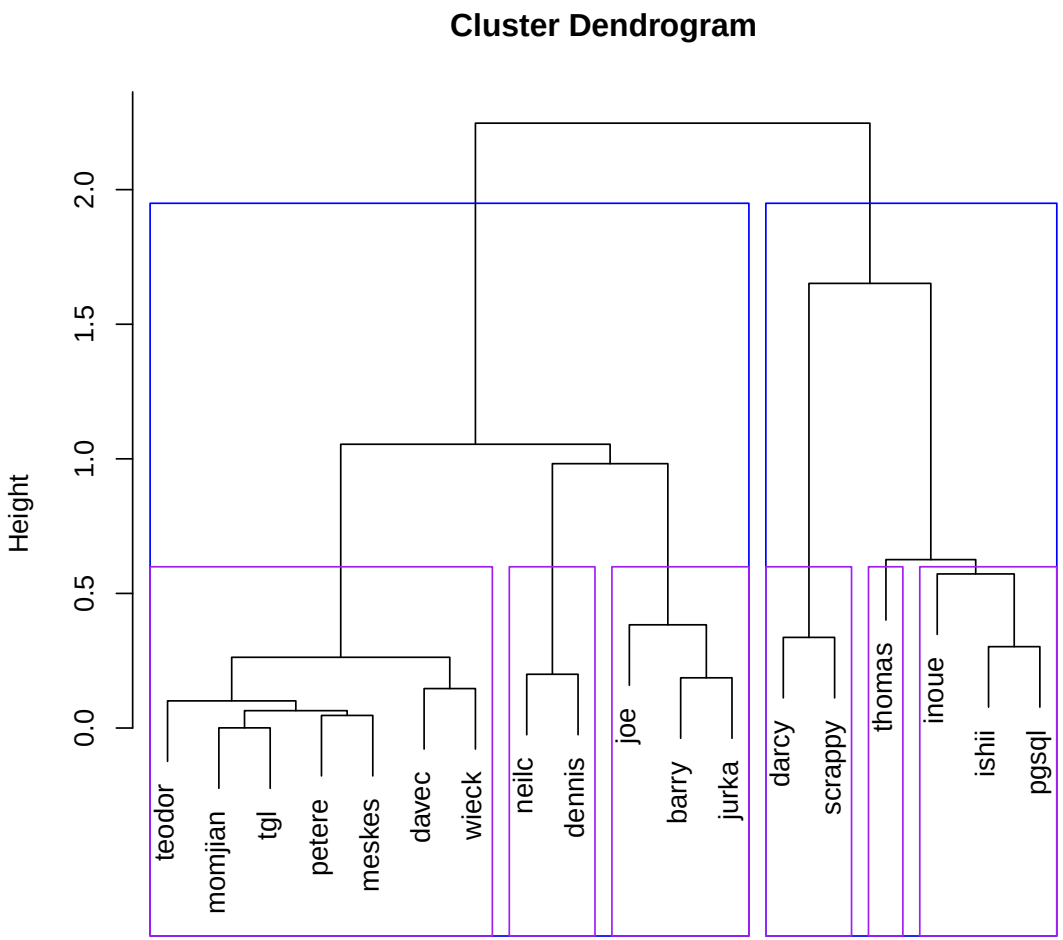
Distance Functions

- Magnitude (Size)
  - Euclidean $\sqrt{p} \cdot p$
  - More so about size in space
  - Less concerned about membership
- Angular (Proportional)
  - Cosine $1 - \frac{A \cdot B}{\|A\|\|B\|}$
  - Correlation - $1 - cor(A, B)$
  - These methods are about content
  - Popular in IR

R code: Pearson Similarity/Distance

```
pdf("oldpostgresql-author-cluster.pdf")
hc <- hclust(as.dist(1-cor(authors)),method="ward")
plot(hc,sub="Organized into 2 and 6 clusters",xlab="Postg
rect.hclust(hc,k=2,border="blue")
rect.hclust(hc,k=6,border="purple")
dev.off()
```

Hierarchical Clustering Dendrogram: Pearson Distance

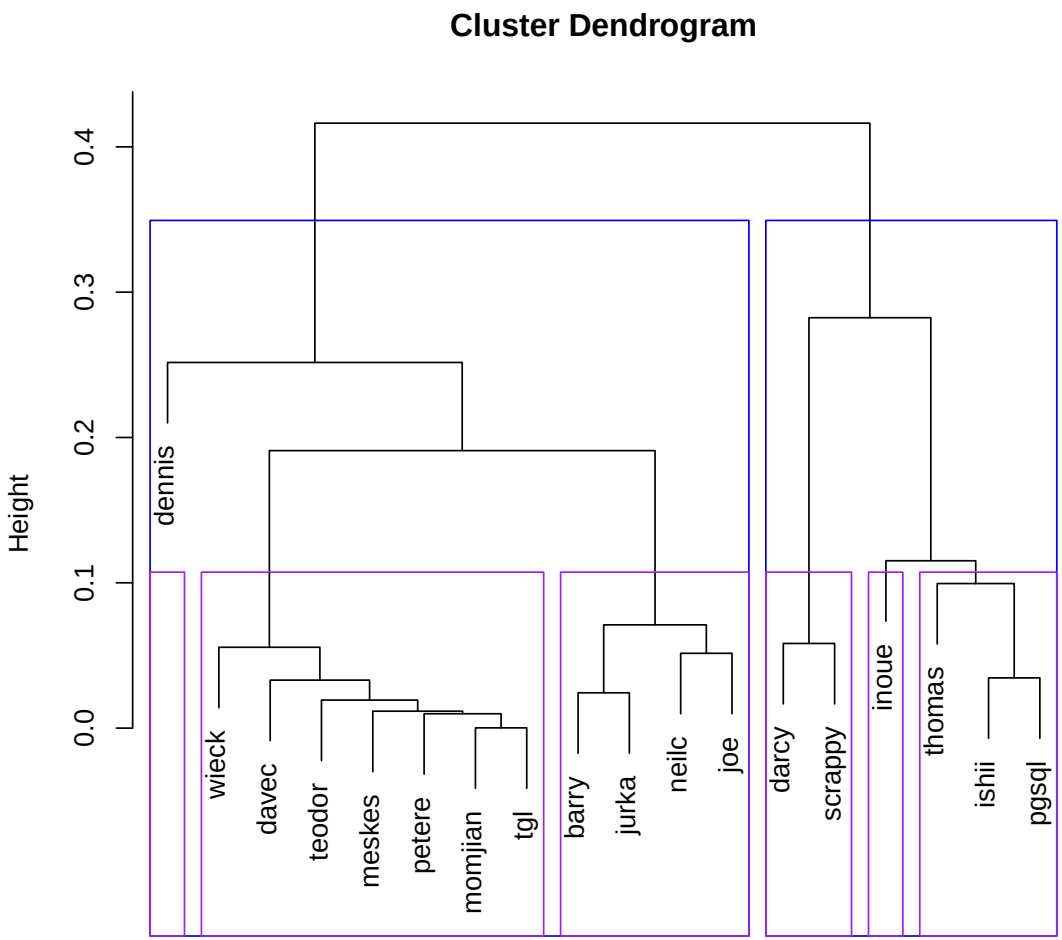


PostgreSQL Authors
Organized into 2 and 6 clusters

R code: Cosine Distance

```
library(lsa)
pdf("postgresql-author-cluster-cosine.pdf")
hc <- hclust(as.dist(1-cosine(authors)),method="ward")
plot(hc,sub="Organized into 2 and 6 clusters",xlab="Postg
rect.hclust(hc,k=2,border="blue")
rect.hclust(hc,k=6,border="purple")
dev.off()
```

Hierarchical Clustering Dendrogram: Cosine Distance



PostgreSQL Authors
Organized into 2 and 6 clusters

Other Clustering Methods

- KMeans
  - centroid based
- DBScan
  - density based

DBScan

- Give each point a radius and then join points who's radius's touch.
- Good for clusters that aren't linearly separable.

DBScan in R

- in R library(fpc) has dbscan
- data.ds = dbscan(data, 0.5)
  - epsilon distance of 0.5
    - warning R's dbscan is slow $O(N^2)$
- data.ds$cluster gives the cluster ID of the element

DBScan in R

- see dbscan.R

KMeans

- Finds K centroids
- reliable and people understand it.
- easy to call in kmeans(data,n) (n is the number of centroids/clusters)

Cluster Stats in R!

- library(fpc)
- cluster.stats computes
  - cluster sizes
  - diameters
  - average distance within and between clusters
  - cluster seperation
  - etc.
- see help(cluster.stats) for references on how to use these tools

Cluster Stats in R

- cluster.stats(distanceMatrix, clustering1)
  - cluster stats of 1 clustering
- cluster.stats(distanceMatrix, clustering1, cluster2)
  - compare clusterings
- see dbscan.R again