

信息组织与检索

第16讲：扁平聚类

主讲人：张蓉

华东师范大学数据科学与工程学院

提纲

- ① 上一讲回顾
- ② 聚类介绍
- ③ 聚类在IR中的应用
- ④ K-均值聚类算法
- ⑤ 聚类评价
- ⑥ 簇个数确定

提纲

- ① 上一讲回顾
- ③ 聚类介绍
- ③ 聚类在IR中的应用
- ④ K-均值聚类算法
- ⑤ 聚类评价
- ⑥ 簇个数确定

特征选择

- 文本分类中，通常要将文本表示在一个高维空间下，每一维对应一个词项
- 本讲义中，我们不特意区分不同的概念：每个坐标轴 = 维 = 词语 = 词项 = 特征
- 许多维上对应是罕见词
- 罕见词可能会误导分类器
- 这些会误导分类器的罕见词被称为噪音特征（noise feature）
- 去掉这些噪音特征会同时提高文本分类的效率和效果
- 上述过程称为特征选择（feature selection）

Reuters 语料中 *poultry*/EXPORT 的 MI 计算

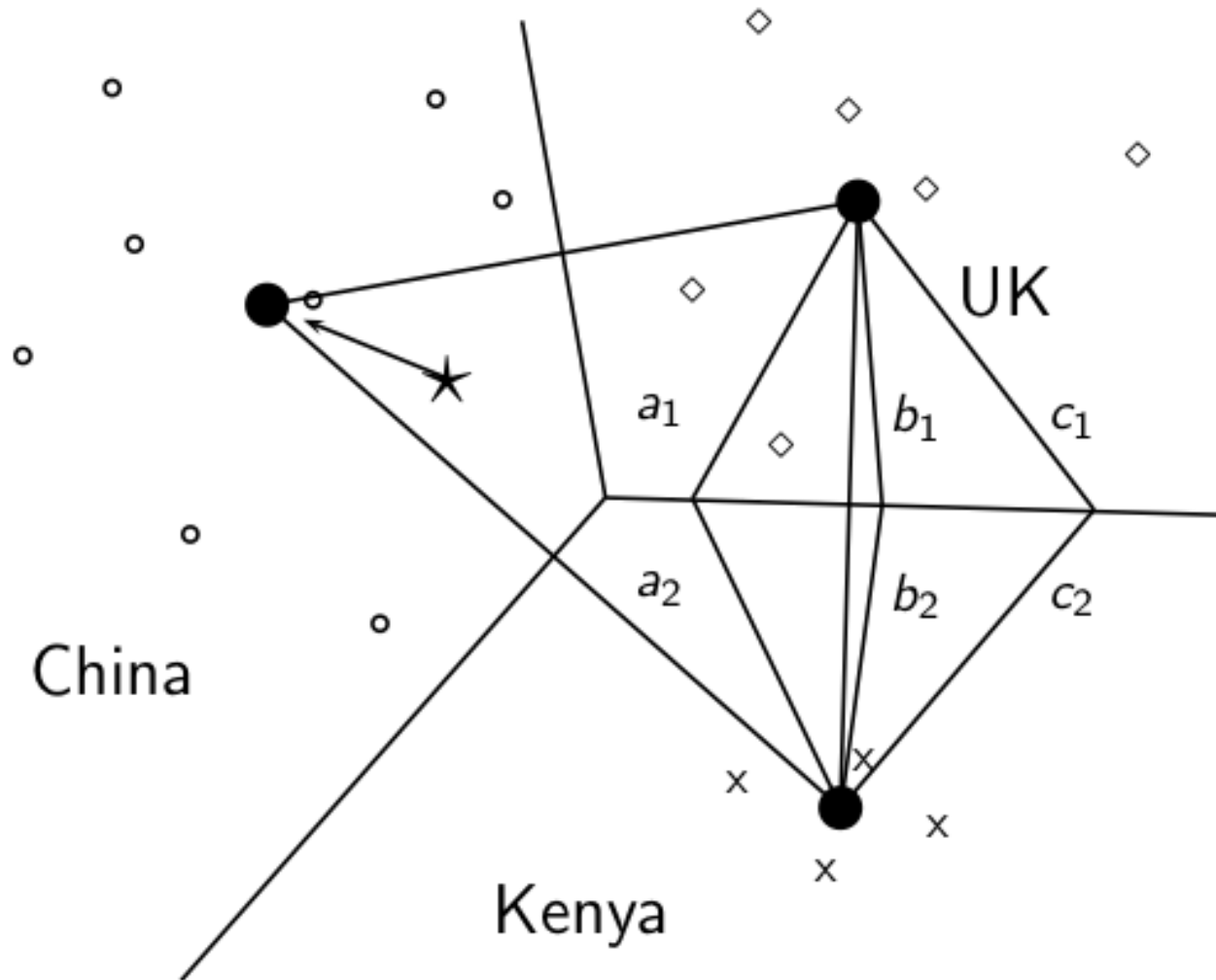
	$e_c = e_{\text{poultry}} = 1$	$e_c = e_{\text{poultry}} = 0$
$e_t = e_{\text{export}} = 1$	$N_{11} = 49$	$N_{10} = 27\ 652$
$e_t = e_{\text{export}} = 0$	$N_{01} = 141$	$N_{00} = 774\ 106$

$$\begin{aligned}
 I(U;C) = & \frac{49}{801\ 948} \log_2 \frac{801\ 948 \times 49}{(49 + 27\ 652)(49 + 141)} \\
 & + \frac{141}{801\ 948} \log_2 \frac{801\ 948 \times 141}{(141 + 774\ 106)(49 + 141)} \\
 & + \frac{27\ 652}{801\ 948} \log_2 \frac{801\ 948 \times 27\ 652}{(49 + 27\ 652)(27\ 652 + 774\ 106)} \\
 & + \frac{774\ 106}{801\ 948} \log_2 \frac{801\ 948 \times 774\ 106}{(141 + 774\ 106)(27\ 652 + 774\ 106)} \\
 \approx & 0.000\ 110\ 5
 \end{aligned}$$

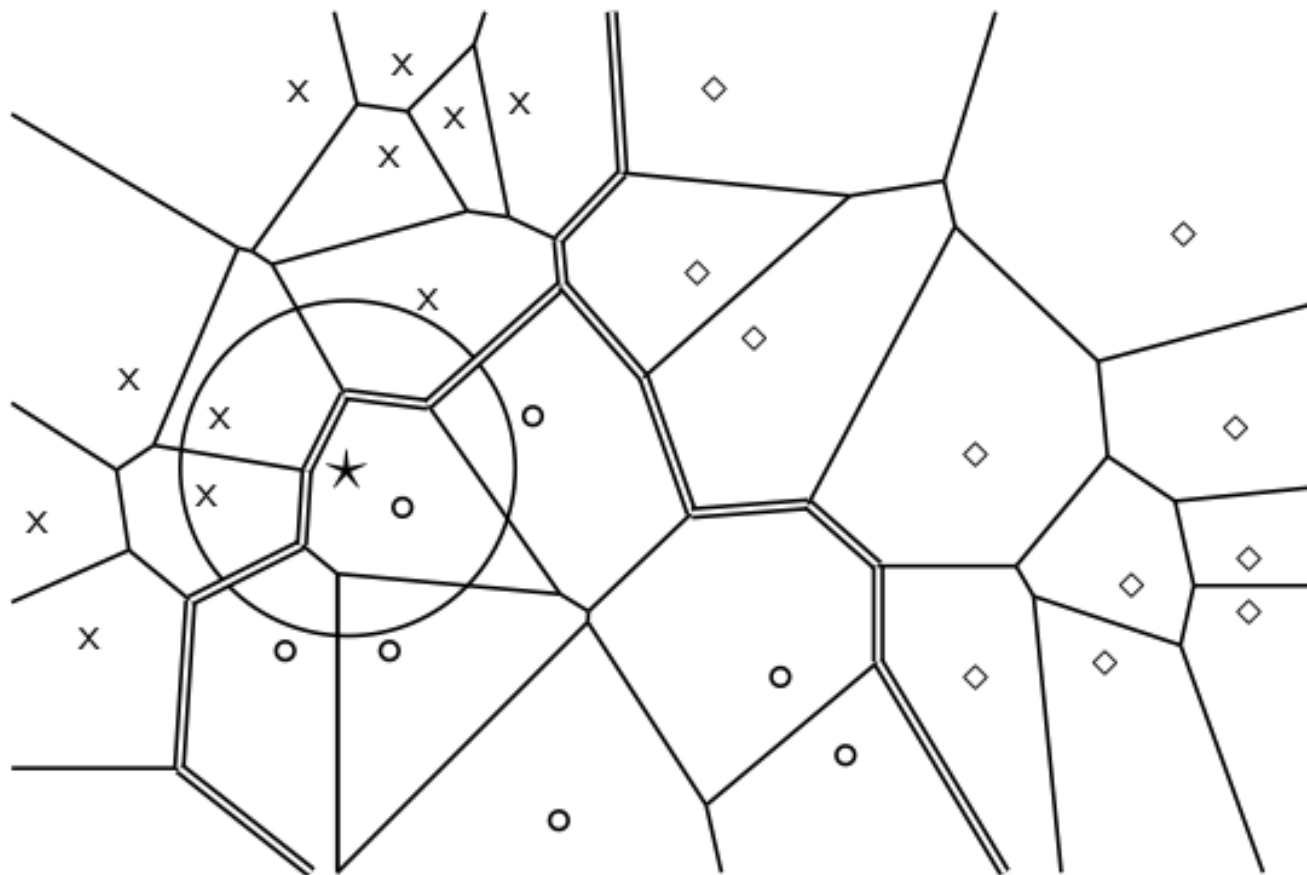
向量空间分类

- 同前面一样，训练集包含一系列文档，每篇都标记着它的类别
- 在向量空间分类中，该集合对应着空间中一系列标记的点或向量。
- 假设 1: 同一类中的文档会构成一片连续区域 (**contiguous region**)
- 假设2: 来自不同类别的文档没有交集
- 接下来我们定义直线、平面、超平面来将上述不同区域分开

Rocchio算法示意图 : $a_1 = a_2, b_1 = b_2, c_1 = c_2$



kNN 算法



对于★ 对应的文档，
在1NN和 3NN下，
分别应该属于哪个类？

线性分类器

- 定义：

- 线性分类器计算特征值的一个线性加权和 $\sum_i w_i x_i$

- 决策规则： $\sum_i w_i x_i > \theta?$

- 其中， θ 是一个参数

- 首先，我们仅考虑二元分类器

- 从几何上说，二元分类器相当于二维平面上的一条直线、三维空间中的一个平面或者更高维下的超平面，称为分类面

- 基于训练集来寻找该分类面

- 寻找分类面的方法：感知机(Perceptron)、 Rocchio, Naïve Bayes – 我们将解释为什么后两种方法也是二元分类器

- 假设：分类是线性可分的

本讲内容

- 聚类的概念(What is clustering?)
- 聚类在IR中的应用
- K -均值(K -Means)聚类算法
- 聚类评价
- 簇(cluster)个数(即聚类的结果类别个数)确定

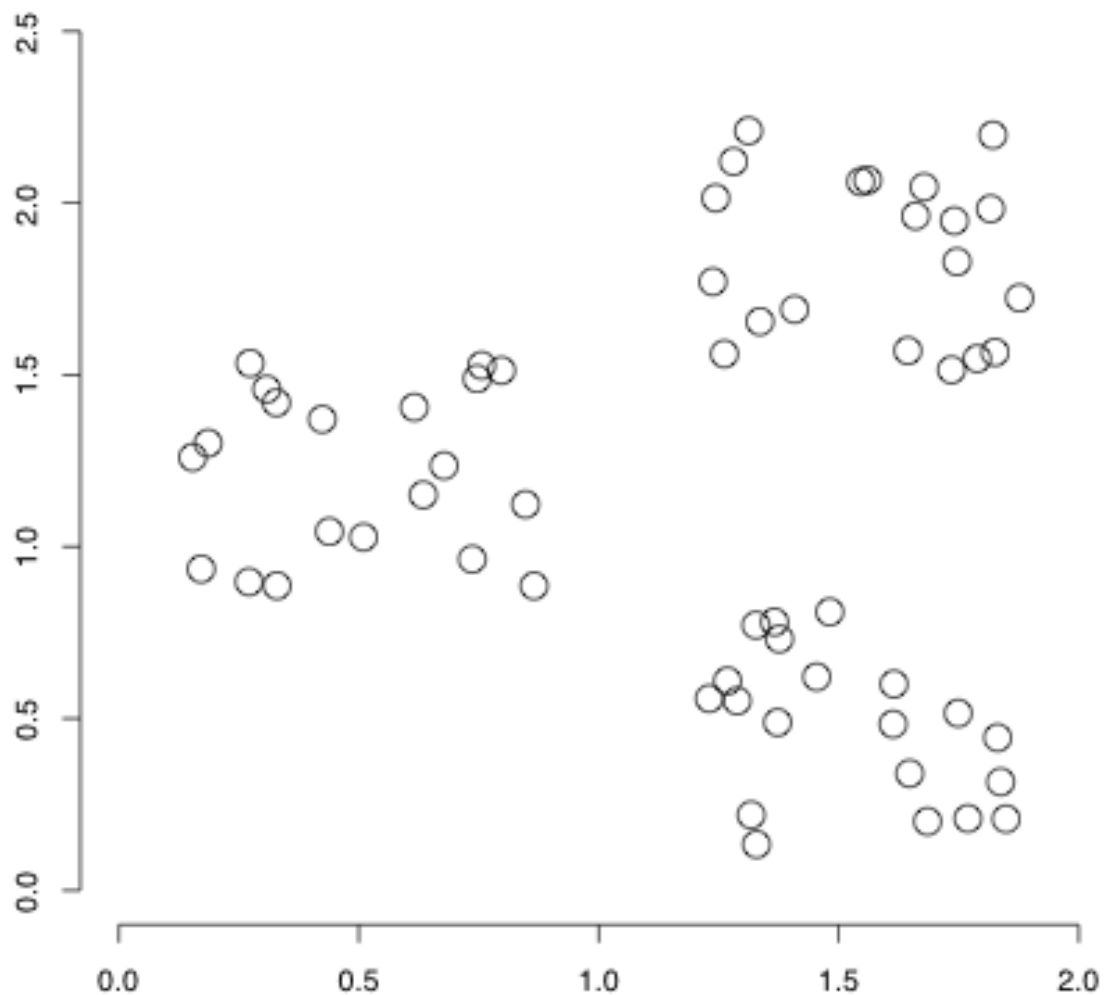
提纲

- ① 上一讲回顾
- ② 聚类介绍
- ③ 聚类在IR中的应用
- ④ K-均值聚类算法
- ⑤ 聚类评价
- ⑥ 簇个数确定

聚类(Clustering)的定义

- (文档)聚类是将一系列文档按照相似性聚团成子集或者簇(cluster)的过程
- 簇内文档之间应该彼此相似
- 簇间文档之间相似度不大
- 聚类是一种最常见的无监督学习(unsupervised learning)方法
- 无监督意味着没有已标注好的数据集

一个具有清晰簇结构的数据集



提出一个算法来寻找该
例中的簇结构

分类 vs. 聚类

- 分类: 有监督的学习
- 聚类: 无监督的学习
- 分类: 类别事先人工定义好, 并且是学习算法的输入的一部分
- 聚类: 簇在没有人工输入的情况下从数据中推理而得
 - 但是, 很多因素会影响聚类的输出结果: 簇的个数、相似度计算方法、文档的表示方式, 等等

提纲

- ① 上一讲回顾
- ② 聚类介绍
- ③ 聚类在IR中的应用
- ④ K-均值聚类算法
- ⑤ 聚类评价
- ⑥ 簇个数确定

聚类假设

聚类假设：在考虑文档和信息需求之间的相关性时，同一簇中的文档表现互相类似.

聚类在IR中的应用：

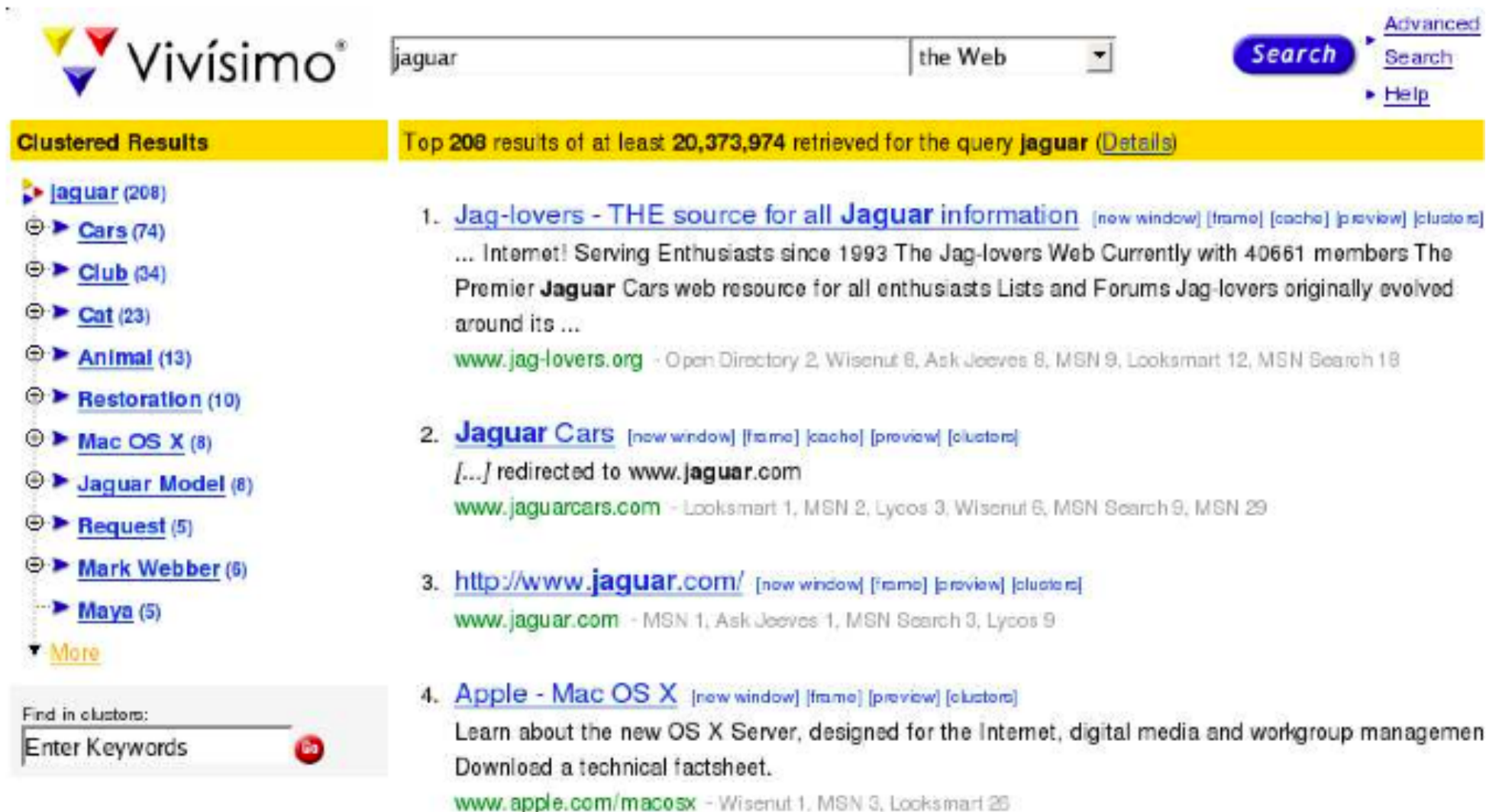
所有应用都直接或间接基于上述聚类假设

Van Rijsbergen的原始定义：“closely associated documents tend to be relevant to the same requests”（彼此密切关联的文档和同一信息需求相关）

聚类在IR中的应用

应 用	聚类对象	优 点
搜索结果聚类	搜索结果	提供面向用户的更有效的展示
“分散—集中”界面	文档集和文档子集	提供了另一种用户界面，即不需要人工输入关键词的搜索界面
文档集聚类	文档集	提供了一种面向探索式浏览的有效信息展示方法
基于语言建模的IR文档集	文档集	提高了正确率和/或召回率
基于聚类的检索	文档集	加快了搜索的速度

搜索结果的聚类：更好地浏览



The screenshot displays the Vivísimo search engine interface. At the top, the Vivísimo logo is on the left, followed by a search bar containing the text 'jaguar' and a dropdown menu set to 'the Web'. To the right of the search bar is a blue 'Search' button and links for 'Advanced Search' and 'Help'.

Below the search bar, a yellow banner indicates 'Top 208 results of at least 20,373,974 retrieved for the query **jaguar** (Details)'. The results are organized into two main sections:


- Clustered Results:** A vertical list on the left side of the results area, featuring expandable categories with minus and plus icons. The categories and their counts are: [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.
- Top 208 results:** A list of search results on the right side. The first four results are:
 - [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 8, 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 25

At the bottom left, there is a 'Find in clusters:' section with a text input field labeled 'Enter Keywords' and a red 'Go' button.

全局浏览: Yahoo

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

Society and Culture
Directory > Society and Culture

 [Culture](http://www.Dealtime.com)
www.Dealtime.com Shop and save on Magazines. SPONSOR RE

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

Most Popular Society and Culture

- [Crime](#) (5453) **NEW**
- [Cultures and Groups](#) (11025) **NEW**
- [Environment and Nature](#) (8659) **NEW**
- [Families](#) (1215)
- [Food and Drink](#) (9775) **NEW**
- [Holidays and Observances](#) (3333)
- [Issues and Causes](#) (4842)
- [Mythology and Folklore](#) (584)
- [People](#) (16361)
- [Relationships](#) (585)
- [Religion and Spirituality](#) (37533)
- [Sexuality](#) (2812) **NEW**

Additional Society and Culture Categories

- [Advice](#) (48)
- [Chats and Forums](#) (27)
- [Cultural Policy](#) (10)
- [Death and Dying](#) (394)
- [Disabilities](#) (1253)
- [Employment and Work](#) (2)
- [Etiquette](#) (54)
- [Events](#) (27)
- [Fashion](#) (2)
- [Gender](#) (21)
- [Home and Garden](#) (1080) **NEW**
- [Magazines](#) (184)
- [Museums and Exhibits](#) (6052)
- [Pets](#) (2)
- [Reunions](#) (228)
- [Social Organizations](#) (336)
- [Web Directories](#) (6)
- [Weddings](#) (371)

SITE LISTINGS By Popularity | [Alphabetical](#) [\(What's This?\)](#) 32/36

全局浏览: MESH (上层目录)

MeSH Tree Structures - 2008

[Return to Entry Page](#)

1. [+](#) Anatomy [A]
2. [+](#) Organisms [B]
3. [+](#) Diseases [C]
 - [Bacterial Infections and Mycoses \[C01\] +](#)
 - [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]

全局浏览: MESH (低层目录)

[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 \[C04.925\] +](#)

浏览的层次结构: 人工构建 vs. 自动构建

- 注意: Yahoo/MESH 并不是聚类的例子, 只是说明聚类的用途
- 但是它们都是根据目录进行浏览的著名的例子
- 也有一些例子是根据聚类进行全局浏览或探索:
 - Cartia
 - Themescapes
 - Google News

全局浏览的例子: Google News



新闻

添加栏目

个性化谷歌新闻 ▼

⚙

中国版 (China) ▼

焦点报道

国际/港台

内地

财经

娱乐

科技

互联网

体育

社会

汽车

房产

教育

热门报道

更新时间: 5分钟前



事业单位公开招聘 既要招到合适的人 又要公平公正
凤凰网 - 58分钟前
中广网北京 11月25日消息 (记者侯艳) 据中国之声《央广新闻》报道, 国务院法制办昨天 (24日) 公布《事业单位人事管理条例 (征求意见稿) 》, 面向社会各界征求意见, 其中有关公开招聘的内容引人关注。 ...
我国拟规范事业单位工资福利 规范薪资与社保 搜狐
事业单位公开招聘 既要招到合适的人 又要公平公正 中国新闻网
新浪网 - 新华网 - 南方报业
此专题所有 1,023 篇报道 »



叙利亚反政府军呼吁外国军队发动空袭 加速政府垮台 搜狐 - 5分钟前
专题报道(600篇) »



速冻食品新国标下月施行 金黄葡萄球菌允许检出 新民网 - 17分钟前
专题报道(2,050篇) »



大连四把火处理结果中石油董事长被处分 14人法办 凤凰网 - 43分钟前 专题报道(755篇) »



销售员坚持投注获双色球二等奖 搜狐 - 20分钟前 专题报道(689篇) »



中国海军西太训练引关注 日媒称系有意对日施压 环球网 - 1小时前
专题报道(449篇) »



住建部研究中心主任陈淮解读房地产 凤凰网 - 19分钟前
专题报道(3,029篇) »



人民日报谈四川藏区多起年轻僧人自焚事件(1) 中华网 - 10分钟前



成都一中学选拔19名“尖子生”与校长共进晚餐 网易 - 13分钟前



午评:金融地产“变脸”拖累大盘反弹 沪指缩量跌0.36% 和讯网 - 39分钟前 - 专题报道(92篇) »



陈浩民就袭胸门痛哭道歉 阿娇陈冠希领衔玩道歉的明星(图) 新民网 - 20分钟前 - 专题报道(530篇) »



传Facebook手机最早明年4月上市 或免费 凤凰网 - 34分钟前 - 专题报道(84篇) »



阿里财报解读: 良无限价值远超一般交易平台 搜狐 - 36分钟前 - 专题报道(464篇) »



坏事也能变好事 韦迪: 未来中超要恢复亚冠满额 搜狐 - 29分钟前 - 专题报道(200篇) »



大四男生校内强奸同学未遂杀人 曾获学校奖学金 新浪网 - 28分钟前 - 专题报道(102篇) »



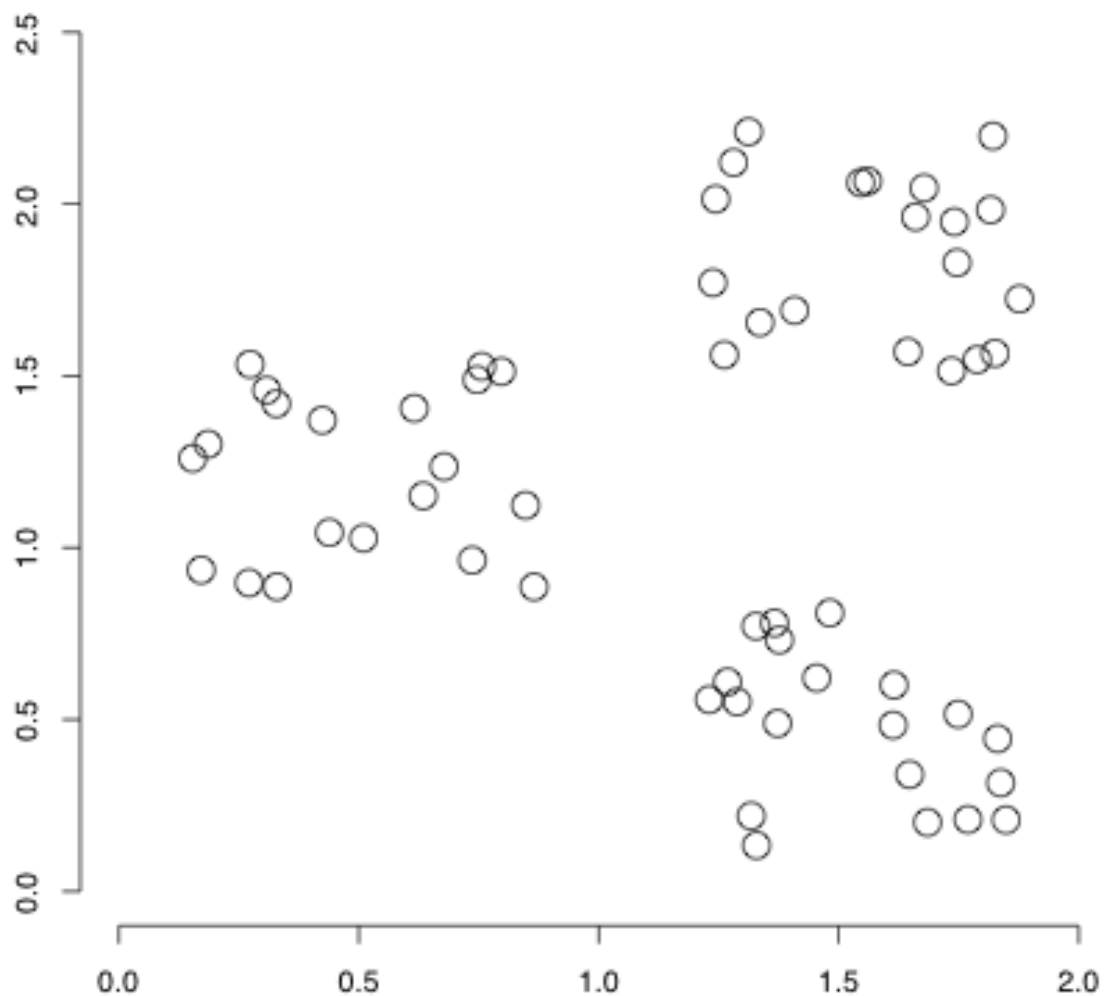
通用中国甘文维: 宝骏630月销量已达5000辆(图) 搜狐 - 46分钟前 - 专题报道(145篇) »

所有内容
新闻标题

文档聚类用于提高召回率

- 为提高搜索召回率：
 - 可以实现将文档集中的文档进行聚类
 - 当文档 d 和查询匹配时，也返回包含 d 的簇所包含的其它文档
 - 我们希望通过上述做法，在输入查询 “car” 时，也能够返回包含 “automobile” 的文档
 - 由于聚类算法会把包含 “car” 的文档和包含 “automobile” 的文档聚在一起
 - 两种文档都包含诸如 “parts”、“dealer”、“mercedes” 和 “road trip” 之类的词语

一个具有清晰簇结构的数据集



提出一个算法来寻找该
例中的簇结构

聚类的要求

- 一般目标：将相关文档放到一个簇中，将不相关文档放到不同簇中
 - 如何对上述目标进行形式化？
- 簇的数目应该合适，以便与聚类的数据集相吻合
 - 一开始，我们假设给定簇的数目为 K 。
 - 后面会介绍确定 K 的半自动的方法
- 聚类的其它目标
 - 避免非常小和非常大的簇
 - 定义的簇对用户来说很容易理解
 - 其它.....

扁平聚类 vs. 层次聚类

- 扁平算法

- 通过一开始将全部或部分文档随机划分为不同的组
- 通过迭代方式不断修正
- 代表算法：K-均值聚类算法

- 层次算法

- 构建具有层次结构的簇
- 自底向上(Bottom-up)的算法称为凝聚式(agglomerative)算法
- 自顶向下(Top-down)的算法称为分裂式(divisive)算法

硬聚类 vs. 软聚类

- 硬聚类(Hard clustering): 每篇文档仅仅属于一个簇
 - 很普遍并且相对容易实现
- 软聚类(Soft clustering): 一篇文档可以属于多个簇
 - 对于诸如浏览目录之类的应用来说很有意义
 - 比如, 将 胶底运动鞋 (sneakers) 放到两个簇中:
 - 体育服装(sports apparel)
 - 鞋类(shoes)
 - 只有通过软聚类才能做到这一点
- 本节课关注扁平的硬聚类算法
- 有关软聚类和层次聚类参考《信息检索导论》其他章节

扁平算法

- 扁平算法将 N 篇文档划分成 K 个簇
- 给定一个文档集合及聚类结果簇的个数 K
- 寻找一个划分将这个文档集合分成 K 个簇，该结果满足某个最优划分准则
- 全局优化：穷举所有的划分结果，从中选择最优的那个划分结果
 - 无法处理
- 高效的启发式方法： K -均值聚类算法

提纲

- ① 上一讲回顾
- ② 聚类介绍
- ③ 聚类在IR中的应用
- ④ K-均值聚类算法
- ⑤ 聚类评价
- ⑥ 簇个数确定

K-均值聚类算法

- 或许是最著名的聚类算法
- 算法十分简单，但是在很多情况下效果不错
- 是文档聚类的默认或基准算法

聚类中的文档表示

- 向量空间模型
- 同基于向量空间的分类一样，这里我们也采用欧氏距离的方法来计算向量之间的相关性.
- ...欧氏距离与余弦相似度差不多等价(如果两个向量都基于长度归一化，那么欧氏距离和余弦相似度是等价的)
- 然而，质心向量通常都没有基于长度进行归一化

K-均值聚类算法

- K-均值聚类算法中的每个簇都定义为其质心向量
- 划分准则：使得所有文档到其所在簇的质心向量的平方和最小
- 质心向量的定义：

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

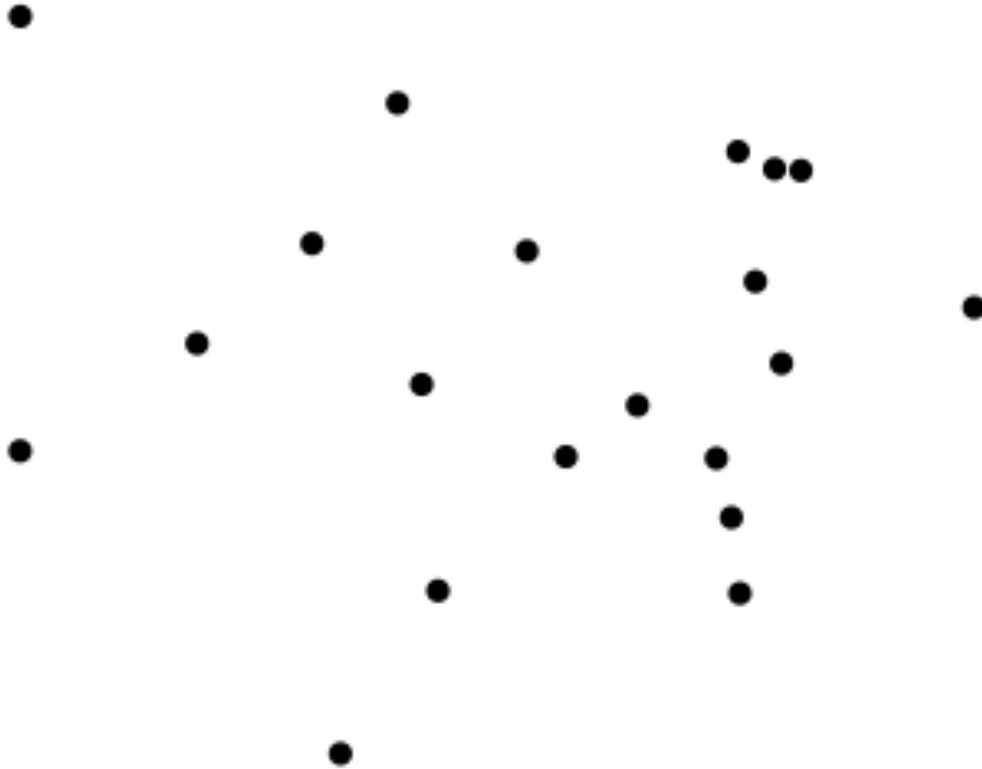
其中 ω 代表一个簇

- 通过下列两步来实现目标优化：
 - 重分配(reassignment): 将每篇文档分配给离它最近的簇
 - 重计算(recomputation): 重新计算每个簇的质心向量

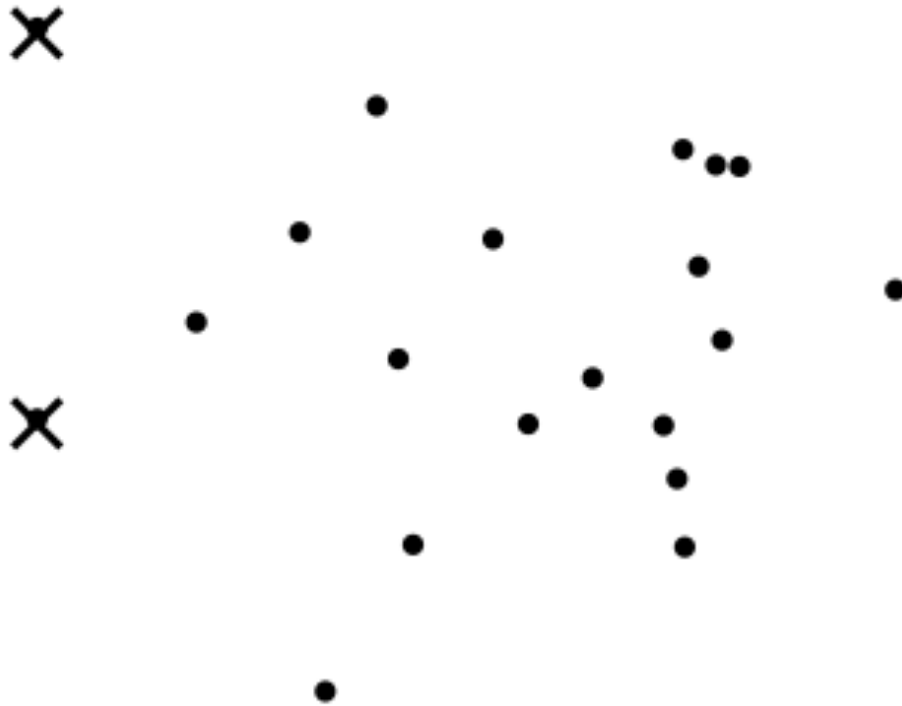
K -均值聚类算法

```
 $K$ -MEANS( $\{\vec{x}_1, \dots, \vec{x}_N\}, K$ )  
1   $(\vec{s}_1, \vec{s}_2, \dots, \vec{s}_K) \leftarrow \text{SELECTRANDOMSEEDS}(\{\vec{x}_1, \dots, \vec{x}_N\}, K)$   
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  $\{\vec{\mu}_1, \dots, \vec{\mu}_K\}$ 
```

例子



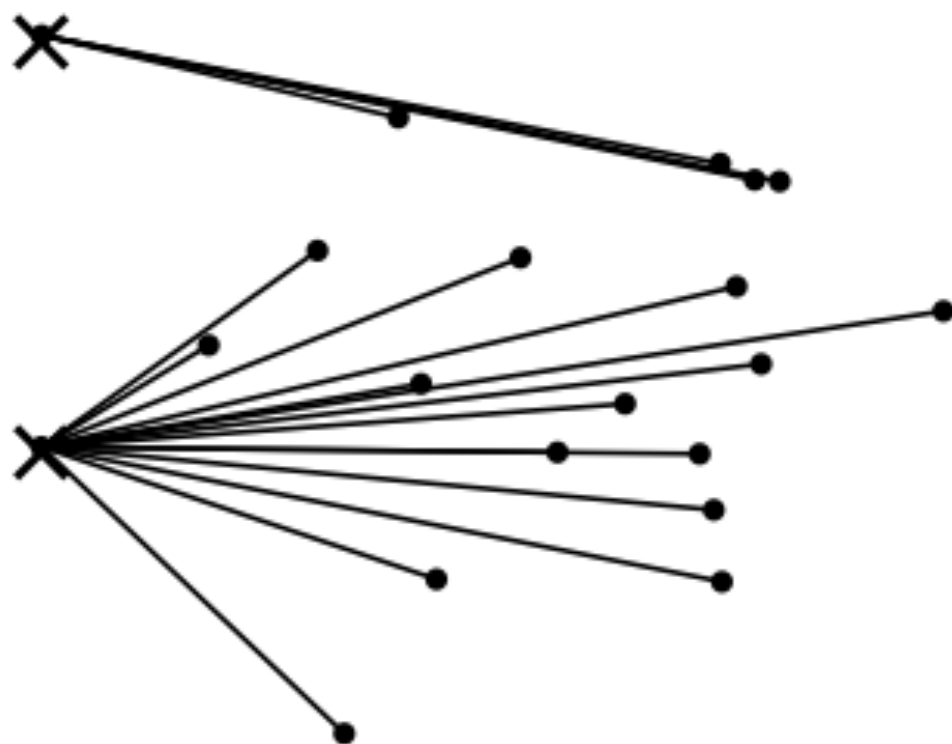
例子：随机选择两个种子($K=2$)



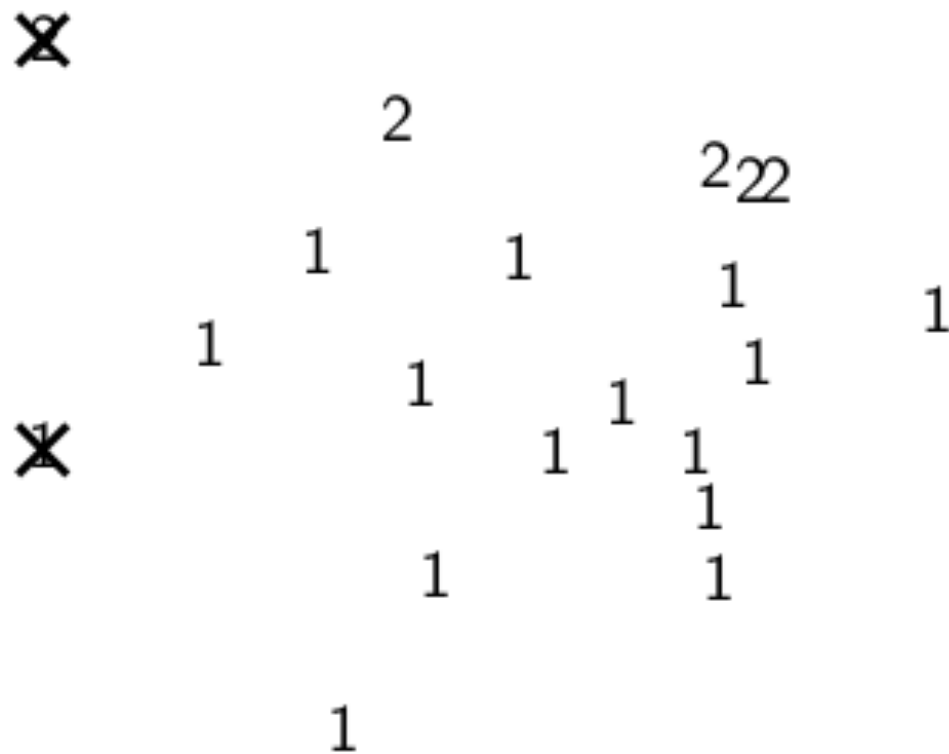
课堂练习: (i) 猜猜最后划分的两个簇是什么?

(ii) 计算簇的质心向量

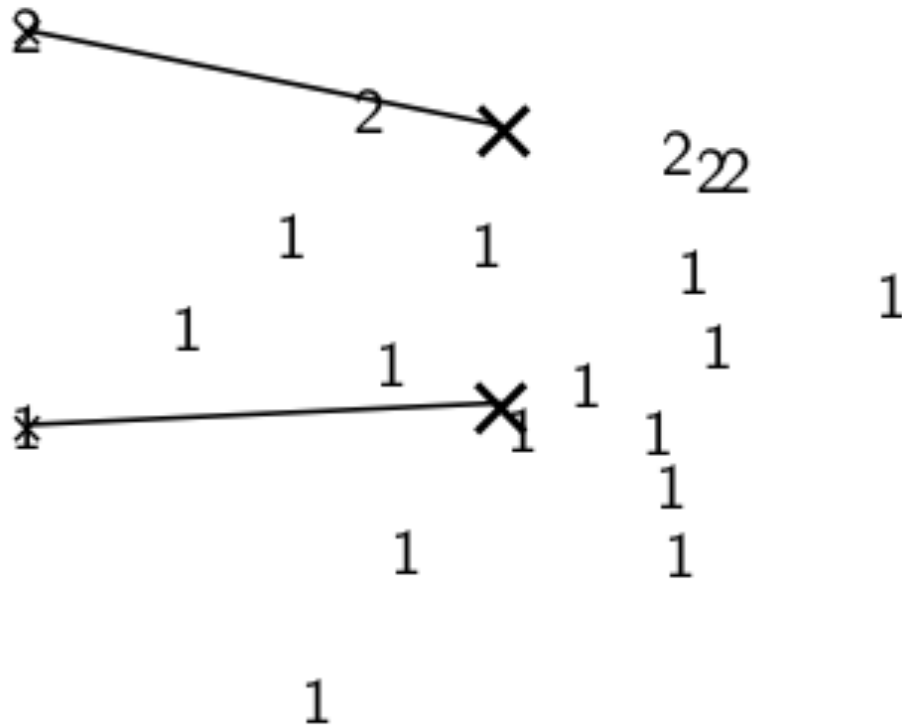
例子：将文档分配给离它最近的质心向量(第一次)



例子：分配后的簇(第一次)



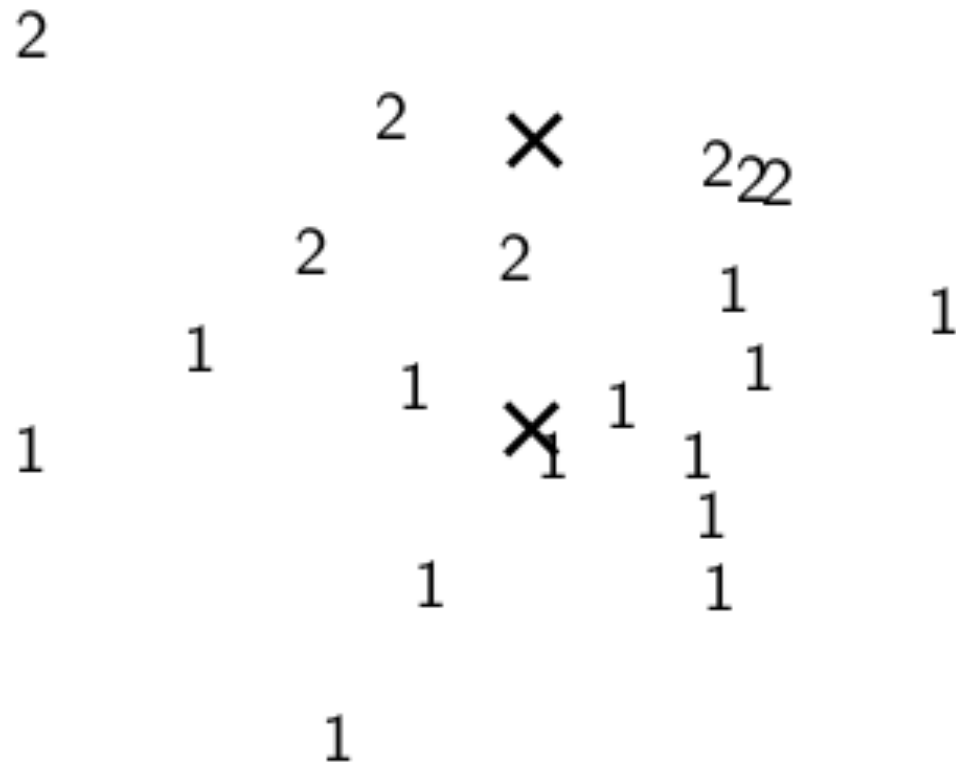
例子：重新计算质心向量



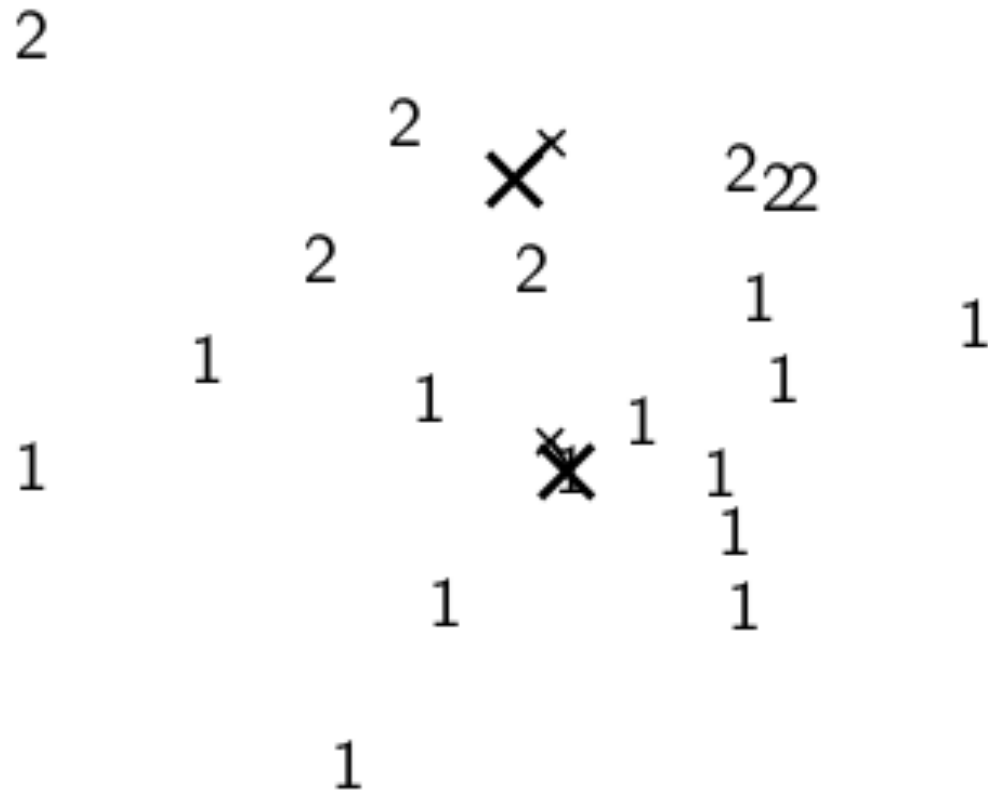
例子：将文档分配给离它最近的质心向量(第二次)



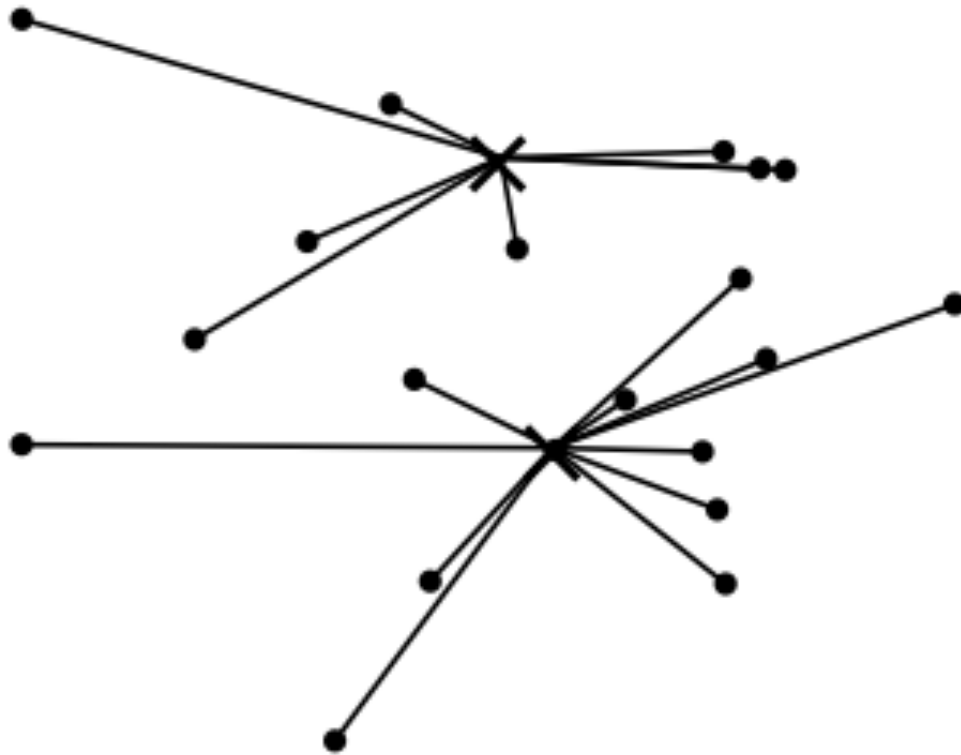
例子：重新分配的结果



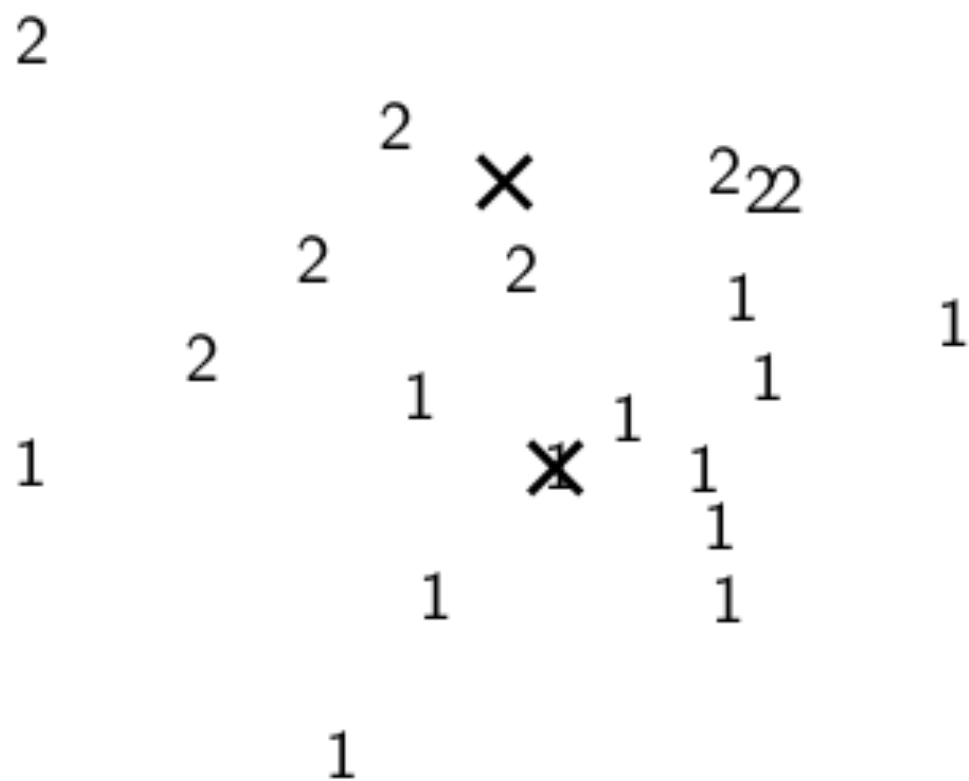
例子：重新计算质心向量



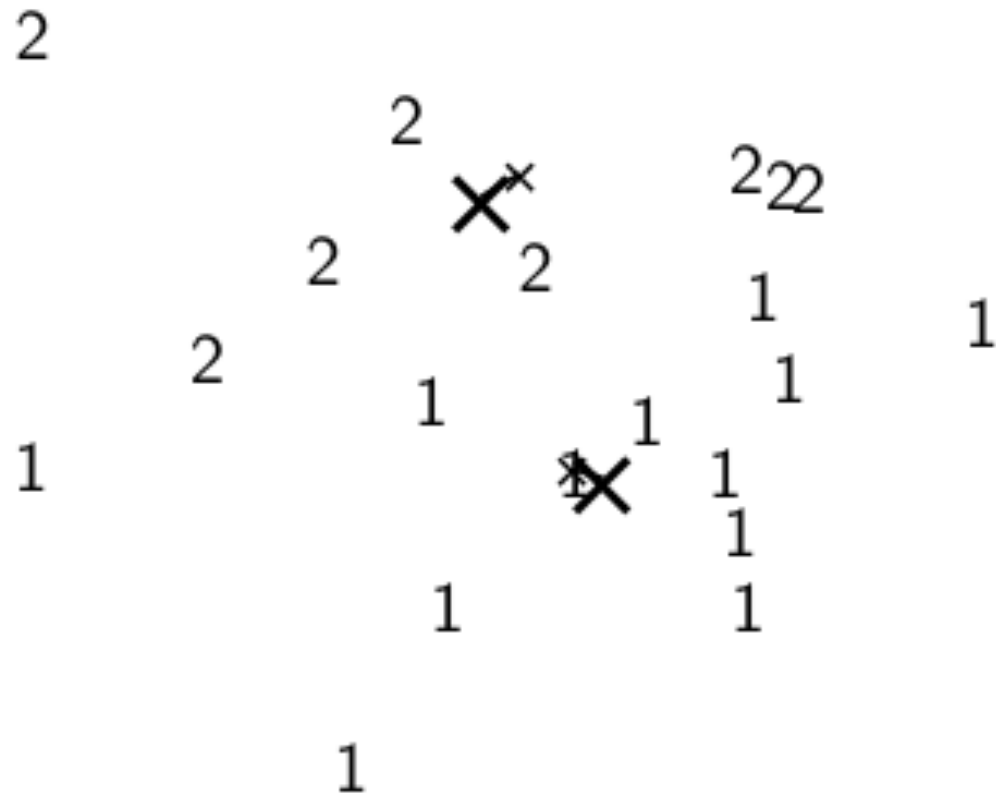
例子：再重新分配(第三次)



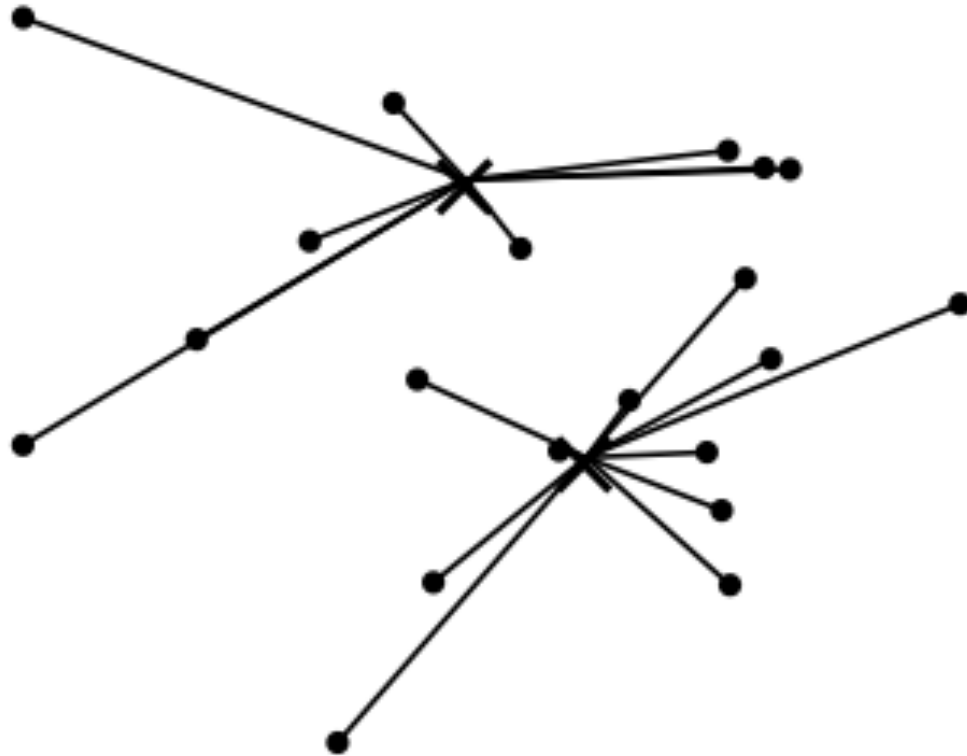
例子：分配结果



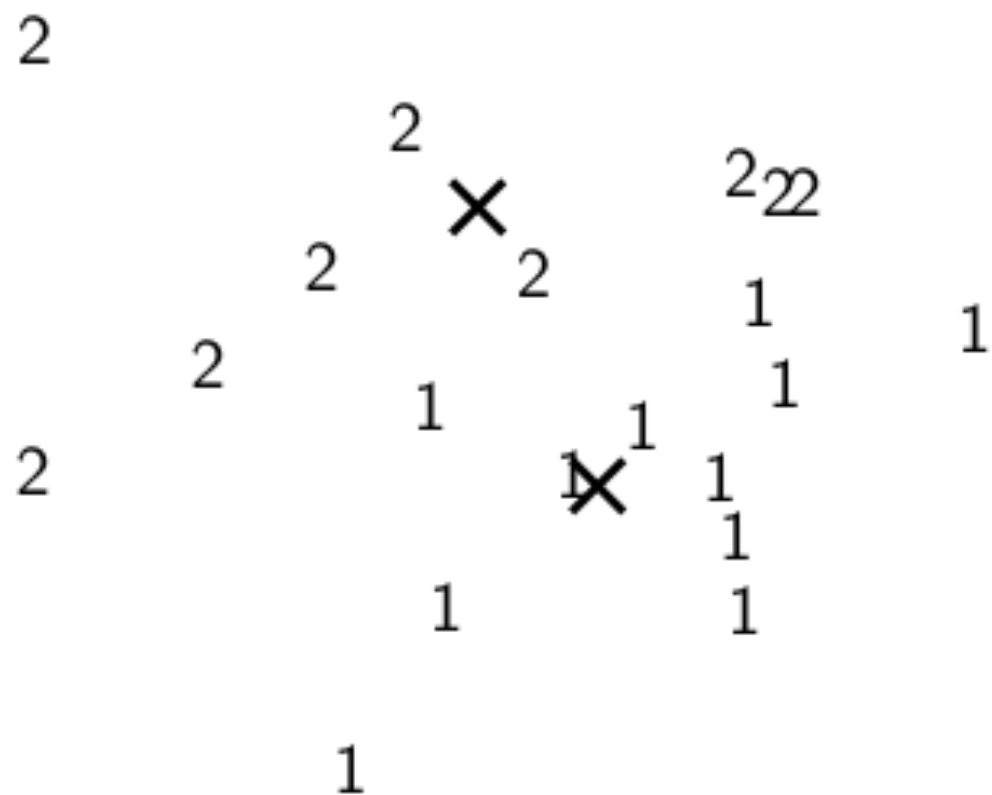
例子：重新计算质心向量



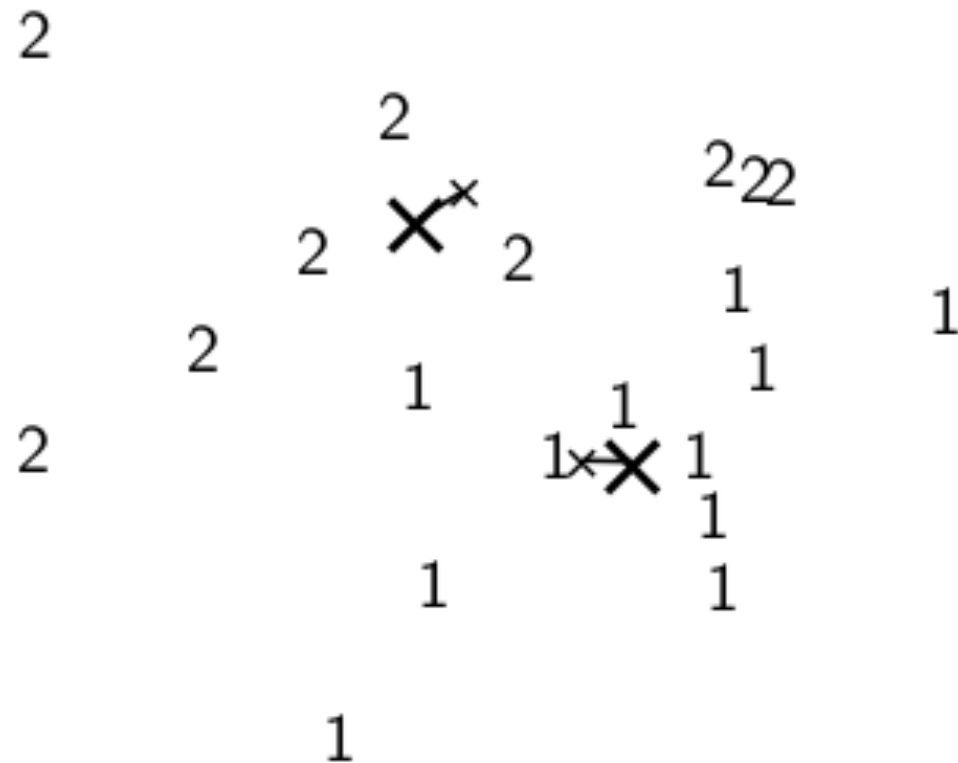
例子：再重新分配(第四次)



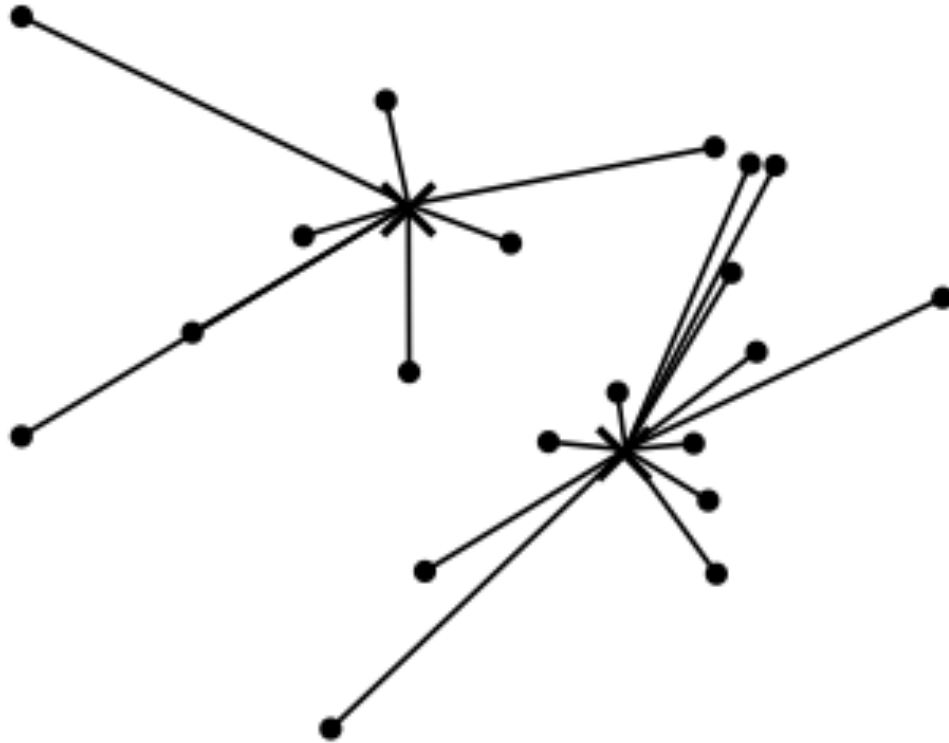
例子：分配结果



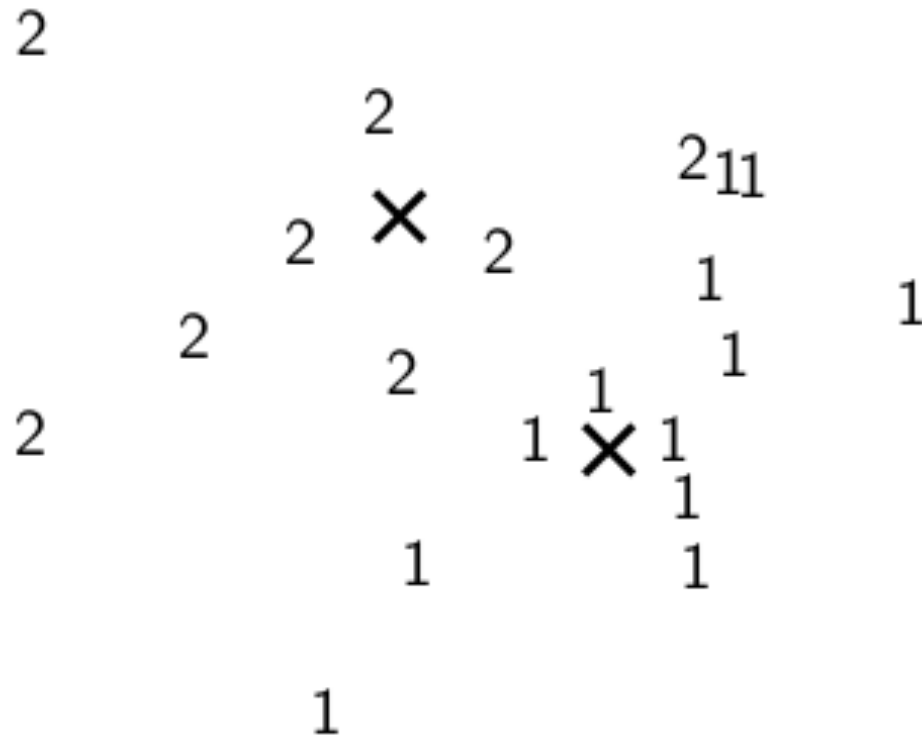
例子：重新计算质心向量



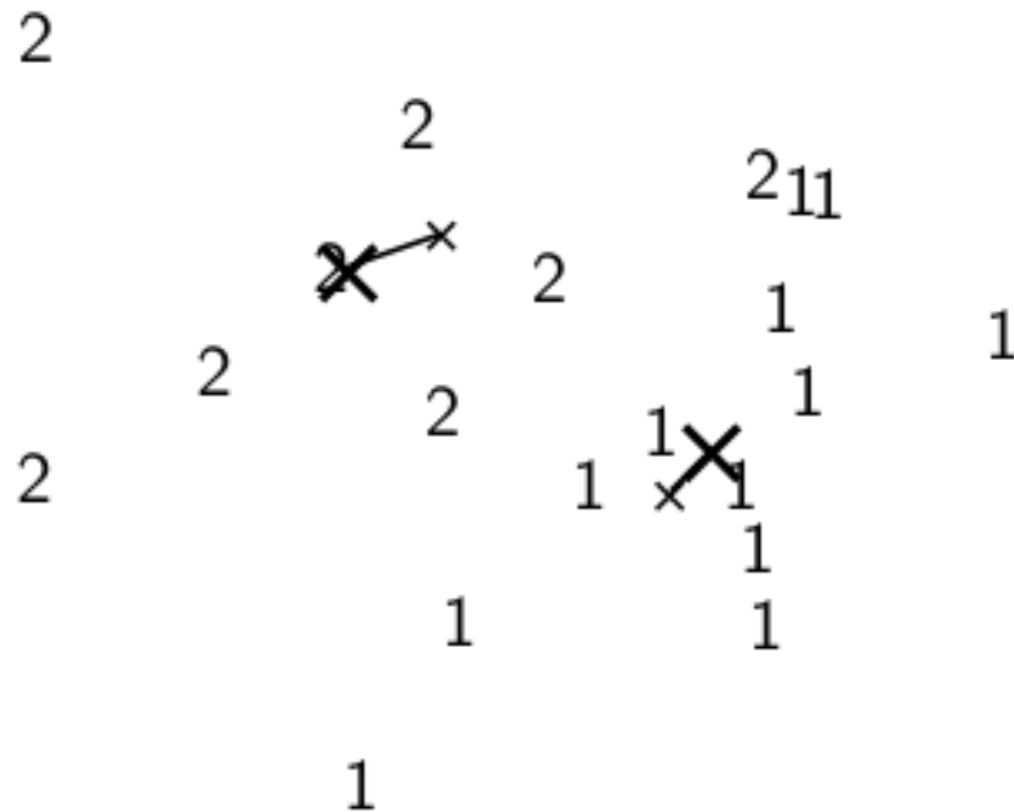
例子：重新分配(第五次)



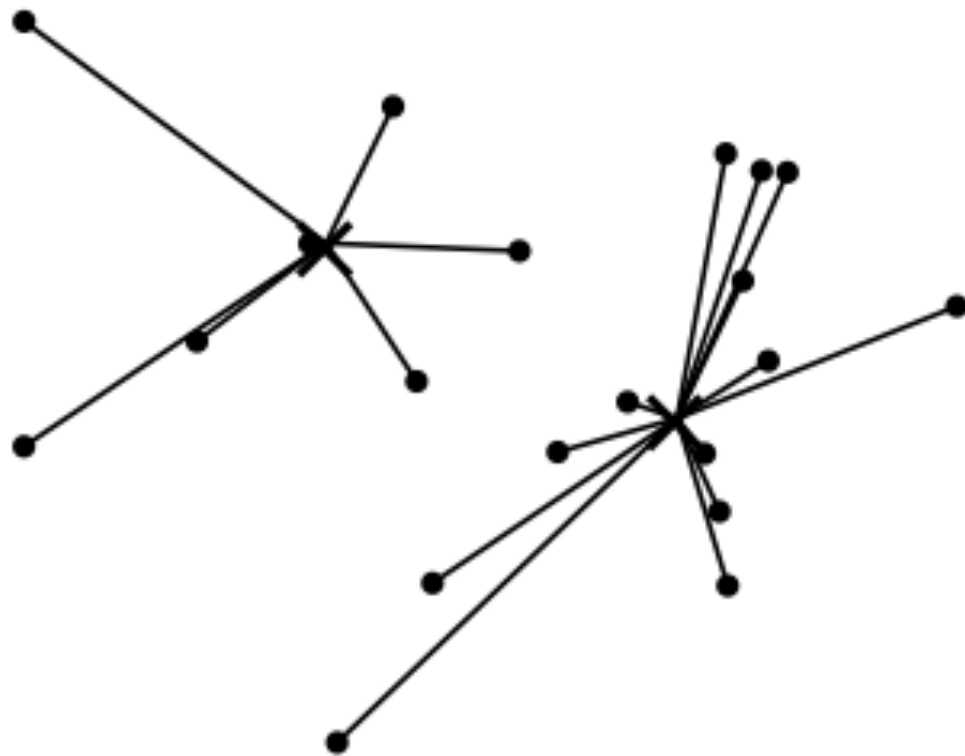
例子：分配结果



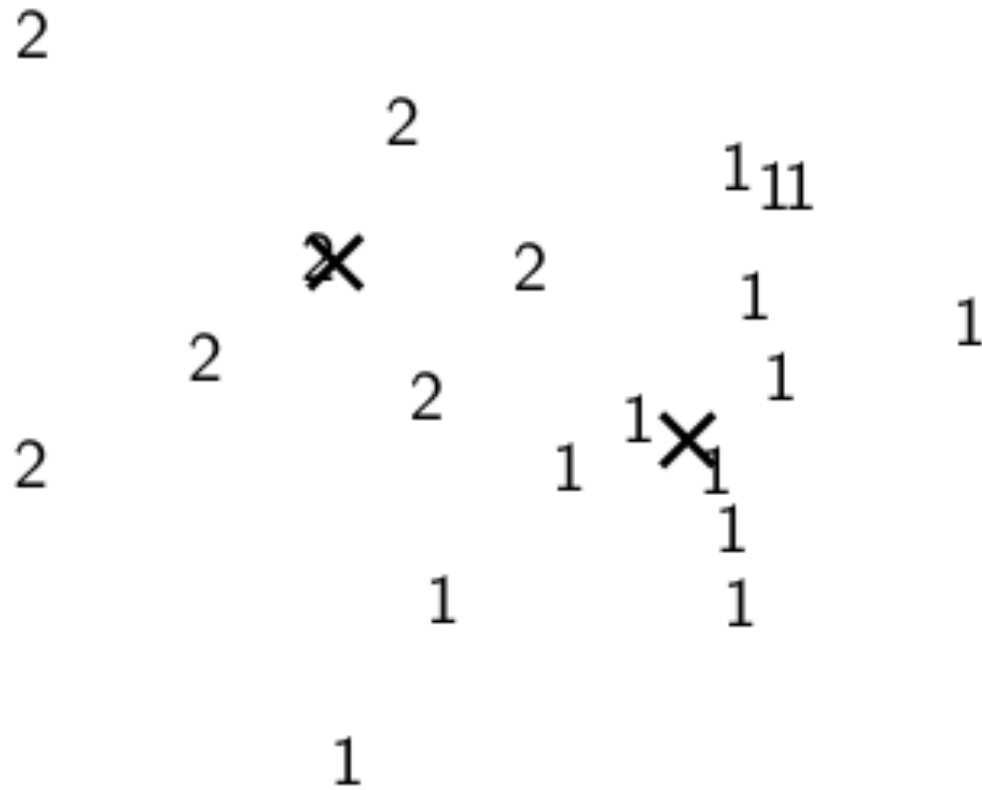
例子：重新计算质心向量



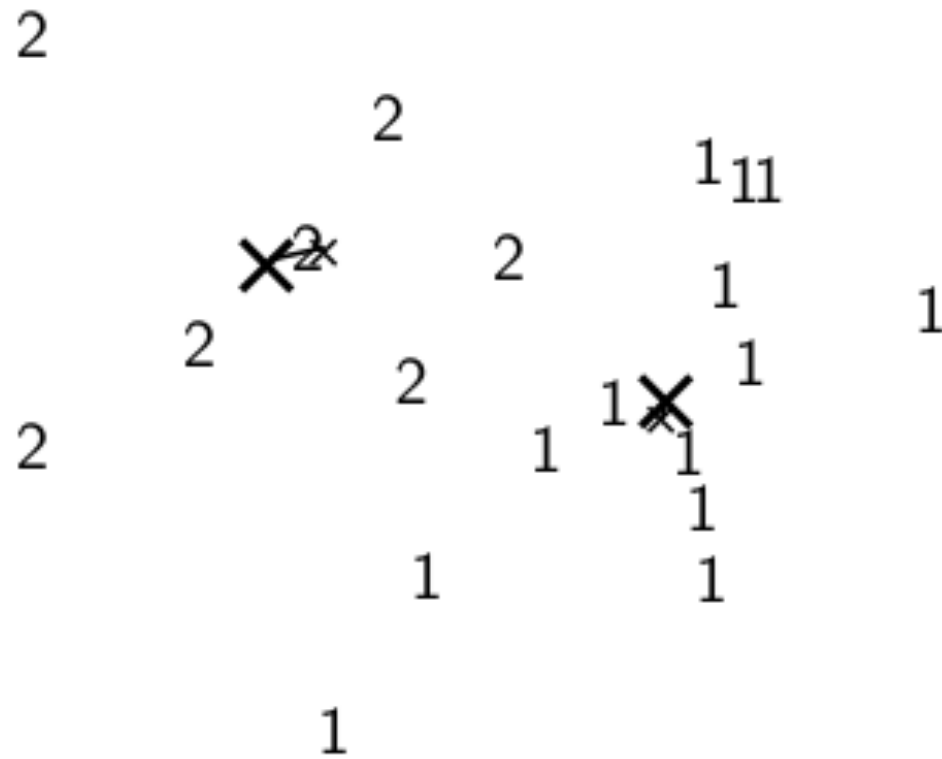
例子：重新分配(第六次)



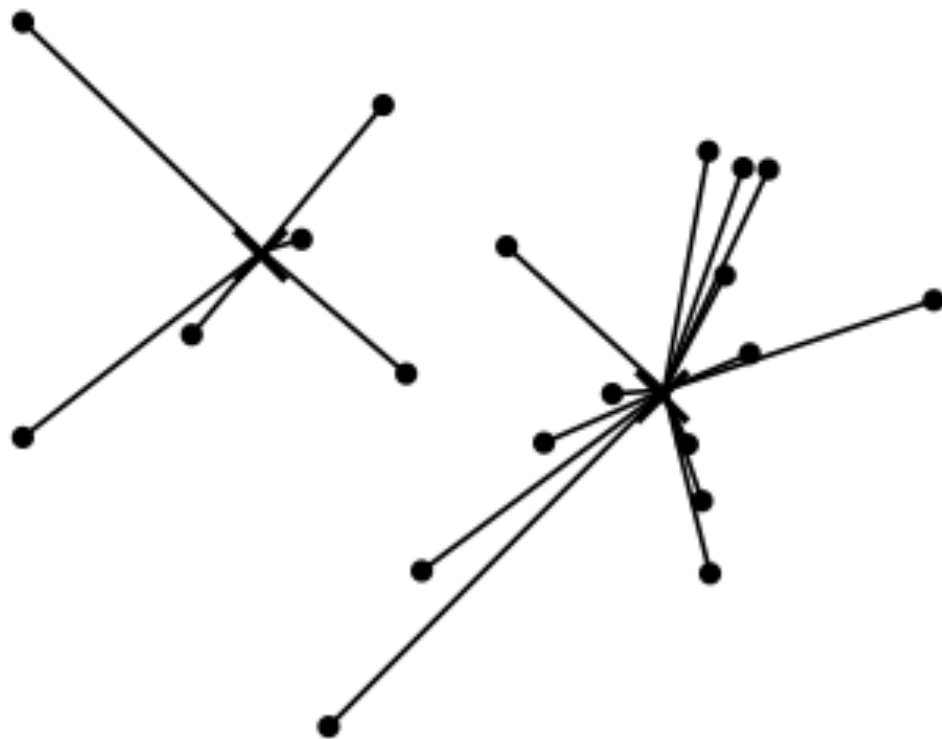
例子：分配结果



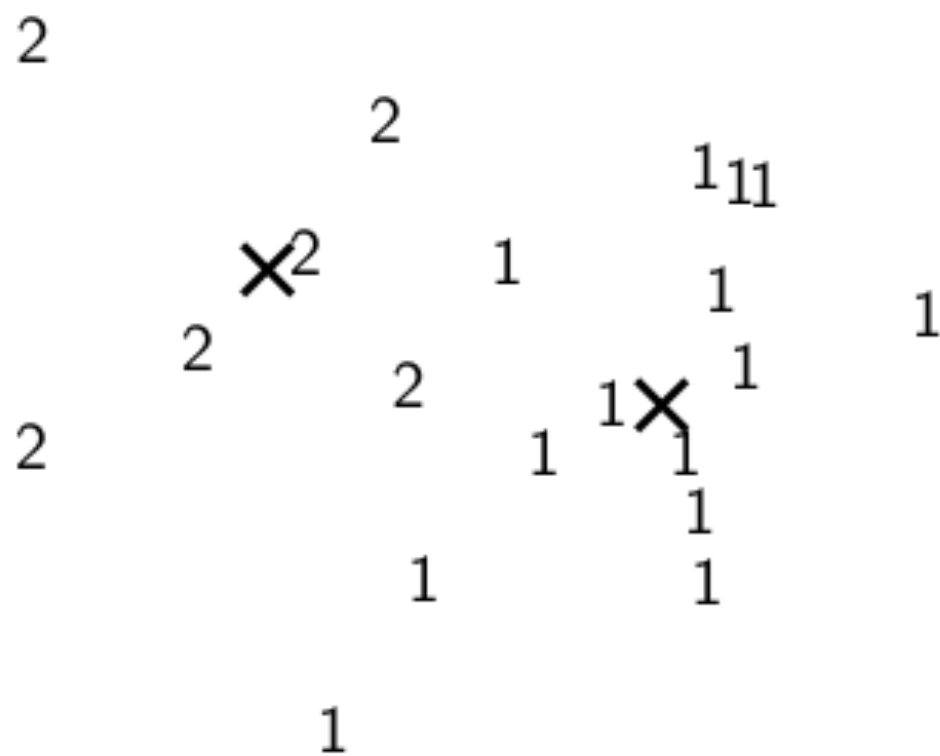
例子：重新计算质心向量



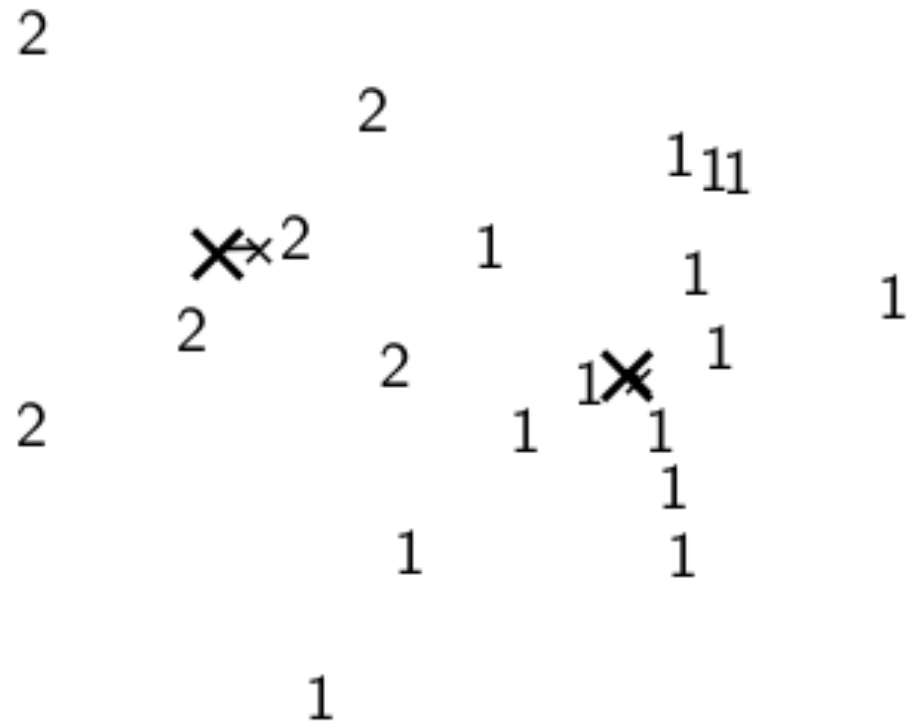
例子：重新分配(第七次)



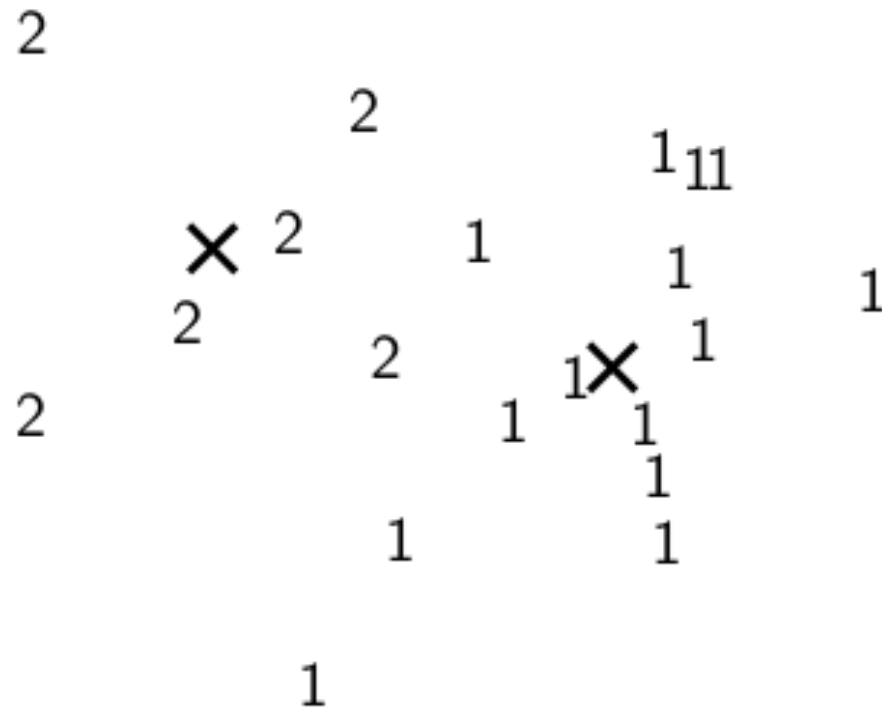
例子：分配结果



例子：重新计算质心向量



质心向量和分配结果最终收敛



K-均值聚类算法一定会收敛: 证明

- RSS = 所有簇上的文档向量到(最近的)质心向量的距离平方和的总和
- 每次重新分配之后 RSS 会下降
 - 这是因为每个向量都被移到离它最近的质心向量所代表的簇中
- 每次重新计算之后 RSS 也会下降
 - 参见下一页幻灯片
- 可能的聚类结果是有穷的
- 因此：一定会收敛到一个固定点
- 当然，这里有一个假设就是假定出现了等值的情况，算法都采用前后一致的方法来处理(比如，某个向量到两个质心向量的距离相等)

重新计算之后RSS也会下降的证明

$$\text{RSS} = \sum_{k=1}^K \text{RSS}_k$$

$$\text{RSS}_k(\vec{v}) = \sum_{\vec{x} \in \omega_k} \|\vec{v} - \vec{x}\|^2 = \sum_{\vec{x} \in \omega_k} \sum_{m=1}^M (v_m - x_m)^2$$

$$\frac{\partial \text{RSS}_k(\vec{v})}{\partial v_m} = \sum_{\vec{x} \in \omega_k} 2(v_m - x_m) = 0$$

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

这正好是基于每个向量分量来计算的质心的定义。因此，当将旧质心替换为新质心时，我们让 RSS_k 极小化。重新计算之后，作为 RSS_k 之和的RSS一定也会下降。

K -均值聚类算法一定是收敛的

- 但是不知道达到收敛所需要的时间!
- 如果不太关心少许文档在不同簇之间来回交叉的话，收敛速度通常会很快 (< 10-20次迭代)
- 但是，完全的收敛需要多得多的迭代过程

K-均值聚类算法的最优性

- 收敛并不意味着会达到全局最优的聚类结果!
- 这是K-均值聚类算法的最大缺点之一
- 如果开始的种子选的不好，那么最终的聚类结果可能会非常糟糕

K-均值聚类算法的初始化

- 种子的随机选择只是K-均值聚类算法的一种初始化方法之一
- 随机选择不太鲁棒：可能会获得一个次优的聚类结果
- 一些确定初始质心向量的更好办法：
 - 非随机地采用某些启发式方法来选择种子(比如，过滤掉一些离群点，或者寻找具有较好文档空间覆盖度的种子集合)
 - 采用层级聚类算法寻找好的种子
 - 选择 i (比如 $i = 10$) 次不同的随机种子集合，对每次产生的随机种子集合运行K-均值聚类算法，最后选择具有最小RSS值的聚类结果

K-均值聚类算法的时间复杂度

- 计算两个向量的距离的时间复杂度为 $O(M)$.
- 重分配过程: $O(KNM)$ (需要计算 KN 个文档-质心的距离)
- 重计算过程: $O(NM)$ (在计算质心向量时, 需要累加簇内的文档向量)
- 假定迭代次数的上界是 I
- 整体复杂度: $O(IKNM)$ - 线性
- 但是, 上述分析并没有考虑到实际中的最坏情况
- 在一些非正常的情况下, 复杂度可能会比线性更糟

提纲

- ① 上一讲回顾
- ② 聚类介绍
- ③ 聚类在IR中的应用
- ④ K-均值聚类算法
- ⑤ 聚类评价
- ⑥ 簇个数确定

怎样判断聚类结果的好坏?

- 内部准则 (Internal criteria)

- 一个内部准则的例子: K-均值聚类算法的RSS值

- 但是内部准则往往不能评价聚类在实际应用中的实际效用

- 替代方法: 外部准则 (External criteria)

- 按照用户定义的分类结果来评价, 即对一个分好类的数据集进行聚类, 将聚类结果和事先的类别情况进行比较, 得到最后的评价结果

外部准则

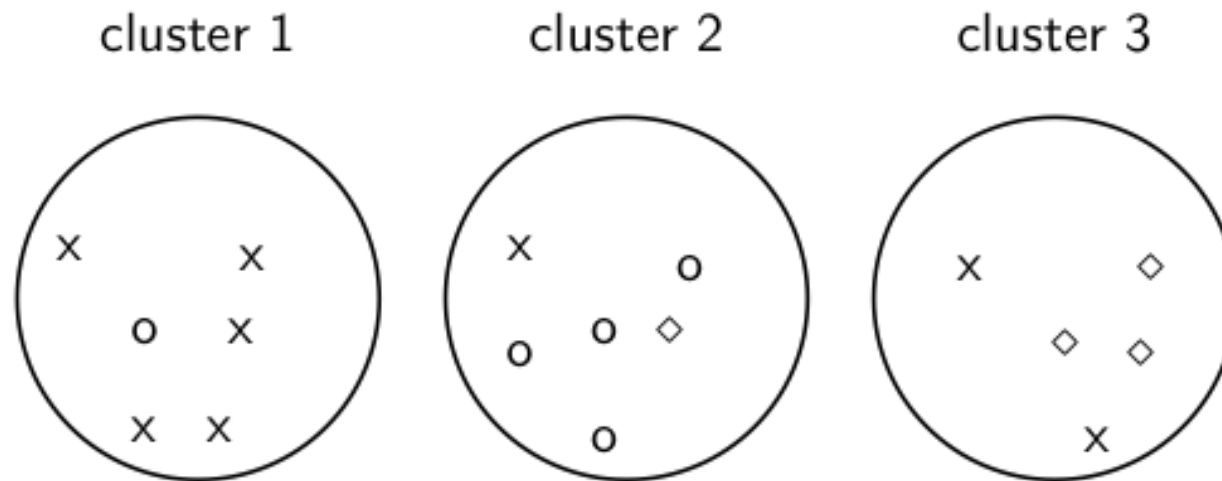
- 基于已有标注的标准数据集(如Reuters语料库)来进行聚类评价
- 目标：聚类结果和给定分类结果一致
- (当然，聚类中我们并不知道最后每个簇的标签，而只是关注如何将文档分到不同的组中)
- 一个评价指标：纯度([purity](#))

外部准则: 纯度

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

- $\Omega = \{\omega_1, \omega_2, \dots, \omega_K\}$ 是簇的集合
- $C = \{c_1, c_2, \dots, c_J\}$ 是类别的集合
- 对每个簇 ω_k : 找到一个类别 c_j , 该类别包含 ω_k 中的元素最多, 为 n_{kj} 个, 也就是说 ω_k 的元素最多分布在 c_j 中
- 将所有 n_{kj} 求和, 然后除以所有的文档数目

纯度计算的例子



纯度?

$\max_j |\omega_1 \cap c_j| = 5$ (class x, cluster 1);

$\max_j |\omega_2 \cap c_j| = 4$ (class o, cluster 2);

$\max_j |\omega_3 \cap c_j| = 3$ (class \diamond , cluster 3)

纯度为 $(1/17) \times (5 + 4 + 3) \approx 0.71$.

提纲

- ① 上一讲回顾
- ② 聚类介绍
- ③ 聚类在IR中的应用
- ④ K-均值聚类算法
- ⑤ 聚类评价
- ⑥ 簇个数确定

簇个数确定

- 在很多应用中，簇个数 K 是事先给定的
 - 比如，可能存在对 K 的外部限制
 - 例子：在“分散-集中”应用中，在显示器上(上世纪90年代)很难显示超过10-20个簇
- 如果没有外部的限制会怎样？是否存在正确的簇个数？
- 一种办法：定义一个优化准则
 - 给定文档，找到达到最优情况的 K 值
 - 能够使用的最优准则有哪些？
 - 我们不能使用前面所提到的RSS或到质心的平均平方距离等准则，因为它们会导致 $K = N$ 个簇

课堂练习

- 你的任务是开发一个聚类算法来和news.google.com竞争
- 你想使用K-均值聚类算法
- 如何确定 K ?

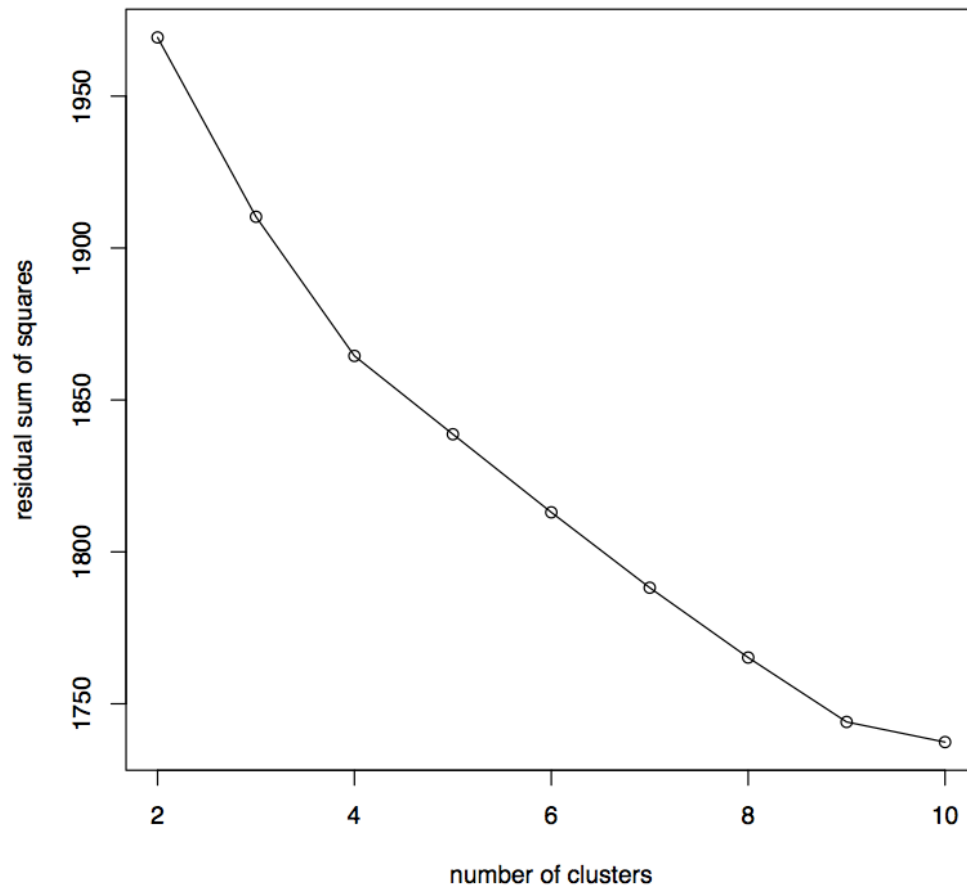
简单的目标函数 (1)

- 基本思路:
 - 从1个簇开始 ($K = 1$)
 - 不断增加簇 (= 不断增大 K)
 - 对每个新的簇增加一个惩罚项
- 在惩罚项和RSS之间折中
- 选择满足最佳折中条件的 K

简单的目标函数 (2)

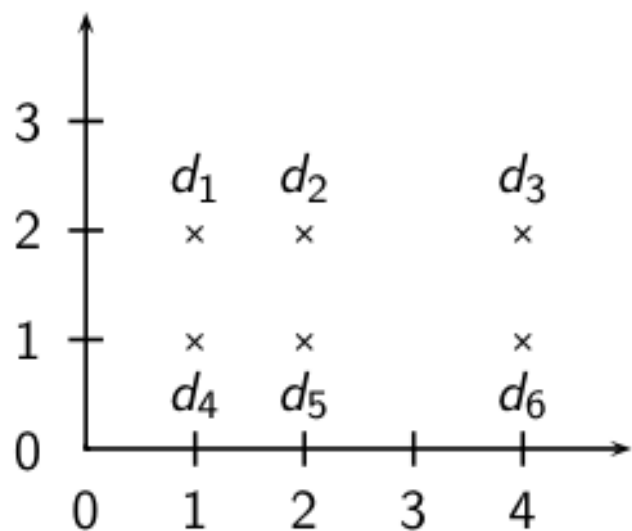
- 给定聚类结果，定义文档的代价为其到质心向量的(平方)距离（失真率）
- 定义全部失真率 $RSS(K)$ 为所有文档代价的和
- 然后：对每个簇一个惩罚项 λ
- 于是，对于具有 K 个簇的聚类结果，总的聚类惩罚项为 $K\lambda$
- 定义聚类结果的所有开销为失真率和总聚类惩罚项的和： $RSS(K) + K\lambda$
- 选择使得 $(RSS(K) + K\lambda)$ 最小的 K 值
- 当然，还要考虑较好的 λ 值 ...

在曲线中寻找拐点



本图中两个拐点：4 和 9

有关收敛性的课堂练习： 次优的聚类结果



- $K=2$ 情况下的最优聚类结果是什么？
- 对于任意的种子 d_i 、 d_j , 我们是否都会收敛于该聚类结果？

本讲小结

- 聚类的概念(What is clustering?)
- 聚类在IR中的应用
- K -均值(K -Means)聚类算法
- 聚类评价
- 簇(cluster)个数(即聚类的结果类别个数)确定

参考资料

- 《信息检索导论》第16章
- <http://ifnlp.org/ir>
 - K -均值聚类算法的例子
 - Keith van Rijsbergen有关聚类假设的论述
 - Bing/Carrot2/Clusty: 搜索结果聚类

致谢

- 本课件参考中科院计算所王斌老师的课件