网络信息检索

第十讲 网络信息检索应用（1）

董守斌
sbdong@scut.edu.cn

华南理工大学计算机学院
广东省计算机网络重点实验室
Communication & Computer Network Laboratory (CCNL)

主要内容

- 什么是分类和聚类?
- 文本分类
- 文本聚类

问题

- 分类/聚类的概念是什么？有什么关系？有什么区别？
- 为什么要进行（文本）分类/聚类？
- （文本）分类/聚类的过程如何？
- （文本）分类/聚类的方法有哪些？
- 如何评价（文本）分类/聚类的效果？

引言

- 分类/聚类是大自然的固有现象：物以类聚、人以群分
  - 相似的对象往往聚集在一起
  - （相对而言）不相似的对象往往分开

什么是分类？

- 简单地说，分类（**Categorization or Classification**）就是按照某种标准给对象贴标签（**label**）
- 性别、籍贯、民族、学历、年龄等等，我们每个人身上贴满了“标签”
- 我们从孩提开始就具有分类能力：爸爸、妈妈；好阿姨、坏阿姨；电影中的好人、坏人等等
- 分类无处不在，从现在开始，我们可以以分类的眼光看世界

什么是聚类?

- 简单地说，聚类是指事先没有“标签”而通过某种成团分析找出事物之间存在聚集性原因的过程
  - 几个自然班在一个大教室上上课，坐在一块的大都是一个班的
  - 事先不知道“标签”，根据对象之间的相似情况进行成团分析

为什么要进行分类和聚类？

- 分类/聚类的根本原因就是因为对象数目太多，处理困难
  - 一些信息处理部门，一个工作人员一天要看上千份信息
  - 分门别类将会大大减少处理难度，提高处理效率和效果

分类/聚类的过程

- 对对象进行表示
  - 表示方法
  - 特征选择
- 根据某种算法进行相似度计算
  - 相似度计算方法
  - 分类/聚类方法

主要内容

- 什么是分类和聚类?
- 文本分类
  - 文本分类的定义
  - 文本分类的方法
  - 文本分类的评估
- 文本聚类

文本分类的定义

- 文本分类（**Text Categorization/Classification**）：事先给定分类体系和训练样例（标注好类别信息的文本），将文本分到某个或者某几个类别中
  - 计算机自动分类，就是根据已经标注好类别信息的训练集合进行学习，将学习到的规律用于新样本(也叫测试样本)的类别判定
  - 分类是有监督/指导学习（**Supervised Learning**）的一种

文本分类的应用

- 网页：
  - 推荐（Recommending）
  - 类似Yahoo的分类目录
- 新闻消息
  - 新闻出版按照栏目分类：类别{政治,体育,军事,...}
  - 个性化推荐
  - 垃圾过滤
- 邮件的判定
  - 垃圾，类别{spam, not-spam}
- 中文分析
  - 词性标注：类别{名词,动词,形容词,...}
  - 词义排歧：类别{词义1,词义2,...}
- 其他：论文分类、产品分类……

分类例子


Sogou搜狗
目录
新闻 网页 音乐 目录 地图 说吧 更多»
搜 索
直通车
50,000主题分类，500,000优选网站，搜狐倾力推荐
娱乐休闲
音乐, 影视, 美女, 写真, MP3...
电脑网络
壁纸, 游戏, 下载, 手机...
卫生健康
减肥, 生殖, 养生, 疾病...
工商经济
农业, 理财, 股票, 外汇, 房产...
教育培训
大学, 高考, 校友录, 英语...
生活服务
美容, 交友, 旅游, 订票...
公司企业
留学, 汽车, 机械, 影院...
艺术
摄影, 风景, 绘画, 书法...
社会文化
宗教, 星座, 环保, 节假日...
文学
小说, 言情, 作家, 武侠...
新闻媒体
报纸, 新闻, 电视, 图片...
政法军事
军事, 飞机, 武器, 军情...
体育健身
足球, 棋牌, NBA, 围棋...
科学技术
数学, 物理, 化学, 生物...
社会科学
心理, 伦理, 测试, 宗教...

分类例子


Make Y! your home p
Yahoo!
http://www.yahoo.com/
YAHOO!
Yahoo! Autos - Get pricing & user reviews today.
Web | Images | Video | Local | Shopping | more
Search:
Web Search
the Web
China only
Yahoo! Home
My Yahoo!
Y! China
Apr 1, 2008
Page Options
Answers
Autos UPDATED!
Finance
Games
Groups
HotJobs
Maps
Mobile Web
Movies
Music
Personals
Real Estate
Shopping
Sports
Tech
Travel
TV
Yellow Pages
Shine NEW!
More Yahoo! Services
Featured
Entertainment
Sports
Video
Texas edge goes to Obama
Hillary Clinton won the popular vote in Texas, but the delegate count swings in her opponent's favor. » Latest total
• Michigan vote compromise offered
• Find more buzzing political stories
Buzzing Now
Texas delegate edge swings to Obama
Tennis star Roddick to marry 20-year-old model
'Star Wars,' as explained by a 3-year-old
1980s R&B singer Sean Levert dies at 39
» More: Featured | Buzz
In the News
World
Local
Finance
As of 9:10 a.m.
• Clinton accuses Obama of trying to stop people from voting
• No sign Iraq clashes will affect U.S. drawdown, Gates says
• Pentagon staffer guilty of passing military secrets to China
• U.S. gas prices hit new high; 58 cents higher than last year
• Postal Service workers learn to defend against dog attacks
• 4,000-year-old gold necklace discovered in Peru burial site
• Heath Ledger may have fathered another child, report says
• MLB · March Madness · NBA · NHL · NASCAR · Soccer · Golf
» More: News | Popular | Election '08
Markets: Dow: +0.4% Nasdaq: +0.8%
Sponsored by: Scottrade
Check your mail status: Sign In
Free mail: Sign Up
Mail
Messenger
Radio
Weather
Local
Horoscopes
SILVERADO
GET THE TRUE LOWDOWN ON SILVERADO
WATCH AND LEARN MORE
Visit chevy.com - Ad Feedback
Inside Yahoo! Music
Watch Hot New Music Videos
Rihanna
Chris Brown
Britney Spears
Taylor Swift
Small Business
• Get a Web Site
• Domain Names
• Sell Online
• Search Ads
Featured Services
• Downloads
• Health
Marketplace
Alibaba.com
Expand your business globally
Buy and sell products on the worlds leading B2B online marketplace. Set up a free account now.
Jewelry and Watches at Buy.com. Shopping for the ultimate sparkler, we bring you the lowest prices.
Scottrade delivers fast trading tools, low fee trades, powerful research tools and top-rated customer service
Pulse - What Yahoos Are Into
Popular Spring Break Beachwear
• Roxy
• Billabong
• O'Neill
• Guess
• Nike
• Shoshanna
• Mossimo
• Hurley
• Speedo
• Quicksilver
» More Yahoo! Shopping

思考题

- 信息检索是否也可以看成分类问题？
- 如果可以看成分类的话，相关反馈的作用是什么？
- 中文分词也是一种分类问题？

主要内容

- 什么是分类和聚类?
- 文本分类
  - 文本分类的定义
  - 文本分类的方法
  - 文本分类的评估
- 文本聚类

分类的定义

- 给定：
  - 一个实例的描述， $x \in X$，这里X是实例空间（instance space）
  - 一系列类别集合： $C=\{c_1, c_2,\ldots c_n\}$
- 决定：
  - x的类别：$c(x) \in C$， 这里 $c(x)$ 是一个分类函数（categorization function），其域是X，取值范围是C

分类问题举例

- 实例空间： <size, color, shape>
  - size ∈ {small, medium, large}
  - color ∈ {red, blue, green}
  - shape ∈ {square, circle, triangle}
- *C* = {positive, negative}
- *D*:

Example	Size	Color	Shape	Category
1	small	red	circle	positive
2	large	red	circle	positive
3	small	red	triangle	negative
4	large	blue	circle	negative

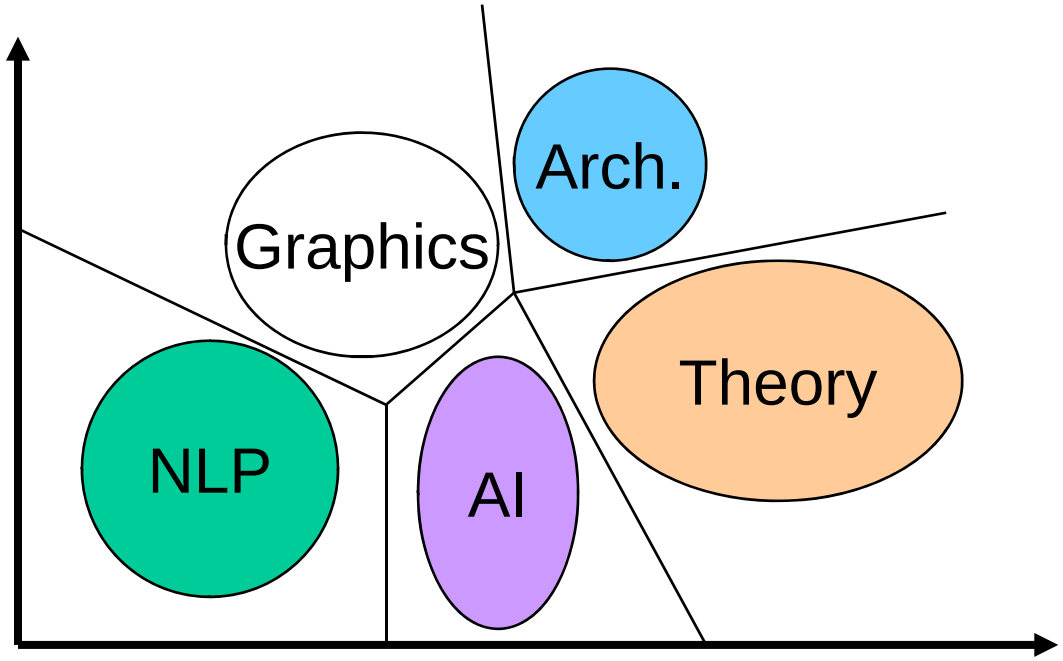
分类的学习方法

- 一个训练例子（**training example**）是一个实例 $x \in X$ 以及正确的类别 $c(x)$ 的值对：$<x, c(x)>$，c 是一个未知的分类函数
- 给定一些训练例子 D，试图找到一个分类函数 $h(x)$，使得：

$$\forall < x, c(x) > \in D : h(x) = c(x)$$

一致性

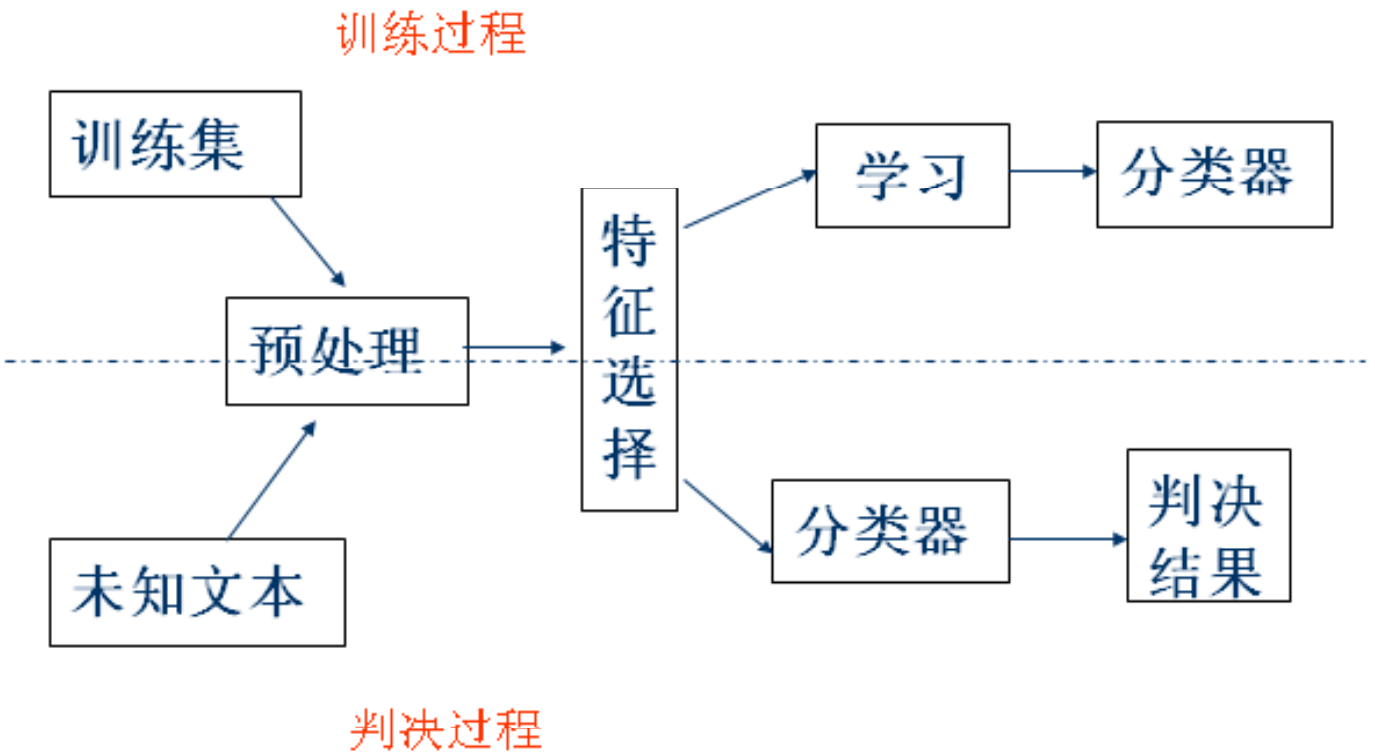
分类问题图示



文本分类的学习方法

- 人工寻找文本分类的函数是非常困难的
- 机器学习（**Machine Learning**）算法：
  - 相关反馈（**Rocchio**）
  - 朴素贝叶斯（**Naïve Bayesian**）
  - 最近邻（**Nearest Neighbor，KNN**）
  - 神经网络（Neural Network，NN）
  - 支持向量机（Support Vector Machines，SVM）
  - 决策树（Decision Tree， DT）

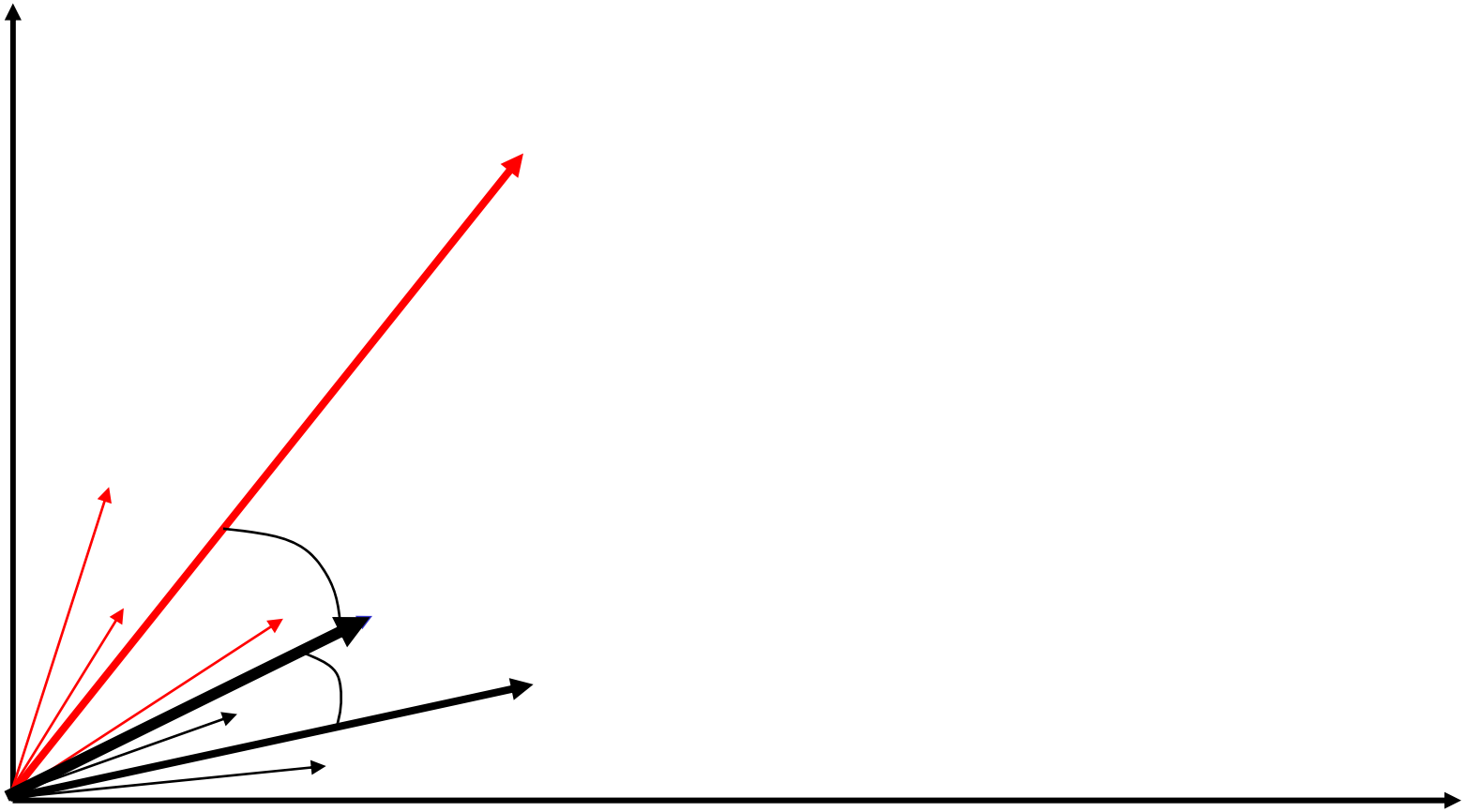
自动分类的一般过程



使用相关反馈（Rocchio）

- 相关反馈方法可以用于文本分类
- 用标准的归一化TF/IDF权重向量表示文本文档
- 对每个类别，通过对计算类别中的训练文档的矢量求和得到一个类向量（prototype vector），或称原型向量
- 基于cosine相似度，如果测试（test）文档与某个类向量最近，则赋予该文档到此类别

Rocchio文本分类图示



Rocchio 的文本分类算法（训练）

Assume the set of categories is $\{c_1, c_2, \ldots c_n\}$
For i from 1 to n let $p_i = <0, 0, \ldots, 0>$ *(init. prototype vectors)*
For each training example $<x, c(x)> \in D$
Let d be the frequency normalized TF/IDF term vector for doc x
Let $i = j: (c_j = c(x))$
(sum all the document vectors in c_i to get p_i)
Let $p_i = p_i + d$

Rocchio 的文本分类算法（判决）

Given test document *x*
Let *d* be the TF/IDF weighted term vector for *x*
Let *m* = –2 ***(init. maximum cosSim)***
For *i* from 1 to *n*:
    (compute similarity to prototype vector)
    Let *s* = cosSim(d, p_i)
    if *s* > *m*
        let *m* = *s*
        let *r* = c_i ***(update most similar class prototype)***
Return class *r*

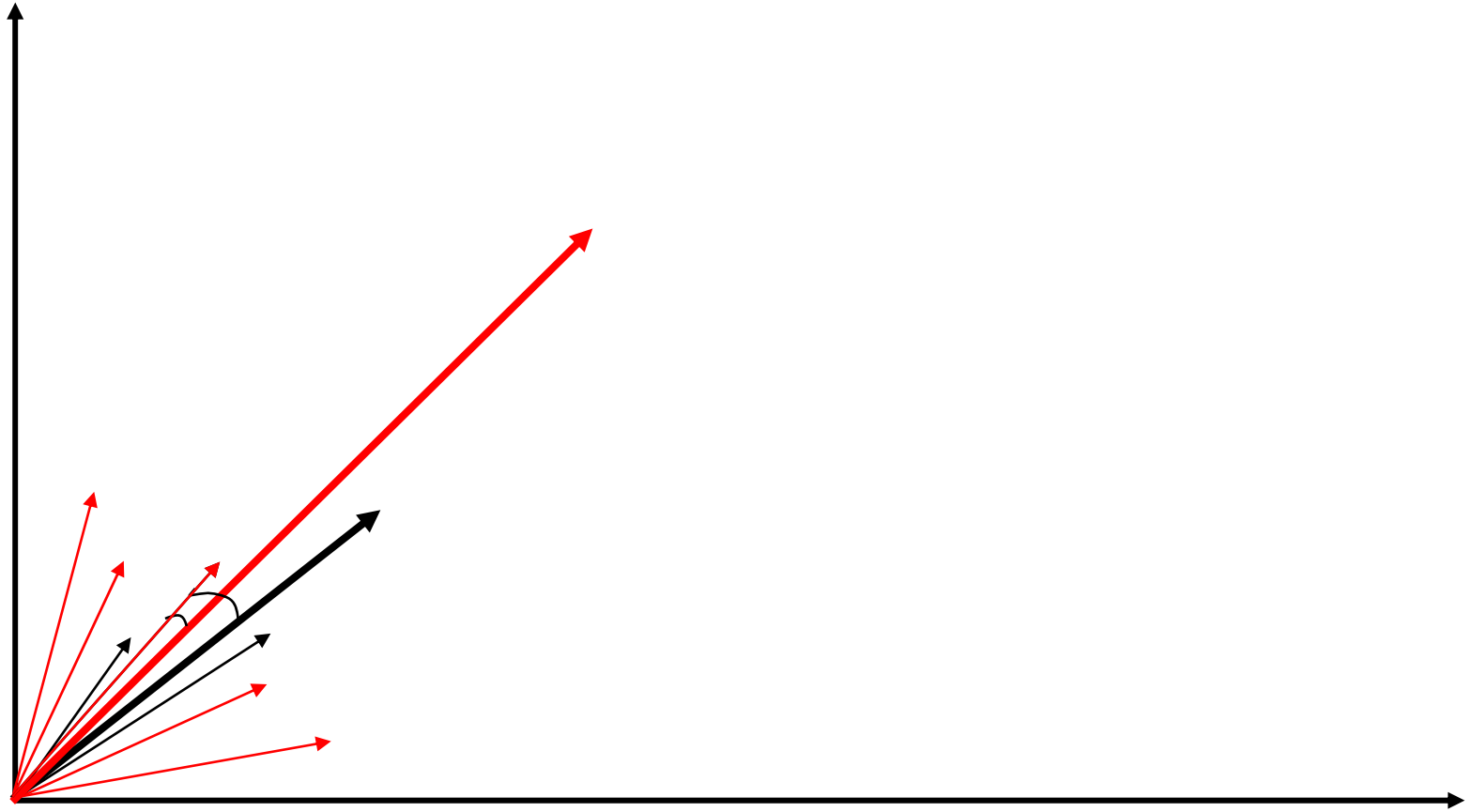
Rocchio的性质

- 对每个类别构建一个类向量
- 类矢量无需归一化因为cosine相似度对长度不敏感
- 分类是基于与类向量的相似度

- 但不能保证满足一致性的要求

Rocchio的异常性

- 在多态类（**polymorphic/disjunctive categories**）的情形下，类向量模型可能有问题



最近邻学习算法
（Nearest-Neighbor Learning Algorithm）

- 学习过程很简单，只是存储***D***中训练样本的表达式
- 对于测试实例***x***：
  - 计算***x*** 和***D***中所有测试实例的相似度
  - 如果某个类别包含那个最相似的实例，则 ***x*** 属于该类
- 也称为：
  - 基于事例（Case-based）的方法
  - 基于内存（Memory-based）的方法
  - 懒惰学习（Lazy learning）方法

K个近邻（Nearest-Neighbor）

- 只用最近的实例来决定类别可能会导致错误：
  - 最近的实例可能只是一个非典型的例子
  - 可能只是噪音导致的标注错误
- 更鲁棒性的选择是找 ***k*** 个最相似的实例，看哪个类包含的最多
- ***k*** 的值要设为奇数， 一般取3或5

相似测度

- 最近邻方法依赖于所采用的相似测度（similarity metric），如
  - 欧拉距离（Euclidian distance ）
  - 汉明距离（Hamming distance ）
- 对于文本分类，TF-IDF 权重向量的cosine相似度通常是最有效的

用于文本分类的K近邻算法

Training:

For each each training example $\langle x, c(x) \rangle \in D$

Compute the corresponding TF-IDF vector, d_x, for document x

Test instance *y*:

Compute TF-IDF vector d for document y

For each $\langle x, c(x) \rangle \in D$

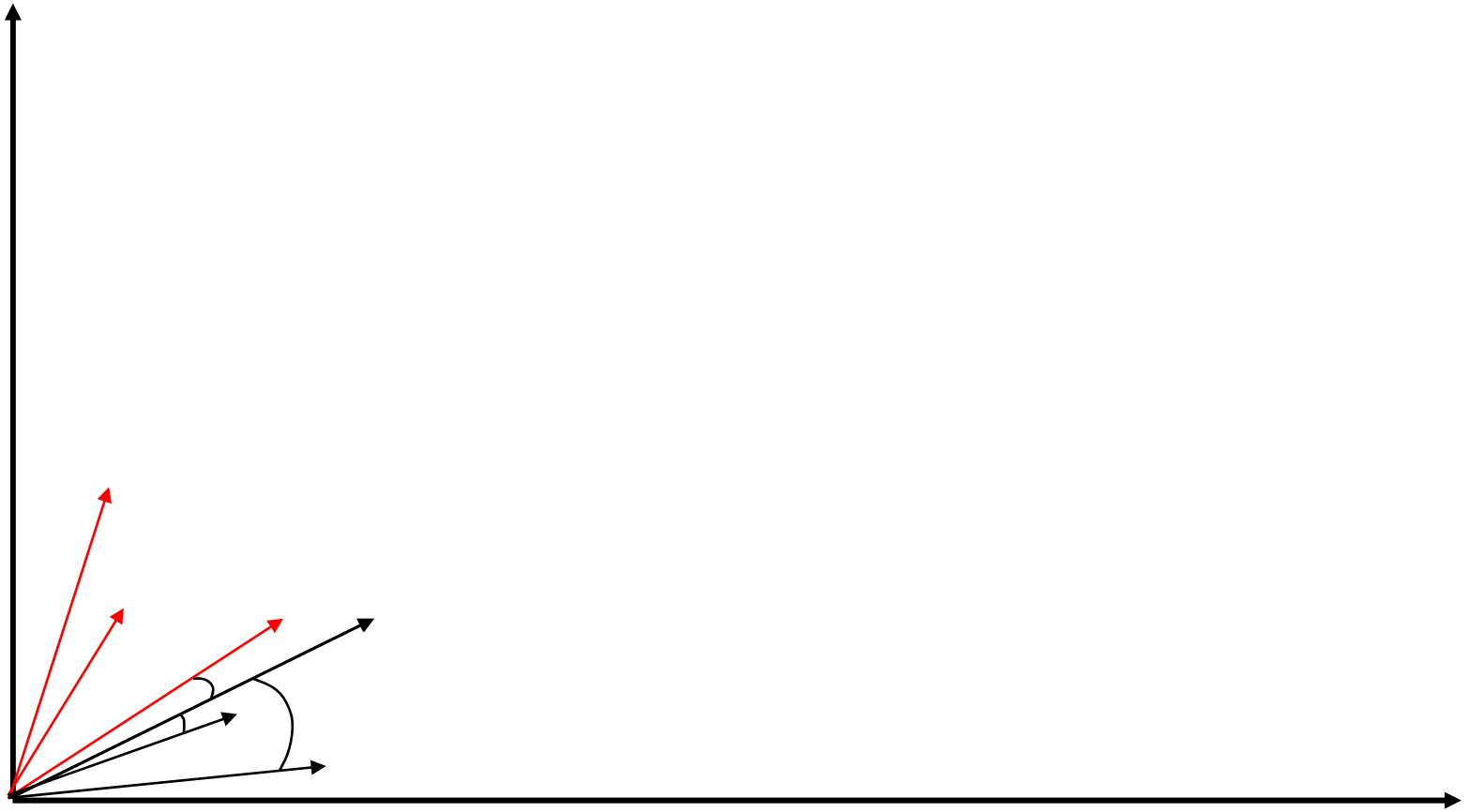
Let $s_x = \text{cosSim}(d, d_x)$

Sort examples, x, in D by decreasing value of s_x

Let N be the first k examples in D. (*get most similar neighbors*)

Return the majority class of examples in N

3近邻方法图示



3近邻方法与Rocchio

- 近邻方法可以解决多态类的问题

贝叶斯方法

- 基于概率理论的学习和分类方法
- 每个类别的先验概率（**prior probability**） 没有任何关于实例的信息
- 分类就是要基于实例的描述生成关于可能类别的后验概率（**posterior probability**）分布

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

贝叶斯分类

- 类别集合： $\{c_1, c_2, \ldots c_n\}$
- E 是实例的描述
- 决定E的类别依赖于对每个c_i每个计算：

$$P(c_i \mid E) = \frac{P(c_i)P(E \mid c_i)}{P(E)}$$

- 由于类别是完全而且不相交（disjoint）的：

$$\sum_{i=1}^{n} P(c_i \mid E) = \sum_{i=1}^{n} \frac{P(c_i)P(E \mid c_i)}{P(E)} = 1$$

$$P(E) = \sum_{i=1}^{n} P(c_i)P(E \mid c_i)$$

贝叶斯分类（2）

- 需要知道：
  - 先验概率： $P(c_i)$
  - 条件概率： $P(E \mid c_i)$
- $P(c_i)$可以从训练数据中估计
  - 如果类型c_i包含D中的 n_i 个实例，则$P(c_i) = n_i / |D|$
- 实例是由一些不相交的特征来描述的：

$$E = e_1 \wedge e_2 \wedge \cdots \wedge e_m$$

Naïve Bayesian Categorization

- 如果假设实例的特征在给定类别c_i下是独立的（条件独立），则

$$P(E \mid c_i) = P(e_1 \wedge e_2 \wedge \cdots \wedge e_m \mid c_i) = \prod_{j=1}^{m} P(e_j \mid c_i)$$

- 因此，对于每个特征和类别，我们只要知道$\mathbf{P(e_j \mid c_i)}$

贝叶斯分类的例子

- C = {allergy, cold, well}
- e_1 = sneeze; e_2 = cough; e_3 = fever

Prob	Well	Cold	Allergy
P(c_i)	0.9	0.05	0.05
P(sneeze|c_i)	0.1	0.9	0.9
P(cough|c_i)	0.1	0.8	0.7
P(fever|c_i)	0.01	0.7	0.4

贝叶斯分类的例子（2）

Probability	Well	Cold	Allergy
P(c_i)	0.9	0.05	0.05
P(sneeze | c_i)	0.1	0.9	0.9
P(cough | c_i)	0.1	0.8	0.7
P(fever | c_i)	0.01	0.7	0.4

E={sneeze, cough, ¬fever}

P(well | E) = (0.9)(0.1)(0.1)(0.99)/P(E)=0.0089/P(E)
P(cold | E) = (0.05)(0.9)(0.8)(0.3)/P(E)=0.01/P(E)
P(allergy | E) = (0.05)(0.9)(0.7)(0.6)/P(E)=0.019/P(E)

最可能的类别： allergy

P(E) = 0.0089 + 0.01 + 0.019 = 0.0379

P(well | E) = 0.23

P(cold | E) = 0.26

P(allergy | E) = 0.50

文本分类中的概率估计

- 在文本分类中，文档被表示为词袋，词汇 $V = \{w_1, w_2, \ldots w_m\}$
- 如果类型c_i包含D中的n_i个实例，这些n_i个实例中有n_{ij}个实例包含词w_j，则：

$$P(w_j \mid c_i) = \frac{n_{ij}}{n_i}$$

- 然而，如果训练集比较小，估计这样概率可能有问题
  - 如果一个罕有词w_k，在训练集中找不到含有该词的文档，即$\forall c_i : P(w_k \mid c_i) = 0$。而某测试样本$E$又含有该词，则有 $\forall c_i$: $P(E \mid c_i) = 0$和$\forall c_i$: $P(c_i \mid E) = 0$，导致无法对E进行分类

平滑（Smoothing）

- 考虑到小样本集的估计问题，通常要采用平滑技术
- **Laplace平滑：假设每个特征有一个先验概率*p*，采用*m*-估计（ *m*-estimate）**

$$P(w_j \mid c_i) = \frac{n_{ij} + mp}{n_i + m}$$

  - $p = 1/|V|$，$m = |V|$

贝叶斯算法（训练）

Let V be the vocabulary of all words in the documents in D
For each category $c_i \in C$
    Let D_i be the subset of documents in D in category c_i
    $P(c_i) = |D_i| / |D|$
    Let T_i be the concatenation of all the documents in D_i
    Let n_i be the total number of word occurrences in T_i
    For each word $w_j \in V$
        Let n_{ij} be the number of occurrences of w_j in T_i
        Let $P(w_i \mid c_i) = (n_{ij} + 1) / (n_i + |V|)$

贝叶斯算法（判决）

Given a test document *X*
Let *n* be the number of word occurrences in *X*
Return the category:

$$\underset{c_i \in C}{\operatorname{argmax}} P(c_i) \prod_{j=1}^{n} P(a_j \mid c_i)$$

where a_j is the word occurring the *j*th position in *X*

贝叶斯文本分类的例子

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Japan Chinese Chinese Chinese Tokyo	?

- **c(5)=?**

贝叶斯文本分类的例子（2）

- 特征似然估计（**Feature likelihood estimate**）

$$\begin{aligned}
\hat{P}(\text{Chinese}|c) &= (5+1)/(8+6) = 6/14 = 3/7 \\
\hat{P}(\text{Tokyo}|c) = \hat{P}(\text{Japan}|c) &= (0+1)/(8+6) = 1/14 \\
\hat{P}(\text{Chinese}|\bar{c}) &= (1+1)/(3+6) = 2/9 \\
\hat{P}(\text{Tokyo}|\bar{c}) = \hat{P}(\text{Japan}|\bar{c}) &= (1+1)/(3+6) = 2/9
\end{aligned}$$

- 后验概率（**Posterior**）

$$\begin{aligned}
\hat{P}(c|d) &\propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003 \\
\hat{P}(\bar{c}|d) &\propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001
\end{aligned}$$

- 分类结果： **c(5) = China**

主要内容

- 什么是分类和聚类?
- 文本分类
  - 文本分类的定义
  - 文本分类的方法
  - 文本分类的评估
- 文本聚类

经典的数据集：Reuters Data Set

- 最常用分类测试数据集，包含21578个文档
- 其中9603个训练（training）文档，3299个测试（test）文档
- 118个类别
  - 一个文档可以属于多个类别
  - 每个文档平均1.24个类别，至少一个
- 118个类别中只有10个类别是大类
- 常用类别 (#train, #test）

- Earn (2877, 1087)
- Acquisitions (1650, 179)
- Money-fx (538, 179)
- Grain (433, 149)
- Crude (389, 189)
- Trade (369,119)
- Interest (347, 131)
- Ship (197, 89)
- Wheat (212, 71)
- Corn (182, 56)

Reuters-21578的文档

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">

<DATE> 2-MAR-1987 16:51:43.42</DATE>

<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off

tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

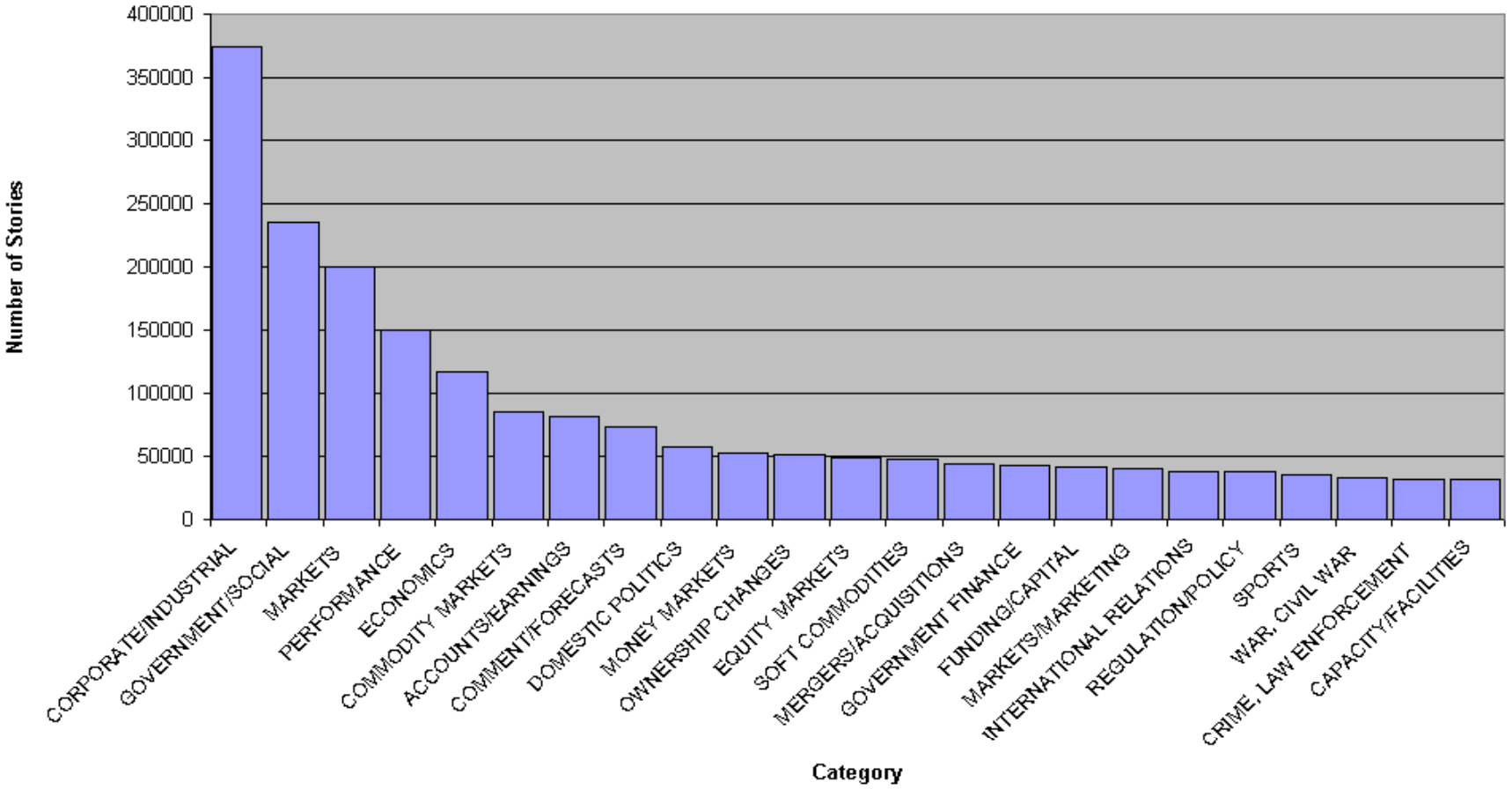
Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

&#3;</BODY></TEXT></REUTERS>

新的Reuters：RCV1

- **Reuters RCV1 （810,000个文档）中的主要类别**



北大天网：中文网页分类

- 通过动员不同专业的几十个学生，人工选取形成了一个基于层次模型的大规模中文网页样本集。
- 包括**12,336**个训练网页实例和**3,269**个测试网页实例，分布在12个大类，共计**733**个类别中，每个类别平均有**17**个训练实例和**4.5**个测试实例
- 中文信息检索论坛**www.cwirf.org**
- 全国搜索引擎和网上信息挖掘学术研讨会**SEWM**上进行分类评测

北大天网：中文网页分类

类别号	类别名称	类别数	训练样本数	测试样本数
1	人文与艺术	24	419	110
2	新闻与媒体	7	125	19
3	商业与经济	48	839	214
4	娱乐与休闲	88	1510	374
5	计算机与因特网	58	925	238
6	教育	18	286	85
7	各国风情	53	891	235
8	自然科学	113	1892	514
9	政府与政治	18	288	84
10	社会科学	104	1765	479
11	医疗与健康	136	2295	616
12	社会与文化	66	1101	301
共计		733	12336	3269

评测指标

- 分类的准确度（accuracy）
  - 学术界的主要评测标准
- 分类器的训练速度
  - 一些方法很快，一些方法开销很大
- 分类的速度（docs/hour）
  - 对大多数算法来说都很快
  - 除了kNN，需要复杂的预处理
- 在特征提取和分类器构建方面的代价
  - 每个主题需要多少人工时间（human hours/topic）

单个类别的评测指标

Actual Class

Predicted class	A	B	C
A			
B			
C			

- 查全率（Recall）： 类 *i* 中被正确分类的文档的比例
- 查准率（Precision）： 被划分为类*i*的文档中正确属于类别*i*的文档比例
- F_1 值： $1/F_1 = ½ (1/P + 1/R)$

评估方法

- **Train + Test**
  - 训练集合和测试集合是不同的集合
  - 用训练集训练，用测试集合测试
- **N次交叉检验（ N fold cross-validation）**
  - 将所有标注文档集合分成N份
  - 用其中N-1训练，其他一份测试
  - 测试N次，求平均

如何合并多次的测度？

- 如果有多个类，如何合并多个测度？

- 宏平均（**Macroaveraging**）
  - 对每个类求值，然后平均。大类小类同等看待
- 微平均（**Microaveraging**）
  - 将所有文档一块儿计算，求值。易受大类的影响

宏平均和微平均的例子

Class 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro.Av. Table

	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

- 宏平均**precision:** **(0.5 + 0.9)/2 = 0.7**
- 微平均**precision:** **100/120 = .83**
- 为什么不一样?

主要内容

- 什么是分类和聚类?
- 文本分类
- 文本聚类
  - 文本聚类的定义
  - 文本聚类的方法
  - 文本聚类的评估

文本聚类定义

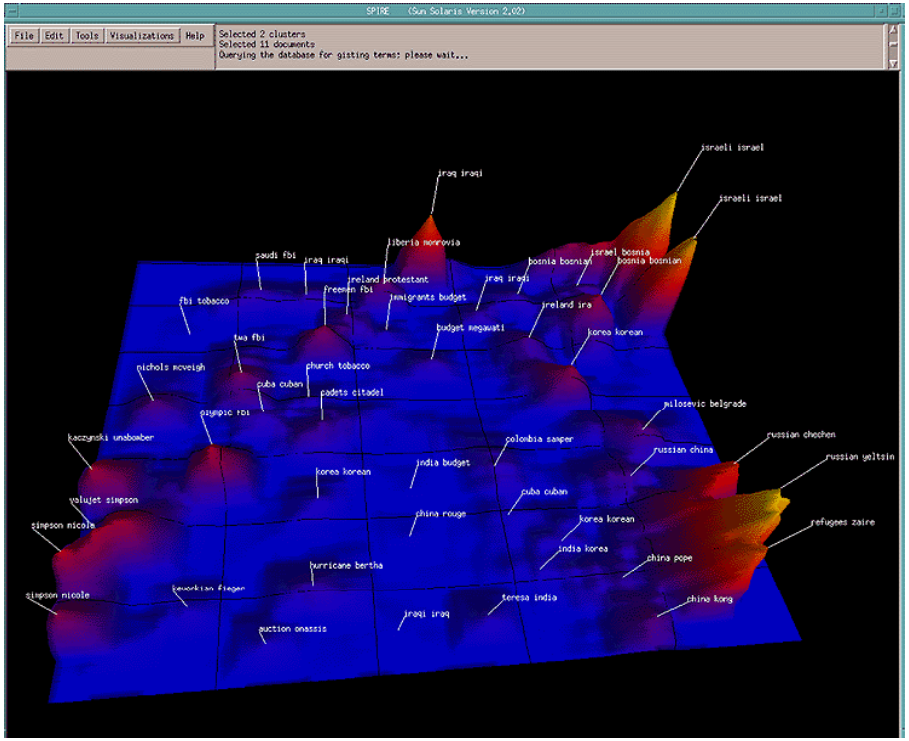
- 聚类是一个无督导的学习过程，它是指根据样本之间的某种距离在无监督（ **unsupervised** ）条件下的聚簇过程
- 利用聚类方法可以把大量的文档划分成用户可迅速理解的簇（**cluster**），从而使用户可以更快地把握大量文档中所包含的内容，加快分析速度并辅助决策
- 大规模文档聚类是解决海量文本中数据理解和信息挖掘的有效解决手段之一

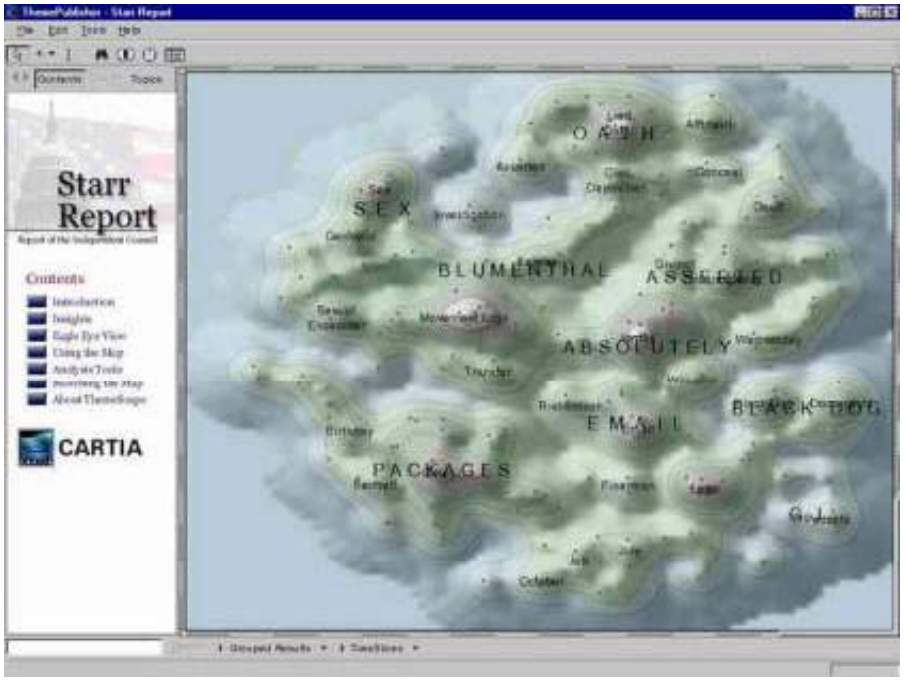
文本聚类的应用

- **TDT（TopicDetection and Tracking）中主题事件的检测**
  - 将文档进行聚类，从聚出的类中发现新的热点主题
- 检索结果的聚类显示
  - 检索结果聚类，以便用户浏览
- 大规模文档的组织和呈现
  - 更好的用户界面
- 加速检索
- 更好的检索结果
  - 提高查全率

可视化文档集及其主题

- **ThemeScapes, Cartia**
  - 山的高度=簇的大小





Wise et al, "Visualizing the non-visual" PNNL

图像聚类

Image Clustering (ICME'04)

1710 JPG images in 1287 pages are crawled within the website
http://www.yahooligans.com/content/animals/

Six Categories

Microsoft Research Asia

Mammal

Fish

Reptile

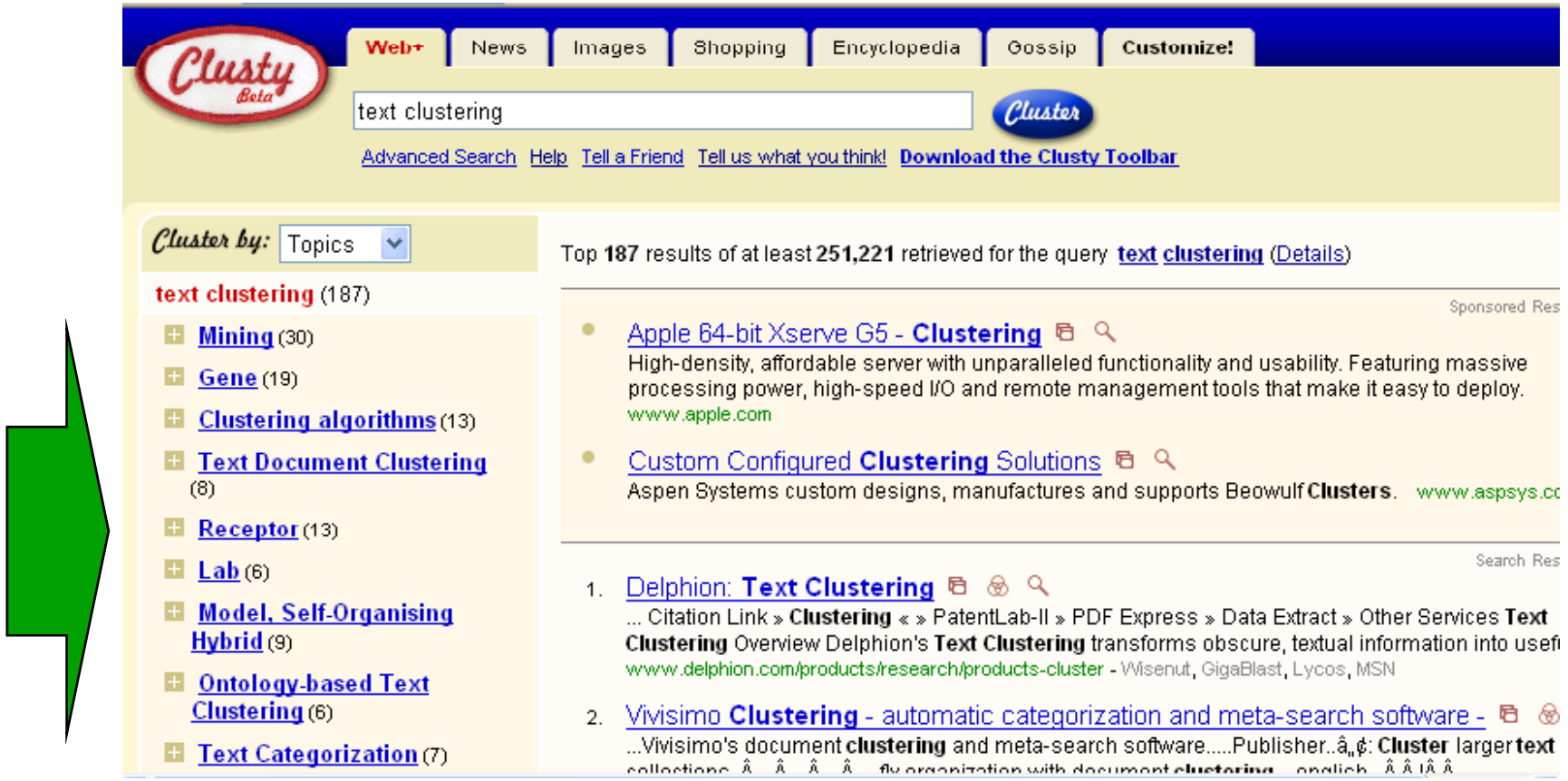
Bird

Amphibian

Insect

检索结果聚类

- 按主题聚合检索结果
  - **clusty.com / Vivisimo**



Vivisimo

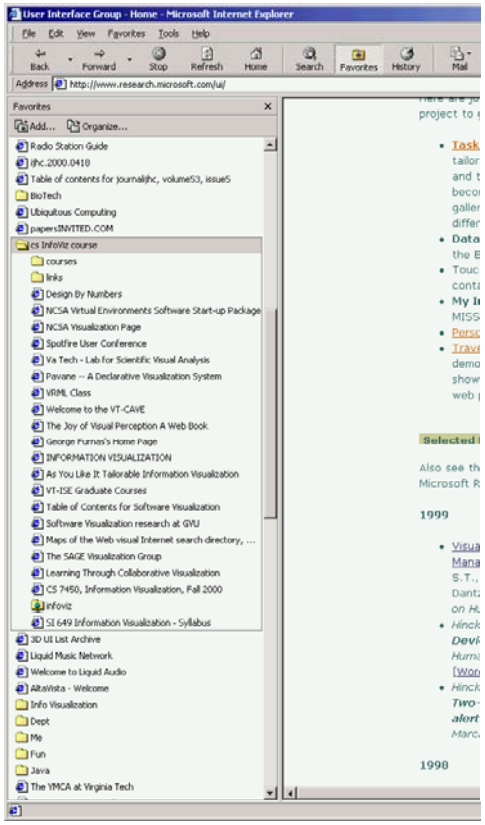

Vivisimo - Clustered search on 王刚 - Microsoft Internet Explorer
company | products | solutions | customers | demos | press
王刚
the Web
Search
Advanced Search
Help
NEW try your query at Clusty.com
Clustered Results
王刚 (131)
张国立 (9)
英格力士 (5)
Sina.com (5)
王刚,男,汉族,1942年10月生,吉林扶余人 (5)
王刚个人档案,最佳男配角 (3)
宰相刘罗锅,铁齿铜牙纪晓岚 (3)
康熙来了,还有张铁林和王刚的加盟,而张国立则缺阵"铁三角"组合,由孙兴来顶替他的位置 (3)
独臂刀,武侠天地 - 司马紫烟作品 (3)
Spaces.msn (2)
人物信息 (2)
人们都会将他与奸猾,尽管王刚试图改变自己的银幕形象,曾扮演过其他角色,但让观众记忆犹新的还是那个狡猾的王刚 (3)
醒狮 第一章 (3)
版权所有 (3)
Top 131 results of at least 106,958 retrieved for the query 王刚 (Details)
1. 王刚 影音娱乐 新浪网 [new window] [frame] [cache] [preview] [clusters]
... 获北京电视艺术春燕奖 获第十四届中国电视金鹰奖"最佳男配角"。 >> 王刚作客新浪访谈实录(图) >> 王刚个人档案 王刚张国立犯下"欺君"之罪 "皇阿玛 ...
ent.sina.com.cn/s/m/f/wangg.html - MSN 1, MSN Search 1
2. 王刚活动报道集 [new window] [frame] [preview] [clusters]
王刚活动报道集 人民网版权所有 未经书面授权禁止使用 Copyright © 2002 by www.people.com.cn. all rights reserved ... 人民网 >> 时政 >> 时政专题 >> 王刚活动报道集 王刚同志简历 王刚活动报道集 国内活动 出访 ...
www.people.com.cn/GB/shizheng/252/9933 - MSN 2
3. 王刚任播音主持委员会副会长 狠批《康熙来了》 影 ... [new window] [frame] [cache] [preview] [clusters]
... 王刚任播音主持委员会副会长 狠批《康熙来了》 http://ent.sina.com.cn 2005年06月25日10:09 北京青年报 会议现场 (图片来源：新浪娱乐) 点击此处查看全部娱乐 ...
ent.sina.com.cn/s/m/2005-06-25/1009762313.html - MSN Search 2
4. 作家王刚作客新浪聊《英格力士》实录 读书频道 新浪网 [new window] [frame] [preview] [clusters]
作家王刚作客新浪聊《英格力士》实录 王刚虽然是实力派作家,但直到《英格力士》才引起众媒体的聚焦
book.sina.com.cn/author/subject/2005-01-05/3/148574.shtml - MSN 3
5. 张铁林：王刚张国立欺负我 "卡通皇帝"一切以 ... [new window] [frame] [cache] [preview] [clusters]
... 张铁林：王刚张国立欺负我 "卡通皇帝"一切以授课为重 05年10月27日 很多观众最初都是通过《大桥下面》这部片子认识张铁林的，不久他就去了 ...
www.chinapressusa.com/yule/200510270195.htm - MSN Search 3
6. 交友档案 - 友缘人 - 雅虎中国 [new window] [frame] [preview] [clusters]
Internet
开始
备课笔记_文...
第8章 文本聚类
Adobe Acrob...
Microsoft P...
Vivisimo - ...
12:49

检索结果的浏览


http://kartoo.com/flash04.php3
KartOO
Web
Images
Videos
Wikipedia
Bill Gates melinda
Search
help
45 700 000 Found results 1 - 20
Topics :
bill melinda gates founda
free encyclopedia
chairman of microsoft
microsoft chairman bill
billion dollar
information about bill ga
interview with bill gates
melinda
founder
wikiquote
time
microsoft
chairman
corporate
biography
information
archival
Add to your search
remove from your search
Erase this map
Modify the links
Statistics (4 weeks)
Rename
www.zpub.com
richest
ei.cs.vt.edu
life
information
news
www.techcrunch.com
www.billgatesisdead.com
microsoft
head
topics.nytimes.com
www.forbes.com
chairman
www.microsoft.com
en.wikiquote.org
founder
time
persian
melinda
corporate
biography
www.answers.com
www.gatesfoundation.org
www.time.com
en.wikipedia.org
next map

可视化书签（bookmarked pages）

- **Robertson, “Data Mountain” (Microsoft)**



提高检索效果

- 具有相似内容的文档被聚合在一起
- 当查询匹配某个文档时，同时返回在同一聚类中的其他文档
  - 可以提高查全率，例如查询 **“car”**可能可以返回包含***automobile***的文档
  - 可以加快检索速度，因为只在聚类文档中查找，可能不是很精确，但避免了大量相似计算

主要内容

- 什么是分类和聚类?
- 文本分类
- 文本聚类
  - 文本聚类的定义
  - 文本聚类的方法
  - 文本聚类的评估

聚类算法

- 层次方法（Hierarchical Methods）
  - 凝聚算法（Agglomerative Algorithms）：自底向上（Bottom-up）
  - 分裂算法（Divisive Algorithms）：至上而下（Top-down）
- 划分方法（Partitioning Methods）
  - K-中心点算法（K-medoidsMethods）
  - K-平均算法（K-means Methods）

划分方法

- 将***n***个文档分到***K***个簇
- 给定：
  - 一系列文档和数字***K***
- 找到：
  - ***K*** 个簇，优化划分准则
    - 全局优化（**Globally optimal**）： 枚举所有的分区（**exhaustively enumerate all partitions**）
    - 启发式规则（**Effective heuristic methods**）： ***K*-means** 或***K*-medoids**算法

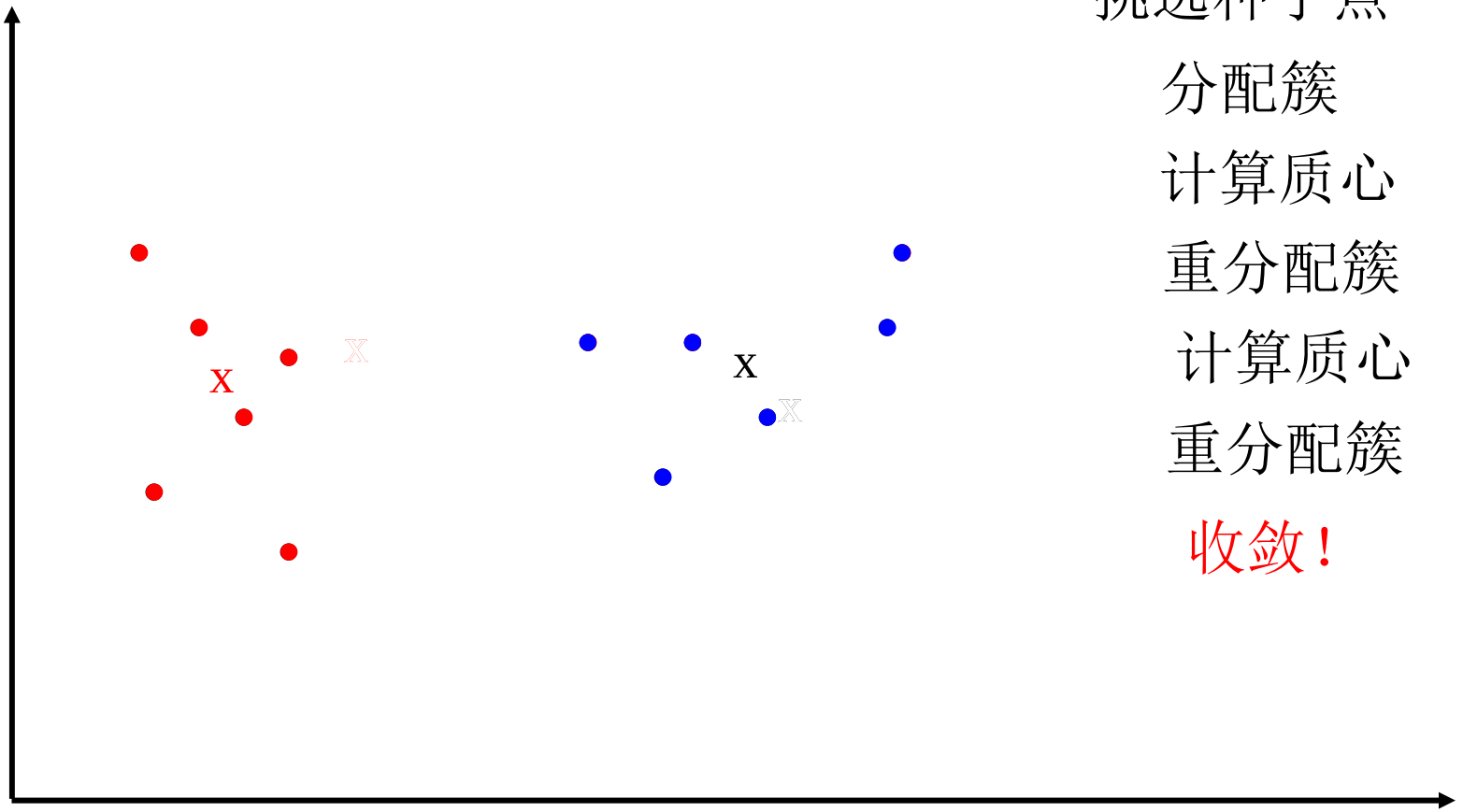
K-Means算法原理

- 文档被表示为实值向量
- 计算一个簇 ω的质心（ centroids ）：

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

- 视实例与当前簇的质心的距离来分配文档到各个簇

K Means聚类的例子 (*K*=2)



挑选种子点

分配簇

计算质心

重分配簇

计算质心

重分配簇

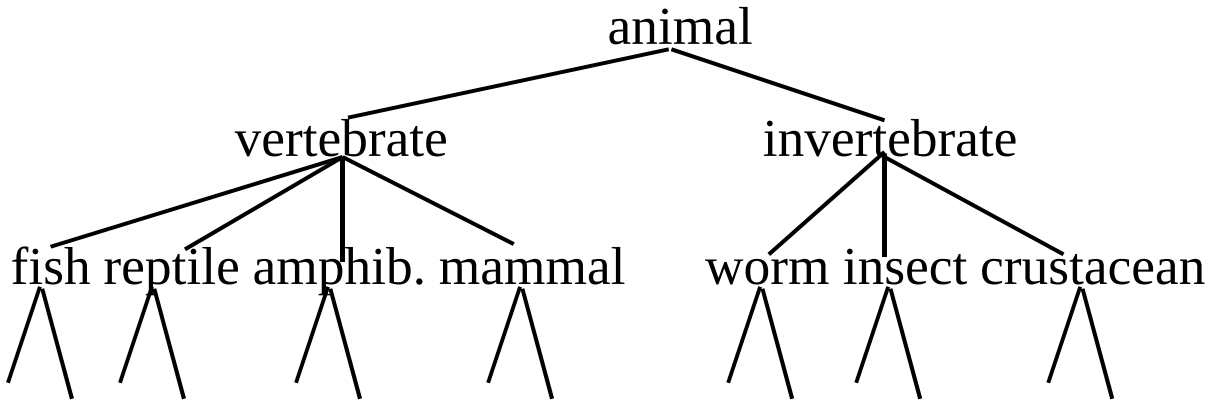
收敛！

K Means算法流程

1. **Step1**：初始化***k***个簇中心；
2. **Step2**：对于每个文档向量，计算该文档向量与**k**个类中心的距离，选择距离最小（相似度最大）的簇将该文档分入该簇；
3. **Step3**：重新计算***k***个簇的中心，中心为该簇内所有点的算术平均。
4. **Step4**：如果簇变化不大或者满足某种退出条件（达到最大迭代次数、满足某种目标函数等），那么结束聚类，否则返回**Step2**。

层次聚类（HAC）

- 根据未标注的实例，构建一个层次的分类（**taxonomy**）树

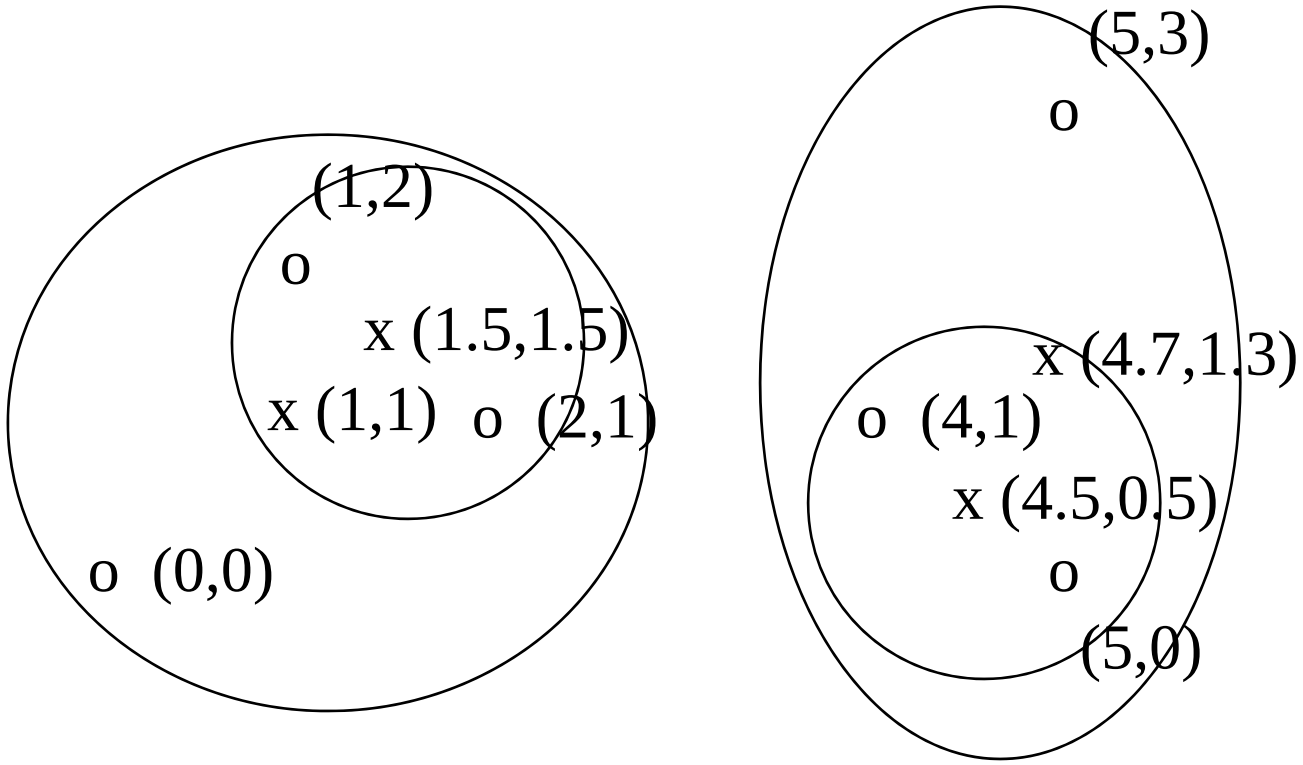


- 生成层次聚类的方法之一是递归地采用划分方法

HAC算法流程

1. **Step1:** 将所有的点各自单独形成一个簇;
2. **Step2:** 从现有所有的簇中选择最近（或者最相似的）两个簇，进行合并;
3. **Step3:** 如果只剩下一个簇或者达到终止条件（比如达到需要的簇的数目），聚类结束,否则返回**Step2.**

HAC的例子



最近簇的定义

- 质心（**center of gravity**）
  - 质心最近（ **cosine**相似）的簇
- 群平均（**average-link**）
  - 任何两对元素之间距离的平均值最小的簇
- 单连接（**single-link**）
  - 距离最近的两个元素的相似度最小的簇
- 全连接（**complete-link**）
  - 距离最远的两个元素的相似度最小的簇

基于单连接的凝聚聚类

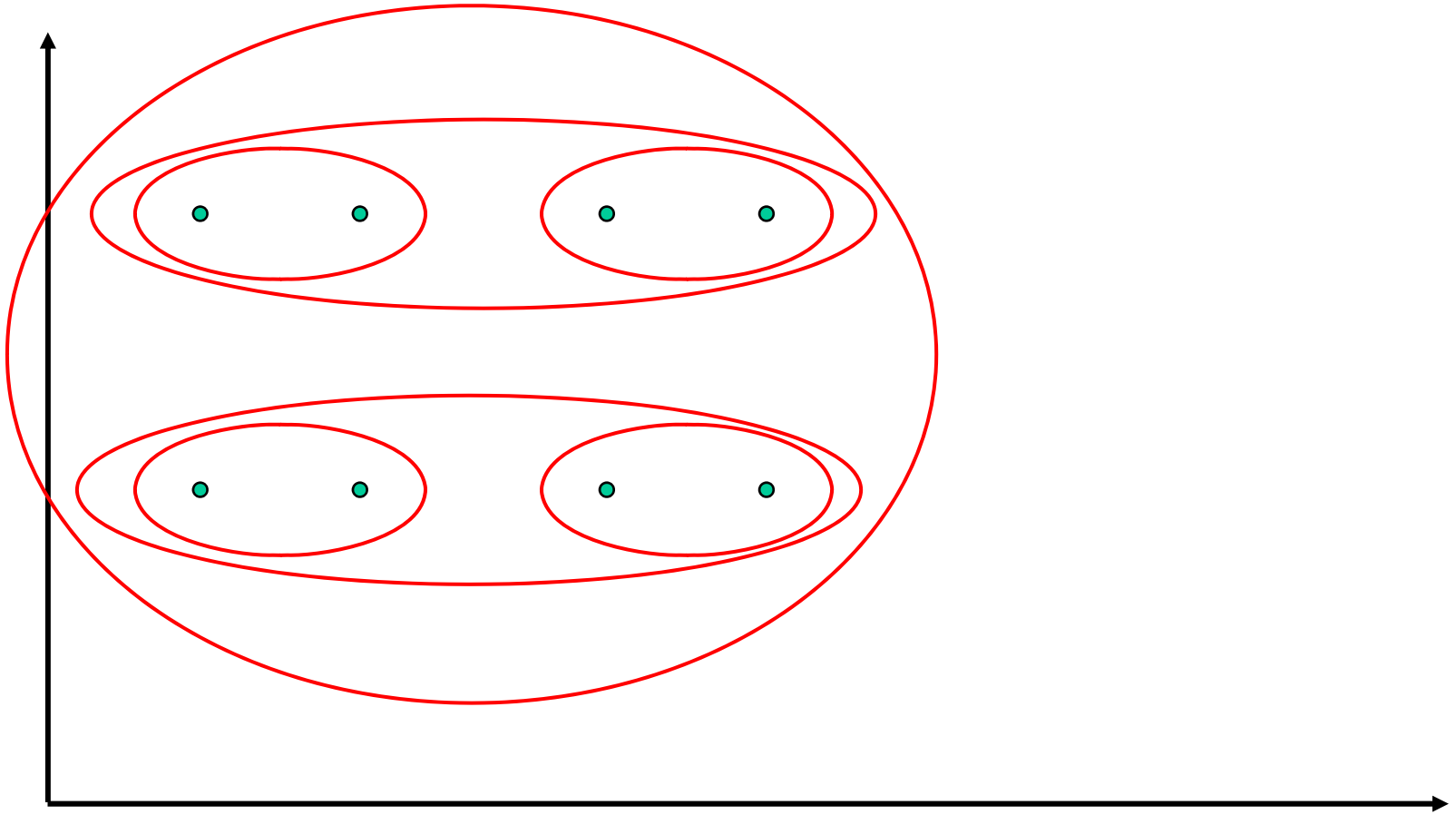
- 寻找具有最大距离的元素对，以它们的相似度作为簇之间的相似度：

$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$

- 把具有最小相似度的簇聚合起来，
  - 可能产生细而长的链，形成聚类岛“**Hawai'i clusters**”
- 在合并簇c_i和c_j后，结果簇和其他簇c_k的相似度可以表示为：

$$sim((c_i \cup c_j), c_k) = \max(sim(c_i, c_k), sim(c_j, c_k))$$

单连接的例子



基于全连接的凝聚聚类

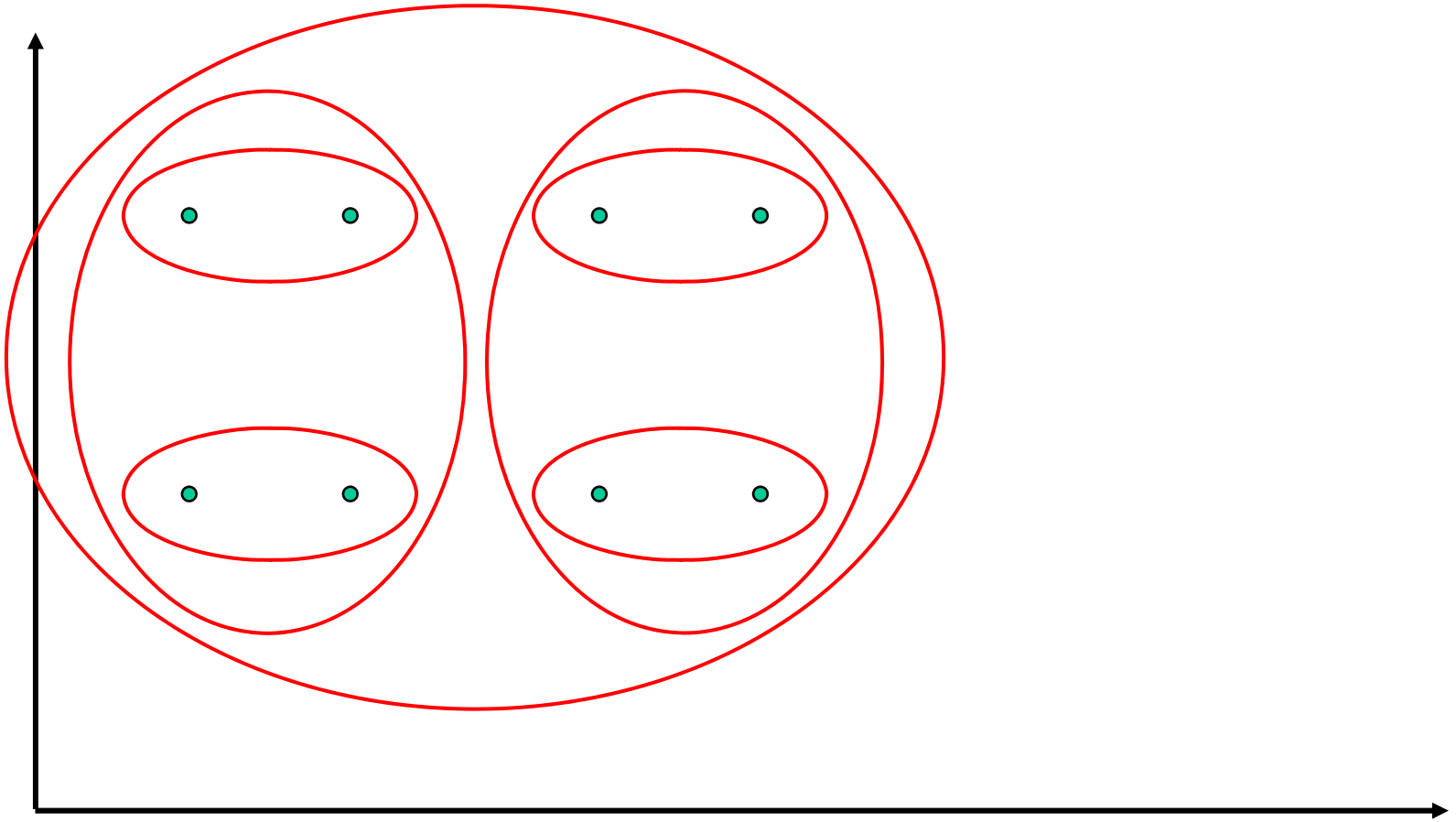
- 用元素对的最小相似度做为簇间的相似度：

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$

- 会产生比较紧密的，球形的簇
- 在合并簇c_i和c_j后，结果簇和其他簇c_k的相似度可以表示为：

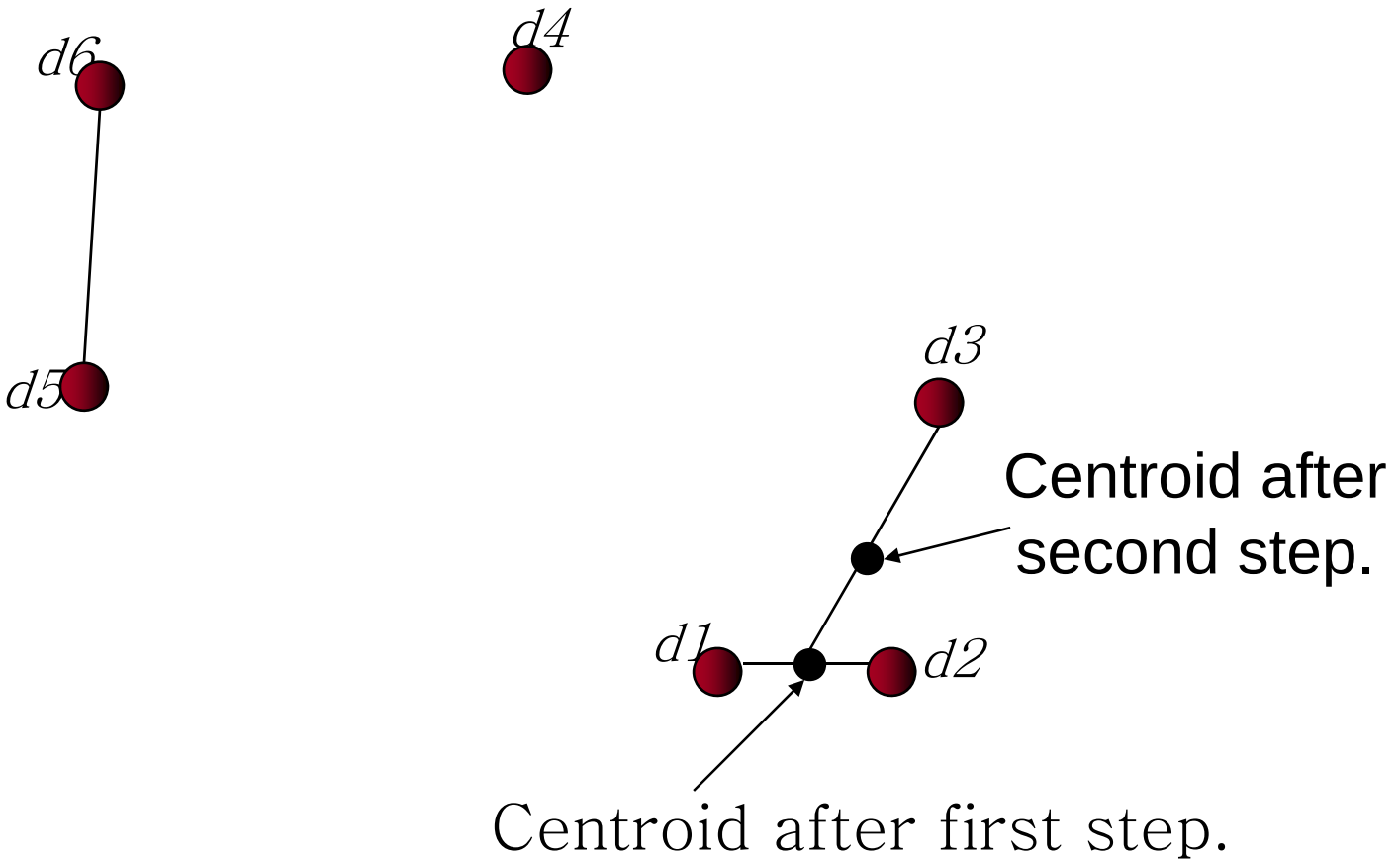
$$sim((c_i \cup c_j), c_k) = \min(sim(c_i, c_k), sim(c_j, c_k))$$

全连接的例子



基于质心的凝聚聚类

例子： ***n=6, k=3,*** **closest pair of centroids**



基于群平均的凝聚聚类

- 首先对两个簇进行并运算，然后计算这个集合中任意两个文档相似度和，再取平均

$$sim(c_i, c_j) = \frac{1}{\left|c_i \cup c_j\right|\left(\left|c_i \cup c_j\right| - 1\right)} \sum_{\vec{x} \in (c_i \cup c_j)} \sum_{\vec{y} \in (c_i \cup c_j): \vec{y} \neq \vec{x}} sim(\vec{x}, \vec{y})$$

- 是在单个链和完全链之间的折中

主要内容

- 什么是分类和聚类?
- 文本分类
- 文本聚类
  - 文本聚类的定义
  - 文本聚类的方法
  - 文本聚类的评估

什么是好的聚类？

- 好的聚类应产生高质量的簇：
  - 簇内（**intra-cluster**）的相似度高
  - 簇间（**inter-cluster**）的相似度低
- 簇的质量测度依赖于文档表示和所用的相似度测度

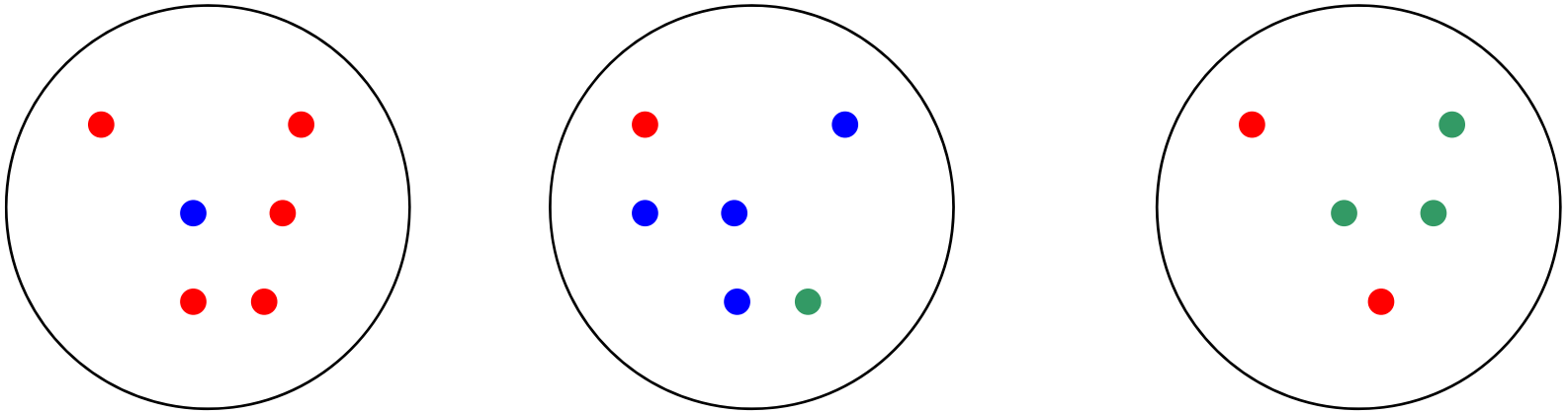
簇的评测指标

- 聚类评测即评测其发现隐含模式或潜在类别的能力
- 假设有J个已经标注好的类（gold standard classes）$\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$，聚类算法产生$K$个簇, $\omega_1, \omega_2, \ldots, \omega_K$
- 单一的 评测指标：纯度（purity）

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

  - N为文档个数

Purity的计算例子



Cluster I: Purity = 1/6 *(max(5, 1, 0)) = 5/6

Cluster II: Purity = 1/6 * (max(1, 4, 1)) = 4/6

Cluster III: Purity = 1/5 * (max(2, 0, 3)) = 3/5

Total: Purity =	1/17 * (5+4+3) = 12/17

Rand Index

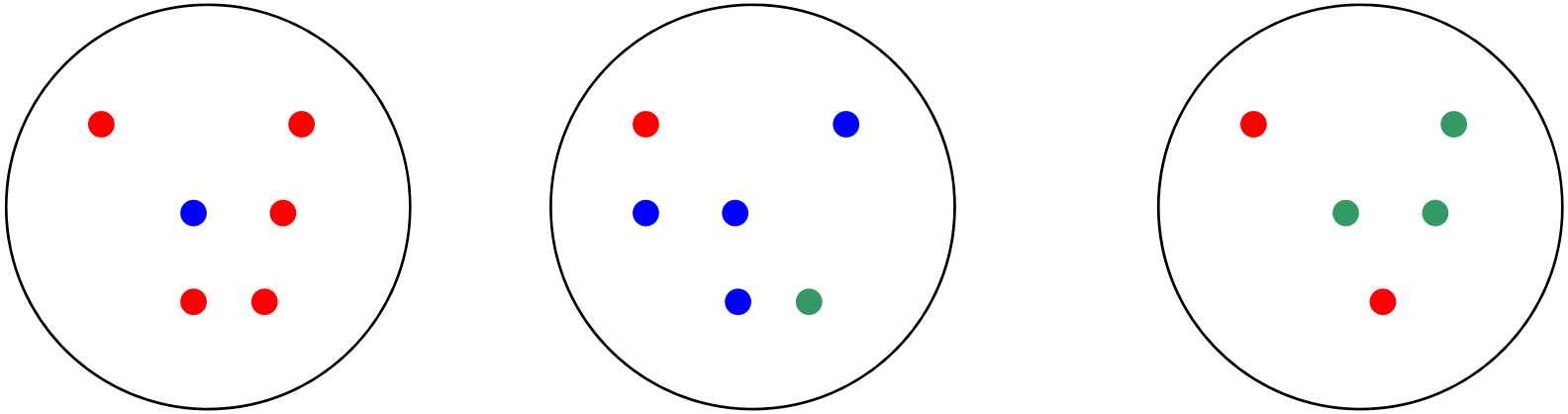
- 把聚类看成是一系列决定的过程，即对集合中 $N(N-1)/2$ 对文档做决定
- **TP**（**true positive**）： 将两个相似的文档放到同一个簇的决定
- **TN**（**true negative** ）： 将两个不相似的文档放到不同簇的决定
- **FP**（**false positive**）： 将两个不相似的文档放到同一簇的决定
- **FN**（**false negative**）： 将两个相似的文档放到不同簇的决定

Rand Index的定义

Number of points	Same Cluster in clustering	Different Clusters in clustering
Same class in ground truth	TP	FN
Different classes in ground truth	FP	TN

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

Rand index的计算例子



Cluster I　　Cluster II　　Cluster III

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

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

	Same cluster	Different clusters
Same class	$TP = 20$	$FN = 24$
Different classes	$FP = 20$	$TN = 72$

本讲小结

- 文本分类的定义
- 算法
  - **Rocchio**
  - **K-Nearest Neighbor**
  - **Naïve Bayesian**
- 分类评估
  - **F1值**
  - 宏平均和微平均
  - **N次交叉校验**

- 文本聚类的定义
- 算法
  - **K-Means**
  - **HAC**
    - 簇的最近测度
- 聚类评估
  - **Purity, Rand Index**

推荐阅读和网站

- 《网络信息检索》第十一章
- **Introduction to Information retrieval**
  - **Ch 13: Text classification and Naïve Bayes**
  - **Ch 14: Vector space classification**
  - **Ch 15: Support Vector Machine and machine learning on documents**
  - **Ch 16: Flat clustering**
  - **Ch 17: Hierarchical clustering**

- **Y. Yang and X. Liu, "A re-examination of text categorization methods," SIGIR, 1999**
- **B. Florian, E. Martin, and X. Xiaowei, "Frequent term-based text clustering," ACM SIGKDD, 2002**