

Introduction to Information Retrieval

<http://informationretrieval.org>

IIR 16: Flat Clustering

Hinrich Schütze

Institute for Natural Language Processing, University of Stuttgart

2011-12-16

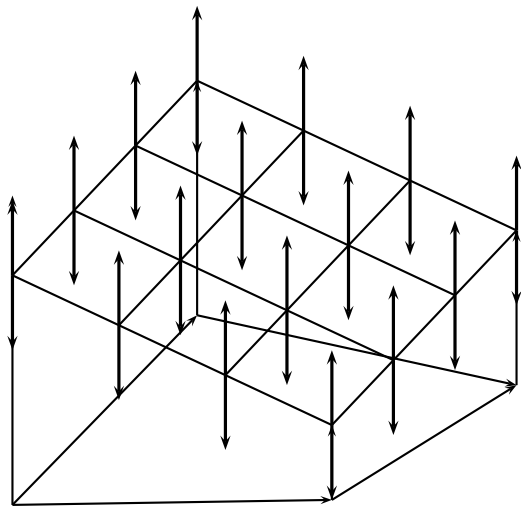
Overview

- 1 Recap
- 2 Clustering: Introduction
- 3 Clustering in IR
- 4 K -means
- 5 Evaluation
- 6 How many clusters?

Outline

- 1 Recap
- 2 Clustering: Introduction
- 3 Clustering in IR
- 4 K -means
- 5 Evaluation
- 6 How many clusters?

A linear classifier in 3D



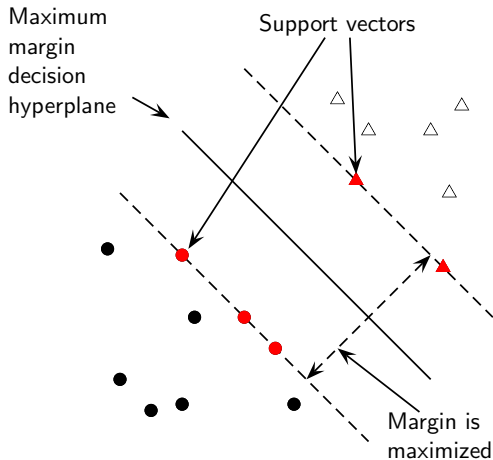
- A linear classifier in 3D is a plane described by the equation
$$w_1 d_1 + w_2 d_2 + w_3 d_3 = \theta$$
- Example for a 3D linear classifier
- Points $(d_1 \ d_2 \ d_3)$ with $w_1 d_1 + w_2 d_2 + w_3 d_3 \geq \theta$ are in the class c .
- Points $(d_1 \ d_2 \ d_3)$ with $w_1 d_1 + w_2 d_2 + w_3 d_3 < \theta$ are in the complement class \bar{c} .

Linear classifiers

- Many common text classifiers are linear classifiers: Naive Bayes, Rocchio, logistic regression, least squares regression, linear support vector machines etc.
- Each method has a different way of selecting the separating hyperplane
 - Huge differences in performance on test documents

Support vector machines

- Binary classification problem
- Simple SVMs are linear classifiers.
- criterion: being maximally far away from any data point → determines classifier **margin**
- linear separator position defined by **support vectors**



Optimization problem solved by SVMs

Find \vec{w} and b such that:

- $\frac{1}{2} \vec{w}^T \vec{w}$ is minimized (because $|\vec{w}| = \sqrt{\vec{w}^T \vec{w}}$), and
- for all $\{(\vec{x}_i, y_i)\}$, $y_i(\vec{w}^T \vec{x}_i + b) \geq 1$

Which machine learning method to choose

- Is there a learning method that is optimal for all text classification problems?
- No, because there is a tradeoff between bias and variance.
- Factors to take into account:
 - How much training data is available?
 - How simple/complex is the problem? (linear vs. nonlinear decision boundary)
 - How noisy is the problem?
 - How stable is the problem over time?
 - For an unstable problem, it's better to use a simple and robust classifier.

Take-away today

- What is clustering?
- Applications of clustering in information retrieval
- K -means algorithm
- Evaluation of clustering
- How many clusters?



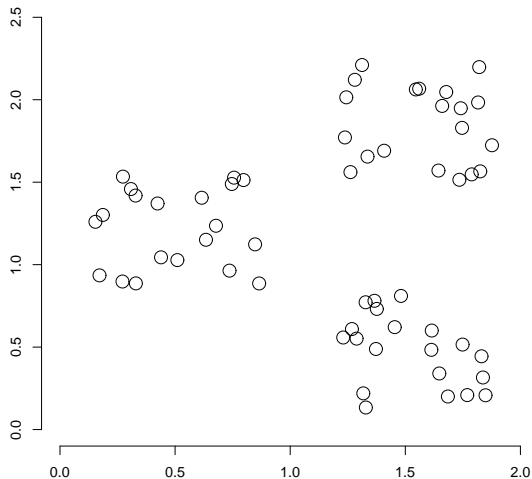
Outline

- 1 Recap
- 2 Clustering: Introduction
- 3 Clustering in IR
- 4 K -means
- 5 Evaluation
- 6 How many clusters?

Clustering: 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. □

Data set with clear cluster structure



Propose
algorithm
for finding
the cluster
structure in
this
example □

Classification vs. Clustering

- Classification: supervised learning
- Clustering: unsupervised learning
- Classification: Classes are **human-defined** and part of the input to the learning algorithm.
- 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, ...



Outline

- 1 Recap
- 2 Clustering: Introduction
- 3 Clustering in IR
- 4 K -means
- 5 Evaluation
- 6 How many clusters?

The cluster hypothesis

Cluster hypothesis. Documents in the same cluster behave similarly with respect to relevance to information needs. All

applications of clustering in IR are based (directly or indirectly) on the cluster hypothesis. Van Rijsbergen's original wording (1979):

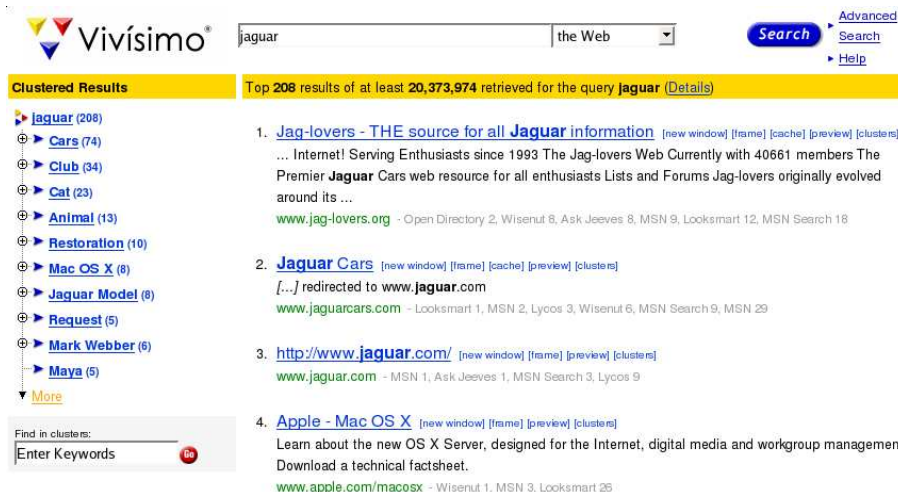
“closely associated documents tend to be relevant to the same requests” .



Applications of clustering in IR

| application | what is clustered? | benefit |
|--------------------------|-------------------------|---|
| search result clustering | search results | more effective information presentation to user |
| Scatter-Gather | (subsets of) collection | alternative user interface: "search without typing" |
| collection clustering | collection | effective information presentation for exploratory browsing |
| cluster-based retrieval | collection | higher efficiency: faster search |

Search result clustering for better navigation



The screenshot displays the Vivísimo search engine interface. At the top left is the Vivísimo logo, a stylized 'V' composed of four triangles in yellow, red, blue, and purple. To its right is a search bar containing the text 'jaguar' and a dropdown menu set to 'the Web'. Further right is a blue 'Search' button and links for 'Advanced Search' and 'Help'. Below the search bar is a yellow banner with the text 'Clustered Results' on the left and 'Top 208 results of at least 20,373,974 retrieved for the query jaguar (Details)' on the right. On the left side, a vertical list of clustered results is shown, each with a plus icon and a link to a category: 'jaguar (208)', 'Cars (74)', 'Club (34)', 'Cat (23)', 'Animal (13)', 'Restoration (10)', 'Mac OS X (8)', 'Jaguar Model (8)', 'Request (5)', 'Mark Webber (6)', and 'Maya (5)'. A 'More' link is at the bottom of this list. Below the clusters is a search box labeled 'Find in clusters:' with the placeholder text 'Enter Keywords' and a red 'Go' button. On the right side, a list of search results is displayed, each with a number, a title, and a description. The results are: 1. 'Jag-lovers - THE source for all Jaguar information' with a description of the website and a link to 'www.jag-lovers.org'; 2. 'Jaguar Cars' with a description of a redirected page and a link to 'www.jaguarcars.com'; 3. 'http://www.jaguar.com/' with a link to 'www.jaguar.com'; 4. 'Apple - Mac OS X' with a description of the operating system and a link to 'www.apple.com/macosx'.

Vivísimo [Advanced Search](#) [Help](#)

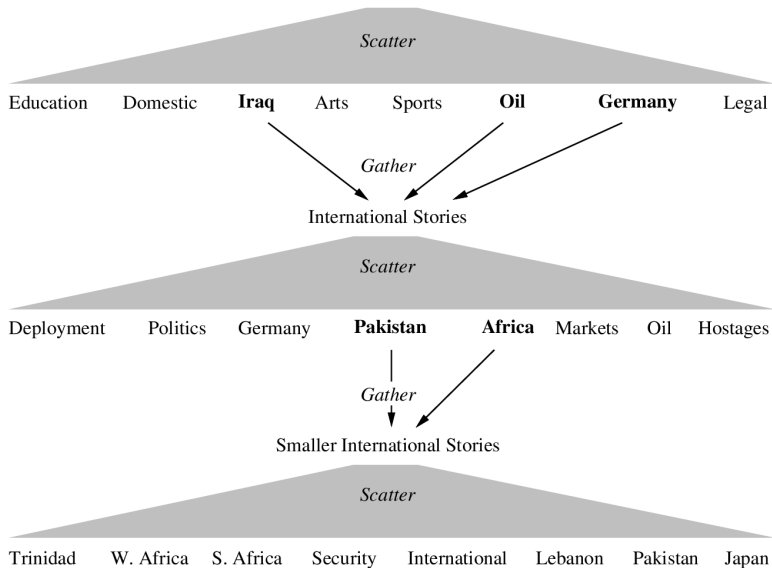
Clustered Results Top 208 results of at least 20,373,974 retrieved for the query **jaguar** ([Details](#))

- [jaguar](#) (208)
 - [Cars](#) (74)
 - [Club](#) (34)
 - [Cat](#) (23)
 - [Animal](#) (13)
 - [Restoration](#) (10)
 - [Mac OS X](#) (8)
 - [Jaguar Model](#) (8)
 - [Request](#) (5)
 - [Mark Webber](#) (6)
 - [Maya](#) (5)
 - [More](#)

Find in clusters:

- [Jag-lovers - THE source for all Jaguar information](#) [\[new window\]](#) [\[frame\]](#) [\[cache\]](#) [\[preview\]](#) [\[clusters\]](#)
... Internet! Serving Enthusiasts since 1993 The Jag-lovers Web Currently with 40661 members The Premier **Jaguar** Cars web resource for all enthusiasts Lists and Forums Jag-lovers originally evolved around its ...
[www.jag-lovers.org](#) - Open Directory 2, Wisenut 6, Ask Jeeves 8, MSN 9, Looksmart 12, MSN Search 18
- [Jaguar Cars](#) [\[new window\]](#) [\[frame\]](#) [\[cache\]](#) [\[preview\]](#) [\[clusters\]](#)
[...] redirected to [www.jaguar.com](#)
[www.jaguarcars.com](#) - Looksmart 1, MSN 2, Lycos 3, Wisenut 6, MSN Search 9, MSN 29
- [http://www.jaguar.com/](#) [\[new window\]](#) [\[frame\]](#) [\[preview\]](#) [\[clusters\]](#)
[www.jaguar.com](#) - MSN 1, Ask Jeeves 1, MSN Search 3, Lycos 9
- [Apple - Mac OS X](#) [\[new window\]](#) [\[frame\]](#) [\[preview\]](#) [\[clusters\]](#)
Learn about the new OS X Server, designed for the Internet, digital media and workgroup management
Download a technical factsheet.
[www.apple.com/macosx](#) - Wisenut 1, MSN 3, Looksmart 26


Scatter-Gather



Global navigation: Yahoo

YAHOO! DIRECTORY Search: the Web | the Directory | this category

Society and Culture
Directory > Society and Culture

 **Culture**
www.Dealtime.com Shop and save on Magazines. SPONSOR RE

CATEGORIES [\(What's This?\)](#)

Most Popular Society and Culture

- [Crime](#) (5453) NEWS!
- [Cultures and Groups](#) (11025) NEWS!
- [Environment and Nature](#) (8558) NEWS!
- [Families](#) (1215)
- [Food and Drink](#) (9776) NEWS!
- [Holidays and Observances](#) (3333)
- [Issues and Causes](#) (4842)
- [Mythology and Folklore](#) (984)
- [People](#) (16351)
- [Relationships](#) (595)
- [Religion and Spirituality](#) (37533)
- [Sexuality](#) (2812) NEWS!

Additional Society and Culture Categories

- [Advice](#) (48)
- [Chats and Forums](#) (27)
- [Cultural Policy](#) (10)
- [Death and Dying](#) (394)
- [Disabilities](#) (1293)
- [Employment and Work@](#)
- [Etiquette](#) (54)
- [Events](#) (27)
- [Fashion@](#)
- [Gender](#) (21)
- [Home and Garden](#) (1080) NEWS!
- [Magazines](#) (164)
- [Museums and Exhibits](#) (6052)
- [Pets@](#)
- [Reunions](#) (228)
- [Social Organizations](#) (338)
- [Web Directories](#) (6)
- [Weddings](#) (371)

SITE LISTINGS By Popularity | [Alphabetical](#) | [\(What's This?\)](#) Site

Global navigation: MESH (upper level)

MeSH Tree Structures - 2008

[Return to Entry Page](#)

1. Anatomy [A]
2. Organisms [B]
3. Diseases [C]
 - [Bacterial Infections and Mycoses \[C011\] +](#)
 - [Virus Diseases \[C02\] +](#)
 - [Parasitic Diseases \[C03\] +](#)
 - [Neoplasms \[C04\] +](#)
 - [Musculoskeletal Diseases \[C05\] +](#)
 - [Digestive System Diseases \[C06\] +](#)
 - [Stomatognathic Diseases \[C07\] +](#)
 - [Respiratory Tract Diseases \[C08\] +](#)
 - [Otorhinolaryngologic Diseases \[C09\] +](#)
 - [Nervous System Diseases \[C10\] +](#)
 - [Eye Diseases \[C11\] +](#)
 - [Male Urogenital Diseases \[C12\] +](#)
 - [Female Urogenital Diseases and Pregnancy Complications \[C13\] +](#)
 - [Cardiovascular Diseases \[C14\] +](#)
 - [Hemic and Lymphatic Diseases \[C15\] +](#)
 - [Congenital, Hereditary, and Neonatal Diseases and Abnormalities \[C16\] +](#)
 - [Skin and Connective Tissue Diseases \[C17\] +](#)
 - [Nutritional and Metabolic Diseases \[C18\] +](#)
 - [Endocrine System Diseases \[C19\] +](#)
 - [Immune System Diseases \[C20\] +](#)
 - [Disorders of Environmental Origin \[C21\] +](#)
 - [Animal Diseases \[C22\] +](#)
 - [Pathological Conditions, Signs and Symptoms \[C23\] +](#)
4. Chemicals and Drugs [D]
5. Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. Psychiatry and Psychology [F]
7. Biological Sciences [G]
8. Natural Sciences [H]
9. Anthropology, Education, Sociology and Social Phenomena [I]
10. Technology, Industry, Agriculture [J]
11. Humanities [K]

Global navigation: MESH (lower level)

[Neoplasms \[C04\]](#)

[Cysts \[C04.182\] +](#)

[Hamartoma \[C04.445\] +](#)

► [Neoplasms by Histologic Type \[C04.557\]](#)

[Histiocytic Disorders, Malignant \[C04.557.227\] +](#)

[Leukemia \[C04.557.337\] +](#)

[Lymphatic Vessel Tumors \[C04.557.375\] +](#)

[Lymphoma \[C04.557.386\] +](#)

[Neoplasms, Complex and Mixed \[C04.557.435\] +](#)

[Neoplasms, Connective and Soft Tissue \[C04.557.450\] +](#)

[Neoplasms, Germ Cell and Embryonal \[C04.557.465\] +](#)

[Neoplasms, Glandular and Epithelial \[C04.557.470\] +](#)

[Neoplasms, Gonadal Tissue \[C04.557.475\] +](#)

[Neoplasms, Nerve Tissue \[C04.557.580\] +](#)

[Neoplasms, Plasma Cell \[C04.557.595\] +](#)

[Neoplasms, Vascular Tissue \[C04.557.645\] +](#)

[Nevi and Melanomas \[C04.557.665\] +](#)

[Odontogenic Tumors \[C04.557.695\] +](#)

[Neoplasms by Site \[C04.588\] +](#)

[Neoplasms, Experimental \[C04.619\] +](#)

[Neoplasms, Hormone-Dependent \[C04.626\]](#)

[Neoplasms, Multiple Primary \[C04.651\] +](#)

[Neoplasms, Post-Traumatic \[C04.666\]](#)

[Neoplasms, Radiation-Induced \[C04.682\] +](#)

[Neoplasms, Second Primary \[C04.692\]](#)

[Neoplastic Processes \[C04.697\] +](#)

[Neoplastic Syndromes, Hereditary \[C04.700\] +](#)

[Paraneoplastic Syndromes \[C04.730\] +](#)

[Precancerous Conditions \[C04.834\] +](#)

[Pregnancy Complications, Neoplastic \[C04.850\] +](#)

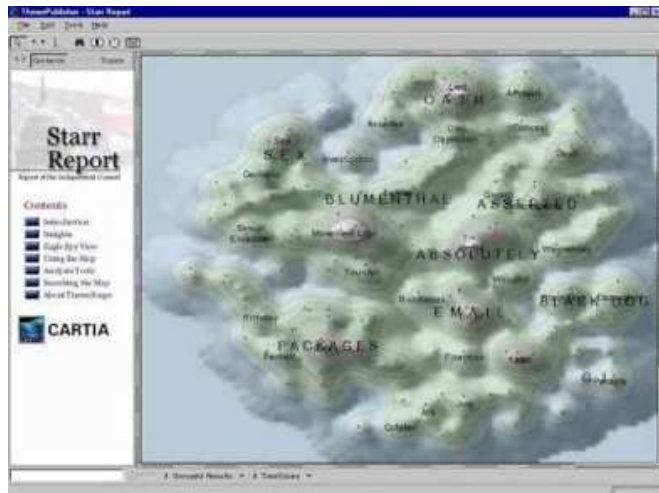
[Tumor Virus Infections \[C04.925\] +](#)

Navigational hierarchies: Manual vs. automatic creation

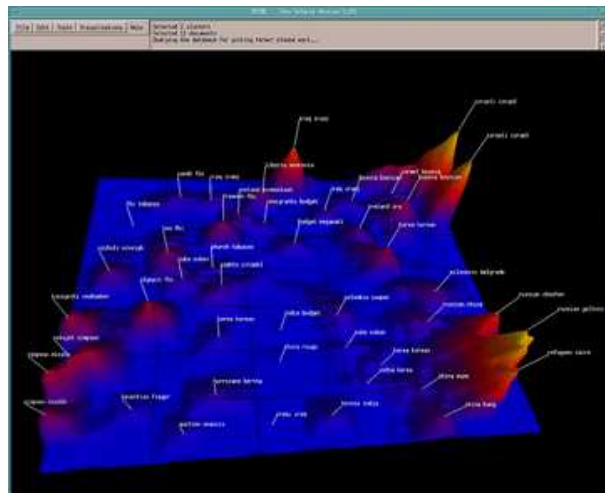
- Note: Yahoo/MESH are **not** examples of clustering.
- But they are well known examples for using a global hierarchy for navigation.
- Some examples for global navigation/exploration based on clustering:
 - Cartia
 - Themescapes
 - Google News



Global navigation combined with visualization (1)



Global navigation combined with visualization (2)



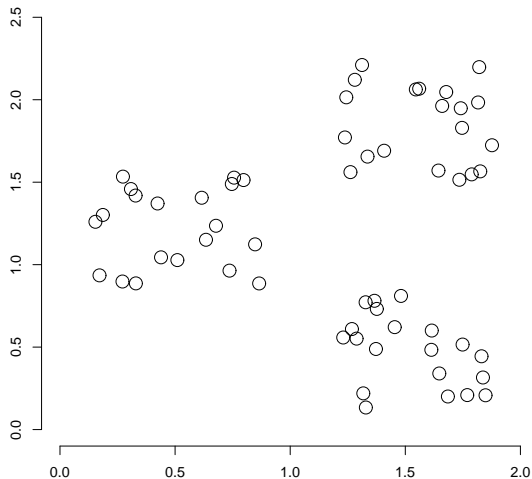
Global clustering for navigation: Google News

<http://news.google.com>

Clustering for improving recall

- To improve search recall:
 - Cluster docs in collection a priori
 - When a query matches a doc d , also return other docs in the cluster containing d
- Hope: if we do this: the query “car” will also return docs containing “automobile”
 - Because the clustering algorithm groups together docs containing “car” with those containing “automobile” .
 - Both types of documents contain words like “parts”, “dealer”, “mercedes”, “road trip” . □

Data set with clear cluster structure



Propose
algorithm
for finding
the cluster
structure in
this
example

Desiderata for clustering

- General goal: put related docs in the same cluster, put unrelated docs in different clusters.
 - We'll see different ways of formalizing this.
- The number of clusters should be appropriate for the data set we are clustering.
 - Initially, we will assume the number of clusters K is given.
 - Later: Semiautomatic methods for determining K
- Secondary goals in clustering
 - Avoid very small and very large clusters
 - Define clusters that are easy to explain to the user
 - Many others . . .



Flat vs. Hierarchical clustering

- Flat algorithms
 - Usually start with a random (partial) partitioning of docs into groups
 - Refine iteratively
 - Main algorithm: K -means
- Hierarchical algorithms
 - Create a hierarchy
 - Bottom-up, agglomerative
 - Top-down, divisive



Hard vs. Soft clustering

- Hard clustering: Each document belongs to **exactly one** cluster.
 - More common and easier to do
- Soft clustering: A document can belong to **more than one** cluster.
 - Makes more sense for applications like creating browsable hierarchies
 - You may want to put *sneakers* in two clusters:
 - sports apparel
 - shoes
 - You can only do that with a soft clustering approach.
- This class: flat, hard clustering
- Next time: hierarchical, hard clustering
- Next week: latent semantic indexing, a form of soft clustering



Flat algorithms

- Flat algorithms compute a partition of N documents into a set of K clusters.
- Given: a set of documents and the number K
- Find: a partition into K clusters that optimizes the chosen partitioning criterion
- Global optimization: exhaustively enumerate partitions, pick optimal one
 - Not tractable
- Effective heuristic method: K -means algorithm □

Outline

- 1 Recap
- 2 Clustering: Introduction
- 3 Clustering in IR
- 4 *K*-means
- 5 Evaluation
- 6 How many clusters?

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.
- Almost: centroids are not length-normalized.



K-means: Basic idea

- Each cluster in K-means is defined by a **centroid**.
- Objective/partitioning criterion: **minimize the average squared difference from the centroid**
- Recall definition of centroid:

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

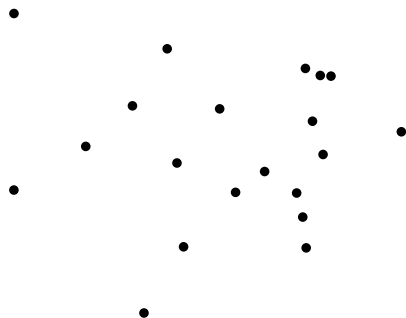
where we use ω to denote a cluster.

- We try to find the minimum average squared difference by iterating two steps:
 - **reassignment**: assign each vector to its closest centroid
 - **recomputation**: recompute each centroid as the average of the vectors that were assigned to it in reassignment □

K-means pseudocode (μ_k is centroid of ω_k)

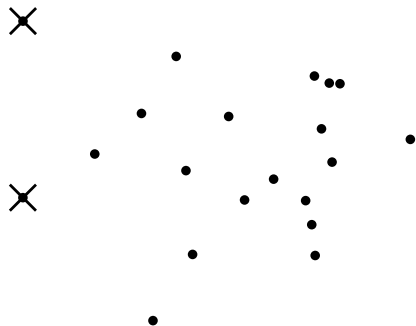
```
K-MEANS( $\{\vec{x}_1, \dots, \vec{x}_N\}, K$ )
1  ( $\vec{s}_1, \vec{s}_2, \dots, \vec{s}_K$ )  $\leftarrow$  SELECTRANDOMSEEDS( $\{\vec{x}_1, \dots, \vec{x}_N\}, K$ )
2  for  $k \leftarrow 1$  to  $K$ 
3  do  $\vec{\mu}_k \leftarrow \vec{s}_k$ 
4  while stopping criterion has not been met
5  do for  $k \leftarrow 1$  to  $K$ 
6    do  $\omega_k \leftarrow \{\}$ 
7    for  $n \leftarrow 1$  to  $N$ 
8    do  $j \leftarrow \arg \min_{j'} |\vec{\mu}_{j'} - \vec{x}_n|$ 
9       $\omega_j \leftarrow \omega_j \cup \{\vec{x}_n\}$  (reassignment of vectors)
10   for  $k \leftarrow 1$  to  $K$ 
11   do  $\vec{\mu}_k \leftarrow \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} \vec{x}$  (recomputation of centroids)
12 return  $\{\vec{\mu}_1, \dots, \vec{\mu}_K\}$ 
```

Worked Example : Set of points to be clustered

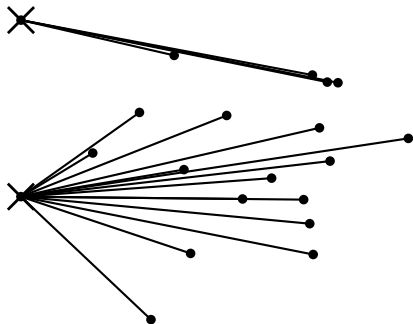


Exercise: (i) Guess what the optimal clustering into two clusters is in this case; (ii) compute the centroids of the clusters □

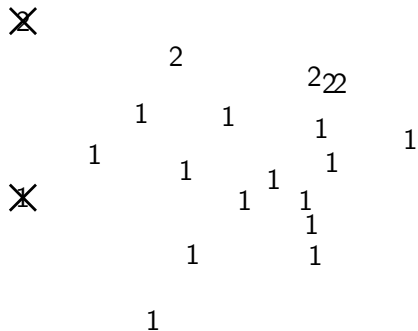
Worked Example: Random selection of initial centroids



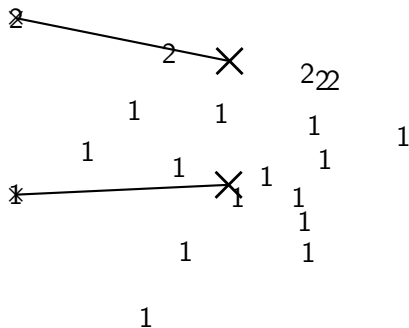
Worked Example: Assign points to closest center



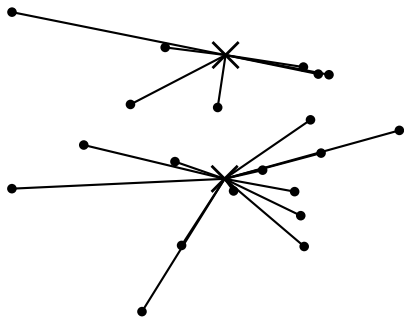
Worked Example: Assignment



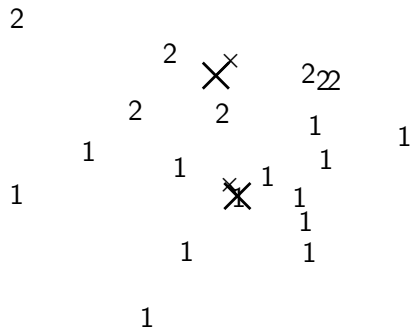
Worked Example: Recompute cluster centroids



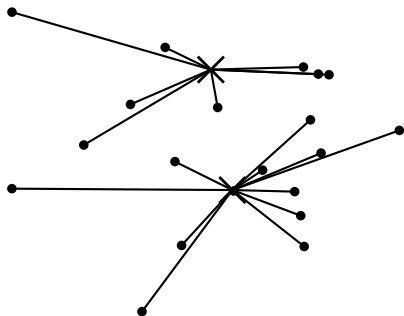
Worked Example: Assign points to closest centroid



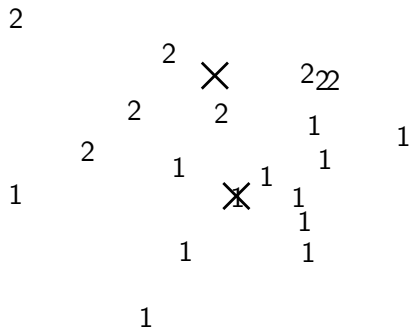
Worked Example: Recompute cluster centroids



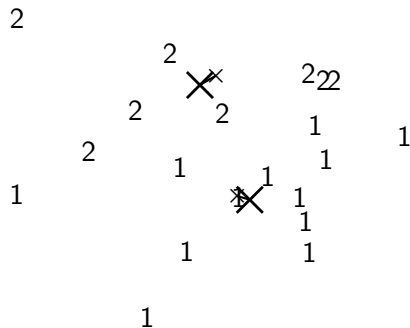
Worked Example: Assign points to closest centroid



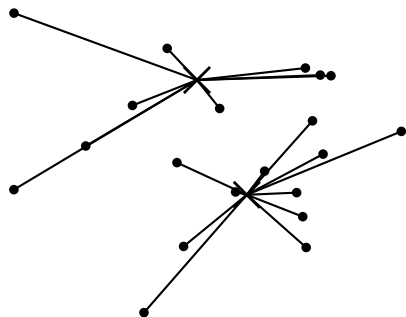
Worked Example: Assignment



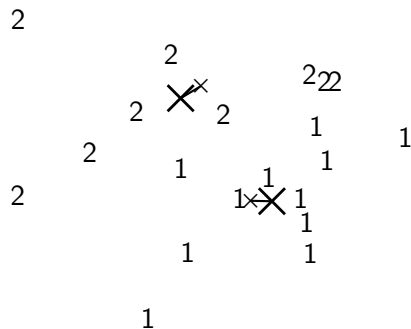
Worked Example: Recompute cluster centroids



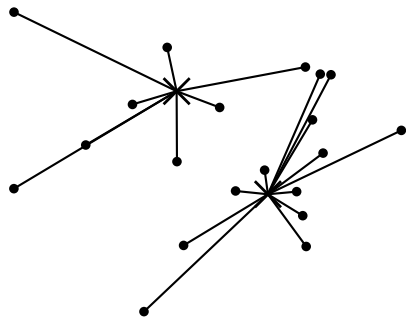
Worked Example: Assign points to closest centroid



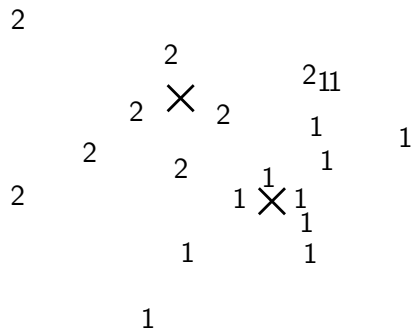
Worked Example: Recompute cluster centroids



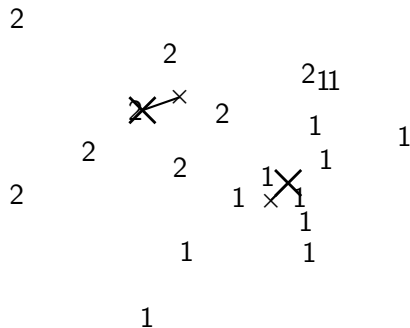
Worked Example: Assign points to closest centroid



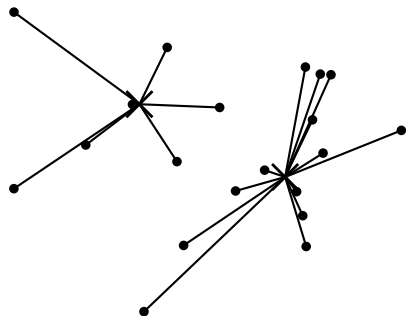
Worked Example: Assignment



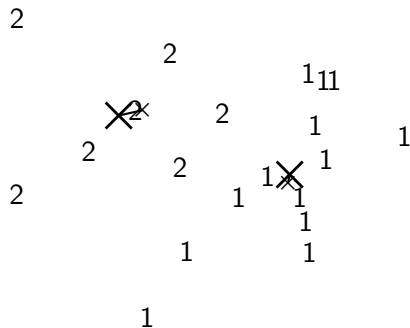
Worked Example: Recompute cluster centroids



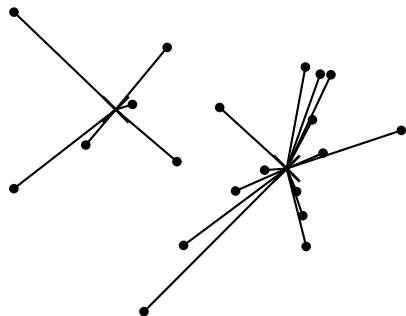
Worked Example: Assign points to closest centroid



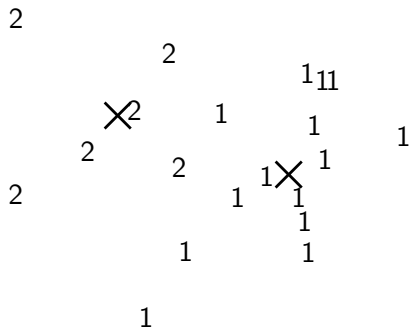
Worked Example: Recompute cluster centroids



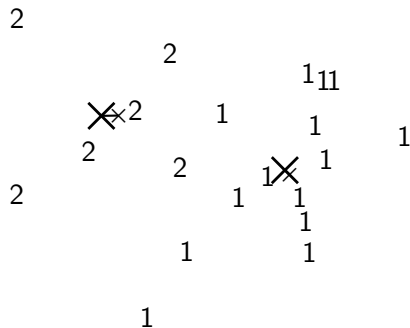
Worked Example: Assign points to closest centroid



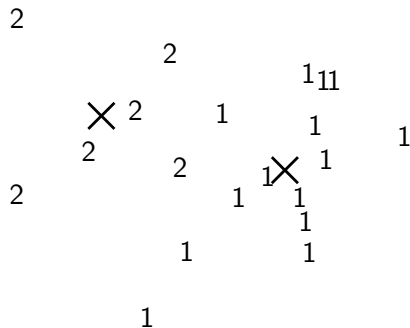
Worked Example: Assignment



Worked Example: Recompute cluster centroids



Worked Ex.: Centroids and assignments after convergence



K -means is guaranteed to converge: Proof

- RSS = sum of all squared distances between document vector and closest centroid
- RSS decreases during each reassignment step.
 - because each vector is moved to a closer centroid
- RSS decreases during each recomputation step.
 - see next slide
- There is only a finite number of clusterings.
- Thus: We must reach a fixed point.
- Assumption: Ties are broken consistently.
- Finite set & monotonically decreasing \rightarrow convergence □

Recomputation decreases average distance

$RSS = \sum_{k=1}^K RSS_k$ – the residual sum of squares (the “goodness” measure)

$$RSS_k(\vec{v}) = \sum_{\vec{x} \in \omega_k} \|\vec{v} - \vec{x}\|^2 = \sum_{\vec{x} \in \omega_k} \sum_{m=1}^M (v_m - x_m)^2$$
$$\frac{\partial RSS_k(\vec{v})}{\partial v_m} = \sum_{\vec{x} \in \omega_k} 2(v_m - x_m) = 0$$

$$v_m = \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} x_m$$

The last line is the componentwise definition of the centroid! We minimize RSS_k when the old centroid is replaced with the new centroid. RSS , the sum of the RSS_k , must then also decrease during recomputation. □

K -means is guaranteed to converge

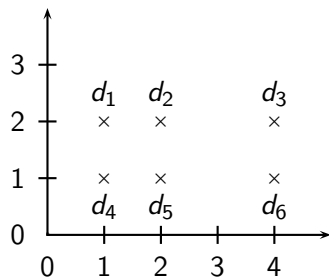
- But we don't know how long convergence will take!
- If we don't care about a few docs switching back and forth, then convergence is usually fast (< 10 - 20 iterations).
- However, complete convergence can take many more iterations.



Optimality of K -means

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

Exercise: Suboptimal clustering



- What is the optimal clustering for $K = 2$?
- Do we converge on this clustering for arbitrary seeds d_i, d_j ?



Initialization of K -means

- Random seed selection is just one of many ways K -means can be initialized.
- Random seed selection is not very robust: It's easy to get a suboptimal clustering.
- Better ways of computing initial centroids:
 - Select seeds not randomly, but using some heuristic (e.g., filter out outliers or find a set of seeds that has “good coverage” of the document space)
 - Use hierarchical clustering to find good seeds
 - Select i (e.g., $i = 10$) different random sets of seeds, do a K -means clustering for each, select the clustering with lowest RSS



Time complexity of K -means

- Computing one distance of two vectors is $O(M)$.
- Reassignment step: $O(KNM)$ (we need to compute KN document-centroid distances)
- Recomputation step: $O(NM)$ (we need to add each of the document's $< M$ values to one of the centroids)
- Assume number of iterations bounded by I
- Overall complexity: $O(IKNM)$ – linear in all important dimensions
- However: This is not a real worst-case analysis.
- In pathological cases, complexity can be worse than linear. □

Outline

- 1 Recap
- 2 Clustering: Introduction
- 3 Clustering in IR
- 4 K -means
- 5 Evaluation
- 6 How many clusters?

What is a good clustering?

- Internal criteria
 - Example of an internal criterion: RSS in K -means
- But an internal criterion often does not evaluate the actual utility of a clustering in the application.
- Alternative: External criteria
 - Evaluate with respect to a human-defined classification □

External criteria for clustering quality

- Based on a gold standard data set, e.g., the Reuters collection we also used for the evaluation of classification
- Goal: Clustering should reproduce the classes in the gold standard
- (But we only want to reproduce how documents are divided into groups, not the class labels.)
- First measure for how well we were able to reproduce the classes: **purity** □

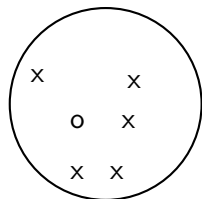
External criterion: Purity

$$\text{purity}(\Omega, C) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

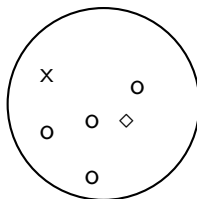
- $\Omega = \{\omega_1, \omega_2, \dots, \omega_K\}$ is the set of clusters and $C = \{c_1, c_2, \dots, c_J\}$ is the set of classes.
- For each cluster ω_k : find class c_j with most members n_{kj} in ω_k
- Sum all n_{kj} and divide by total number of points □

Example for computing purity

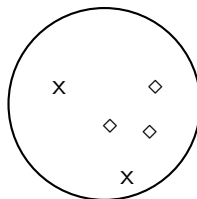
cluster 1



cluster 2



cluster 3



To compute

purity: $5 = \max_j |\omega_1 \cap c_j|$ (class x, cluster 1); $4 = \max_j |\omega_2 \cap c_j|$ (class o, cluster 2); and $3 = \max_j |\omega_3 \cap c_j|$ (class \diamond , cluster 3).
Purity is $(1/17) \times (5 + 4 + 3) \approx 0.71$. □

Another external criterion: Rand index

- Purity can be increased easily by increasing K – a measure that does not have this problem: Rand index.

- Definition: $RI = \frac{TP+TN}{TP+FP+FN+TN}$

- Based on 2x2 contingency table of all pairs of documents:

| | same cluster | different clusters |
|-------------------|----------------------|----------------------|
| same class | true positives (TP) | false negatives (FN) |
| different classes | false positives (FP) | true negatives (TN) |

- $TP+FN+FP+TN$ is the total number of pairs.
- $TP+FN+FP+TN = \binom{N}{2}$ for N documents.
- Example: $\binom{17}{2} = 136$ in o/◇/x example
- Each pair is either positive or negative (the clustering puts the two documents in the same or in different clusters) ...
- ...and either “true” (correct) or “false” (incorrect): the clustering decision is correct or incorrect. □

Rand Index: Example

As an example, we compute RI for the o/◇/x example. We first compute TP + FP. The three clusters contain 6, 6, and 5 points, respectively, so the total number of “positives” or pairs of documents that are in the same cluster is:

$$TP + FP = \binom{6}{2} + \binom{6}{2} + \binom{5}{2} = 40$$

Of these, the x pairs in cluster 1, the o pairs in cluster 2, the ◇ pairs in cluster 3, and the x pair in cluster 3 are true positives:

$$TP = \binom{5}{2} + \binom{4}{2} + \binom{3}{2} + \binom{2}{2} = 20$$

Thus, $FP = 40 - 20 = 20$. FN and TN are computed similarly. □

Rand measure for the o/◇/x example

| | same cluster | different clusters | |
|-------------------|--------------|--------------------|------------|
| same class | TP = 20 | FN = 24 | RI is then |
| different classes | FP = 20 | TN = 72 | |

$$(20 + 72)/(20 + 20 + 24 + 72) \approx 0.68.$$



Two other external evaluation measures

- Two other measures
- Normalized mutual information (NMI)
 - How much information does the clustering contain about the classification?
 - Singleton clusters (number of clusters = number of docs) have maximum MI
 - Therefore: normalize by entropy of clusters and classes
- F measure
 - Like Rand, but “precision” and “recall” can be weighted □

Evaluation results for the o/◇/x example

| | purity | NMI | RI | F_5 | |
|-------------------|--------|------|------|-------|----------|
| lower bound | 0.0 | 0.0 | 0.0 | 0.0 | All four |
| maximum | 1.0 | 1.0 | 1.0 | 1.0 | |
| value for example | 0.71 | 0.36 | 0.68 | 0.46 | |

measures range from 0 (really bad clustering) to 1 (perfect clustering).



Outline

- 1 Recap
- 2 Clustering: Introduction
- 3 Clustering in IR
- 4 K -means
- 5 Evaluation
- 6 How many clusters?

How many clusters?

- Number of clusters K is given in many applications.
 - E.g., there may be an external constraint on K . Example: In the case of Scatter-Gather, it was hard to show more than 10–20 clusters on a monitor in the 90s.
- What if there is no external constraint? Is there a “right” number of clusters?
- One way to go: define an optimization criterion
 - Given docs, find K for which the optimum is reached.
 - What optimization criterion can we use?
 - We can't use RSS or average squared distance from centroid as criterion: always chooses $K = N$ clusters. □

Exercise

- Your job is to develop the clustering algorithms for a competitor to news.google.com
- You want to use K -means clustering.
- How would you determine K ?



Simple objective function for K : Basic idea

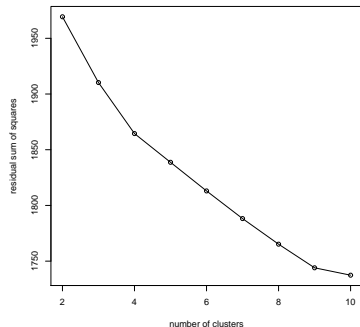
- Start with 1 cluster ($K = 1$)
- Keep adding clusters (= keep increasing K)
- Add a penalty for each new cluster
- Then trade off cluster penalties against average squared distance from centroid
- Choose the value of K with the best tradeoff



Simple objective function for K : Formalization

- Given a clustering, define the cost for a document as (squared) distance to centroid
- Define total **distortion** $RSS(K)$ as sum of all individual document costs (corresponds to average distance)
- Then: penalize each cluster with a cost λ
- Thus for a clustering with K clusters, total cluster penalty is $K\lambda$
- Define the total cost of a clustering as distortion plus total cluster penalty: $RSS(K) + K\lambda$
- Select K that minimizes $(RSS(K) + K\lambda)$
- Still need to determine good value for $\lambda \dots$ □

Finding the “knee” in the curve



Pick the number of clusters where
curve “flattens”. Here: 4 or 9. □

Take-away today

- What is clustering?
- Applications of clustering in information retrieval
- K -means algorithm
- Evaluation of clustering
- How many clusters?



Resources

- Chapter 16 of IIR
- Resources at <http://ifnlp.org/ir>
 - Keith van Rijsbergen on the cluster hypothesis (he was one of the originators)
 - Bing/Carrot2/Clusty: search result clustering systems
 - Stirling number: the number of distinct k -clusterings of n items