COEN 169

Text Clustering

Yi Fang

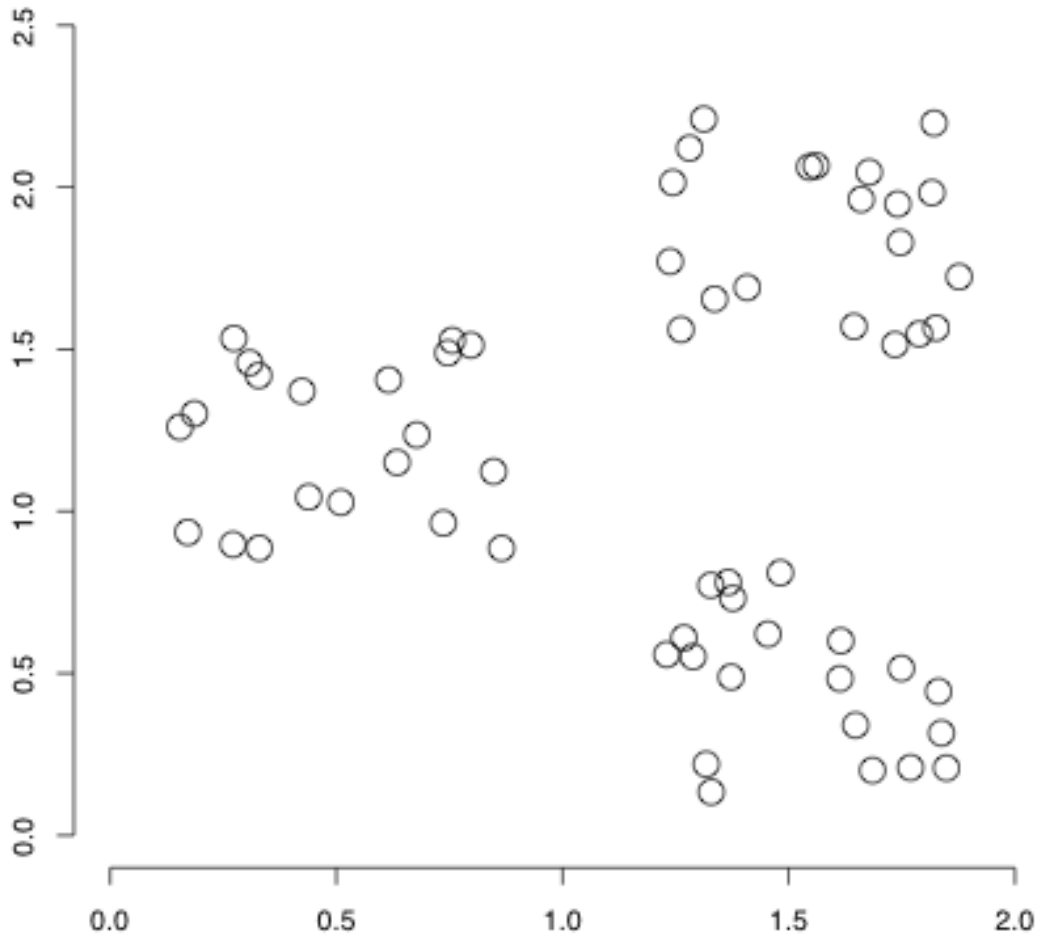
Department of Computer Engineering

Santa Clara University

Definition

- Document clustering is the process of grouping a set of documents into clusters of similar documents.
- Documents within a cluster should be similar.
- Documents from different clusters should be dissimilar.
- Clustering is the most common form of unsupervised learning.
- Unsupervised = there are no labeled or annotated data.
- Classification is the most common form of supervised learning.

Data set with clear cluster structure



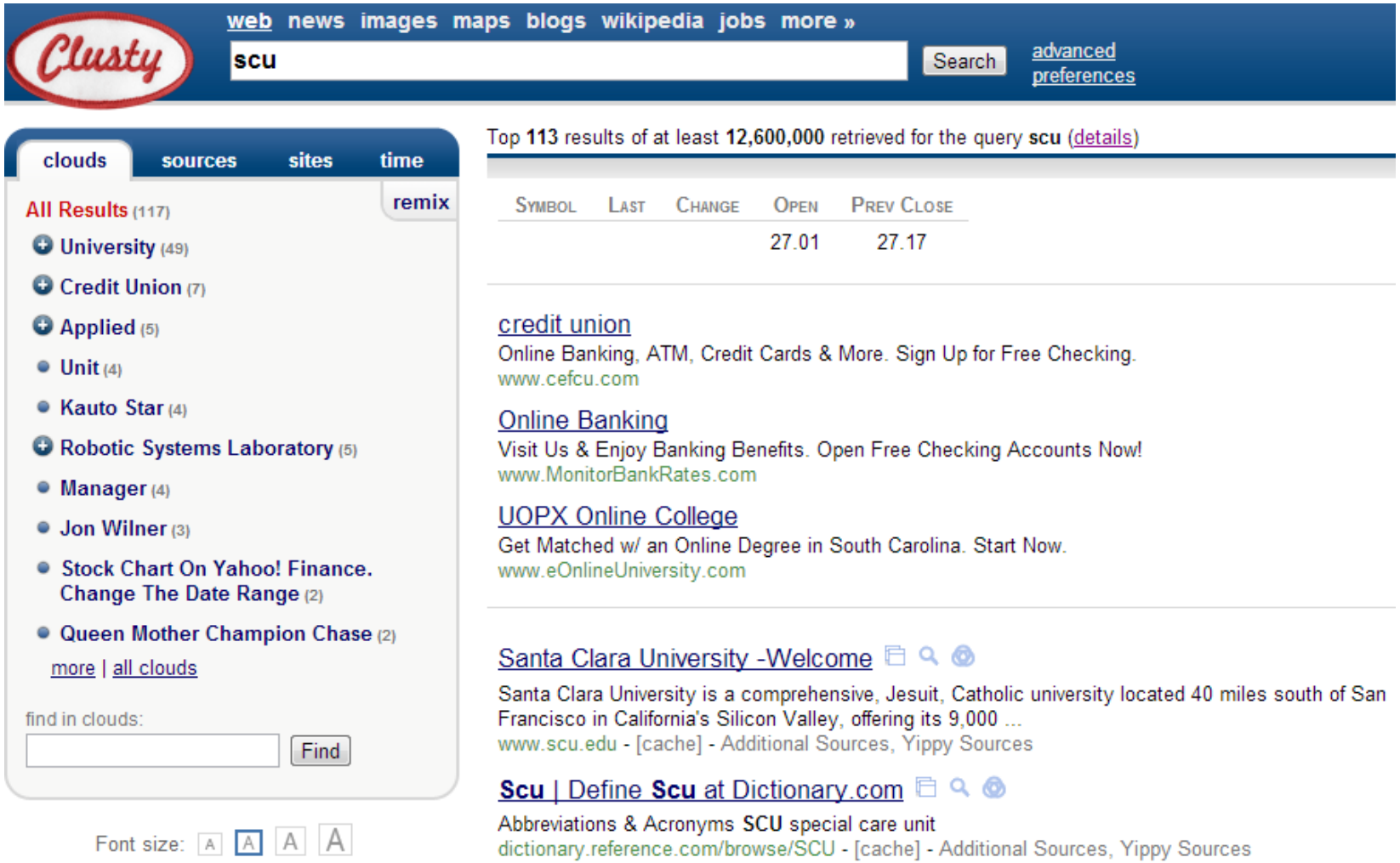
Propose algorithms for finding the cluster structure in this example

Classification vs. Clustering

- Classification: supervised learning
- Clustering: unsupervised learning
- Classification: Classes are human-defined
- Clustering: Clusters are inferred from the data without human input.

- However, there are many ways of influencing the outcome of clustering: number of clusters, similarity measure, representation of documents, . . .

Application I: Search results clustering



Application II: Related articles

Web Images More...

Google

pagerank

Scholar

About 27,500 results (**0.07** sec)

Articles

Legal documents

Any time
Since 2012
Since 2011
Since 2008
Custom range...

Sort by relevance
Sort by date

include patents
include citations

Create alert

The **PageRank** citation ranking: bringing order to the web.
L Page, S Brin, R Motwani, T Winograd - 1999 - ilpubs.stanford.edu
The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes **PageRank**, ...
Cited by 5406 Related articles All 24 versions Cite

Topic-sensitive **pagerank**: A context-sensitive ranking algorithm for web search
TH Haveliwala - ... and Data Engineering, IEEE Transactions on, 2003 - ieeexplore.ieee.org
Abstract The original **PageRank** algorithm for improving the ranking of search-query results computes a single vector, using the link structure of the Web, to capture the relative" importance" of Web pages, independent of any particular search query. To yield more ...
Cited by 1442 Related articles BL Direct All 124 versions Cite

Adaptive methods for the computation of **PageRank**
S Kamvar, T Haveliwala, G Golub - Linear Algebra and its Applications, 2004 - Elsevier
We observe that the convergence patterns of pages in the **PageRank** algorithm have a nonuniform distribution. Specifically, many pages converge to their true **PageRank** quickly, while relatively few pages take a much longer time to converge. Furthermore, we observe ...
Cited by 202 Related articles All 60 versions Cite

Application III: Google News

Web Images Maps Shopping **News** More Search tools

About 5,310 results (0.24 seconds)

Add "**marissa mayer**" section to my Google News homepage

"I like to stay in the rhythm of things," ... "My maternity leave will be a few weeks long, and I'll work throughout it."

Nov 6, 2012 Investing Daily **Marissa Mayer**

Times of India

Yahoo shares reach highest price since 2010 as CEO **Marissa** ...
San Jose Mercury News - 7 hours ago
... as Yahoo Inc. buys back its own stock and more investors bet on CEO **Marissa Mayer's** ability to turn around the long-struggling company.

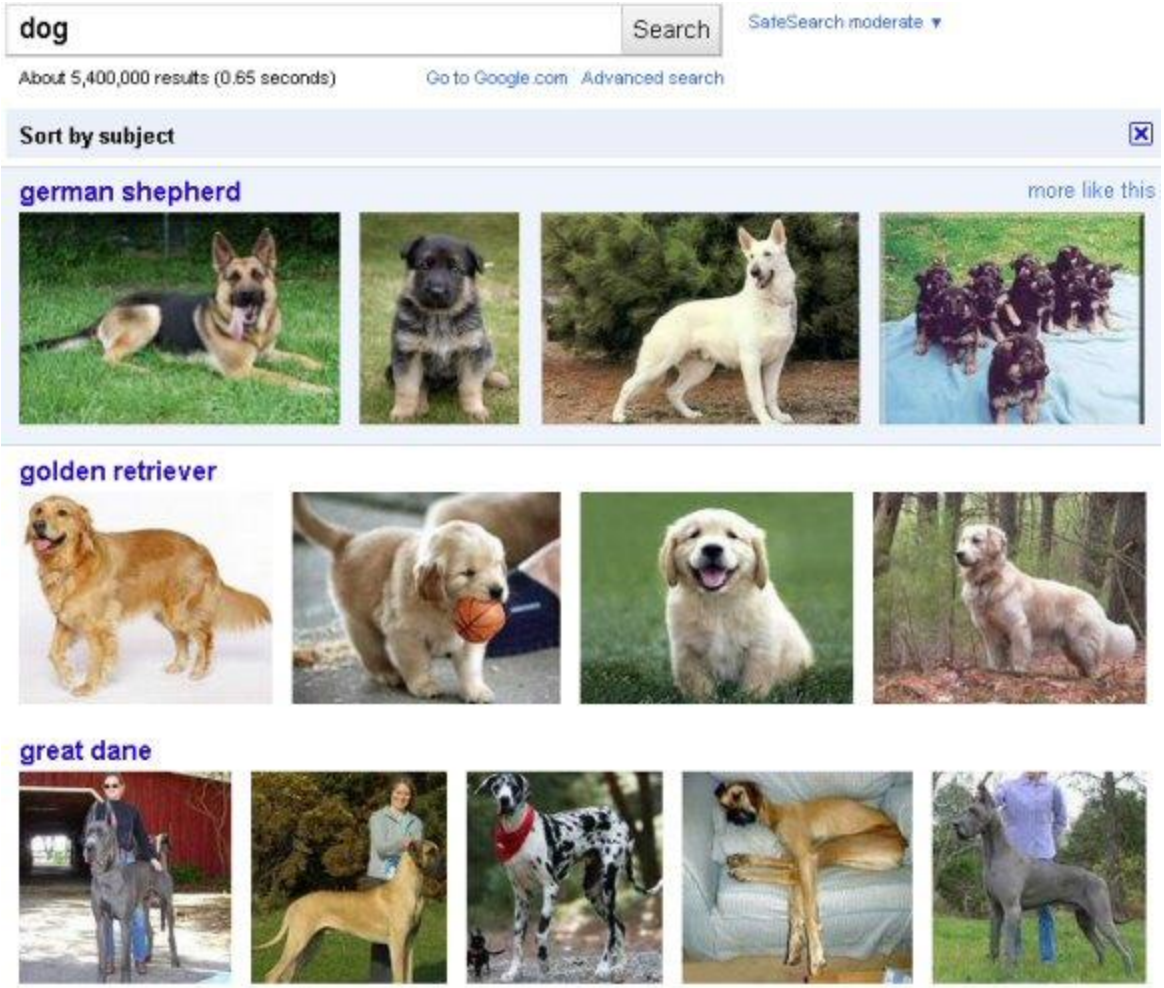
Investors cheer Yahoo for CEO **Marissa Mayer**
USA TODAY - 6 hours ago
Outsider or Insider CEO: Why Yahoo and Citigroup Leaders Have ...
Huffington Post (blog) - 7 hours ago
all 157 news sources »

ValueWalk

Marissa Mayer's Yahoo Was Just Made A Goldman Sachs ...
San Francisco Chronicle - by Nicholas Carlson - 13 hours ago
Marissa Mayer's Yahoo Was Just Made A Goldman Sachs 'Conviction Buy' (YHOO, GS) ... Related: **Marissa Mayer**'s Plan To Shrink Yahoo ...
+ Show more

The **Marissa Mayer** effect? Yahoo shares hit highest point in more ...
San Jose Mercury News - by Alexei Oreskovic - Nov 19, 2012
The **Marissa Mayer** effect? ... a half, as investor confidence grows that new CEO **Marissa Mayer** can

Application IIII: Image clustering



K-means

- Perhaps the best known clustering algorithm
- Simple, works well in many cases
- Use as default / baseline for clustering documents

Document representations in clustering

- Vector space model
- As in vector space classification, we measure relatedness between vectors by Euclidean distance . . .
- which is almost equivalent to cosine similarity.

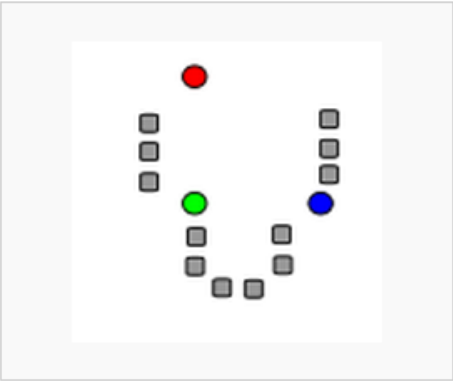
K-Means

- An iterative algorithm
- Clusters based on centroids (aka the mean) of points in a cluster, *c*:

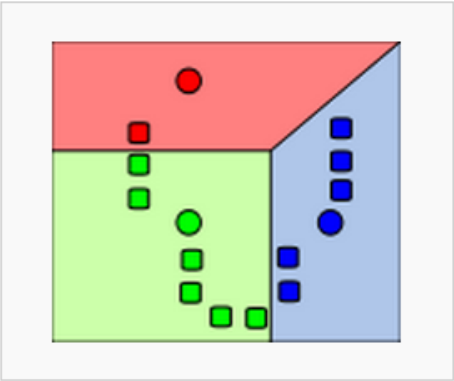
$$\vec{\mu}(c) = \frac{1}{|c|}\sum_{\vec{x}\in c}\vec{x}$$

- Reassignment of documents to clusters is based on the nearest distance to the current cluster centroids
- Recompute the centroids of the clusters based on the new membership of documents

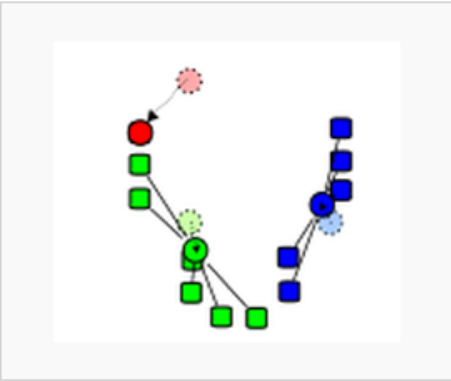
An example



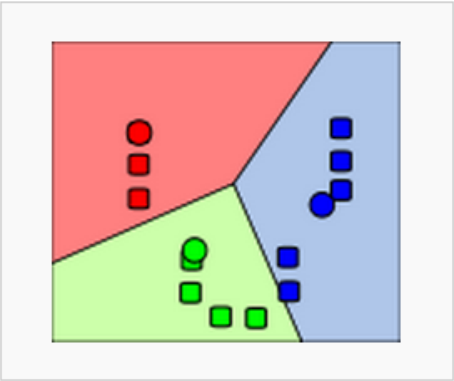
1) *k* initial "means" (in this case *k*=3) are randomly generated.



2) *k* clusters are created by associating every observation with the nearest mean.



3) The centroid of each of the *k* clusters becomes the new mean.



4) Steps 2 and 3 are repeated until convergence has been reached.

K-Means Algorithm

Select *K* random docs $\{s_1, s_2, \ldots s_K\}$ as seeds.

Until clustering *converges*:

For each doc d_i:

Assign d_i to the cluster c_j such that $dist(x_i, s_j)$ is minimal.

For each cluster c_j

$s_j = \mu(c_j)$

Reassignment of cluster membership

Recomputation of cluster centroids

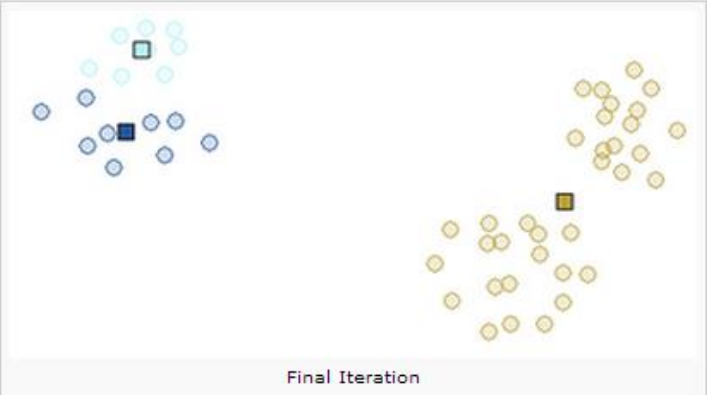
K-means is guaranteed to converge

- But we don't know how long convergence will take!
- If we don't care about a few docs switching back and forth, then convergence is usually fast (< 10-20 iterations).

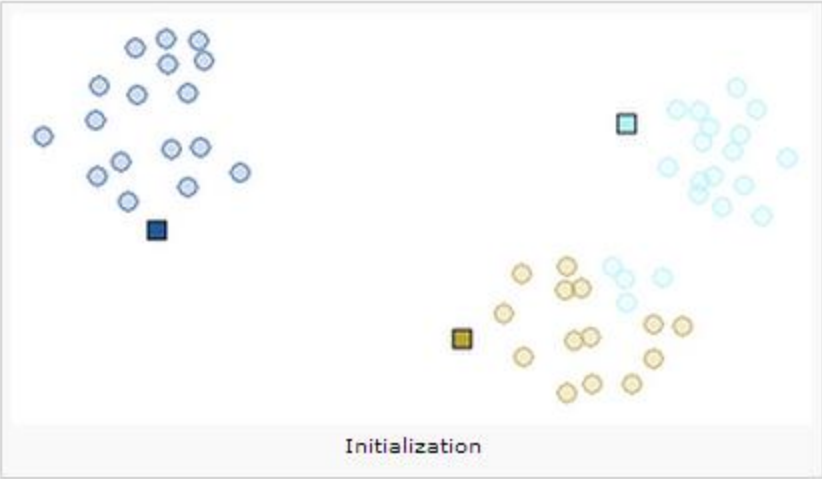
Optimality of *K*-means

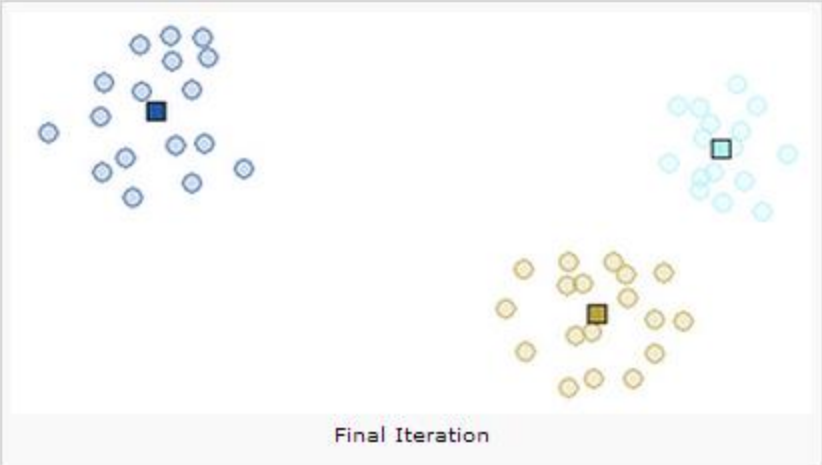
- Convergence does not mean that we converge to the optimal clustering!
- This is the weakness of K-means.
- If we start with a bad set of seeds, the resulting clustering can be horrible.

(a) Bad initialization



(b) Better initialization





Rule of thumb for *K*

One simple Rule of thumb sets *K* to

$$k \approx \sqrt{n/2}$$

with *n* as the number of data points

When does K-means not work?