

# 信息检索与数据挖掘

---

## 第7章 相关反馈和查询扩展

# 课程内容

- 第1章 绪论
- 第2章 布尔检索及倒排索引
- 第3章 词典查找及扩展的倒排索引
- 第4章 索引构建和索引压缩
- 第5章 向量模型及检索系统
- 第6章 检索的评价
- 第7章 相关反馈和查询扩展
- 第8章 概率模型
- 第9章 基于语言建模的检索模型
- 第10章 文本分类
- 第11章 文本聚类
- 第12章 Web搜索
- 第13章 多媒体信息检索
- 第14章 其他应用简介

# 提纲

- ① 上一讲回顾
- ② 动机
- ③ 相关反馈基础
- ④ 相关反馈详细介绍
- ⑤ 查询扩展

# 提纲

- ① 上一讲回顾
- ② 动机
- ③ 相关反馈基础
- ④ 相关反馈详细介绍
- ⑤ 查询扩展

# 上一讲回顾

- 信息检索的评价方法
  - 不考虑序的评价方法(即基于集合): P、R、F
  - 考虑序的评价方法: P/R曲线、MAP、NDCG
- 相关评测
- 检索结果的摘要

# 正确率(Precision)和召回率(Recall)

- 正确率(Precision, 简写为 $P$ ) 是返回文档中真正相关的比率

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

- 召回率(Recall,  $R$ ) 是返回结果中的相关文档占所有相关文档(包含返回的相关文档和未返回的相关文档)的比率

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

# 正确率 vs. 召回率

	相关(relevant)	不相关(nonrelevant)
返回(retrieved)	真正例(true positives, <i>tp</i> )	伪正例(false positives, <i>fp</i> )
未返回(not retrieved)	伪反例(false negatives, <i>fn</i> )	真反例(true negatives, <i>tn</i> )

$$P = TP / ( TP + FP )$$

$$R = TP / ( TP + FN )$$

# 正确率和召回率相结合的指标：F值

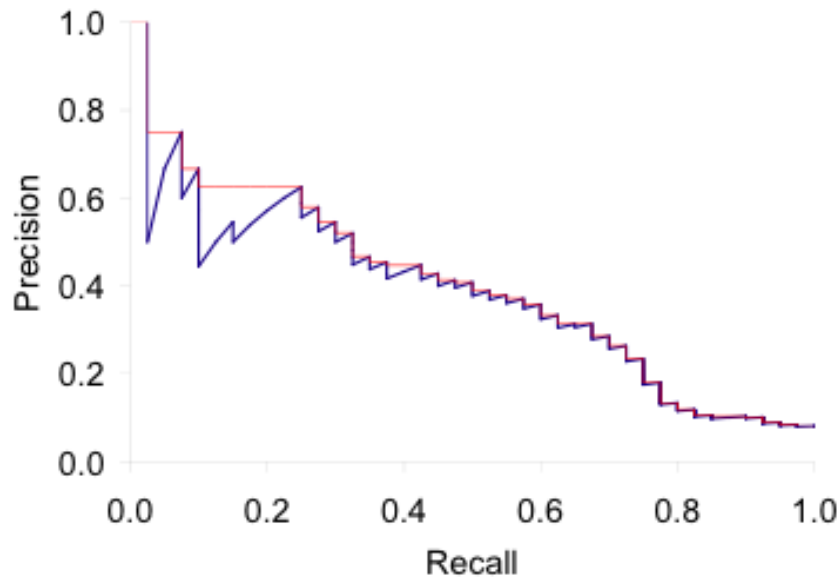
- $F$  允许正确率和召回率的折中

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \text{ where } \beta^2 = \frac{1 - \alpha}{\alpha}$$

- $\alpha \in [0, 1]$  ,  $\beta^2 \in [0, \infty]$
- 常用参数: **balanced  $F$**  ,  $\beta = 1$  or  $\alpha = 0.5$ 
  - 实际上是正确率和召回率的调和平均数 (**harmonic mean**)  
 $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$

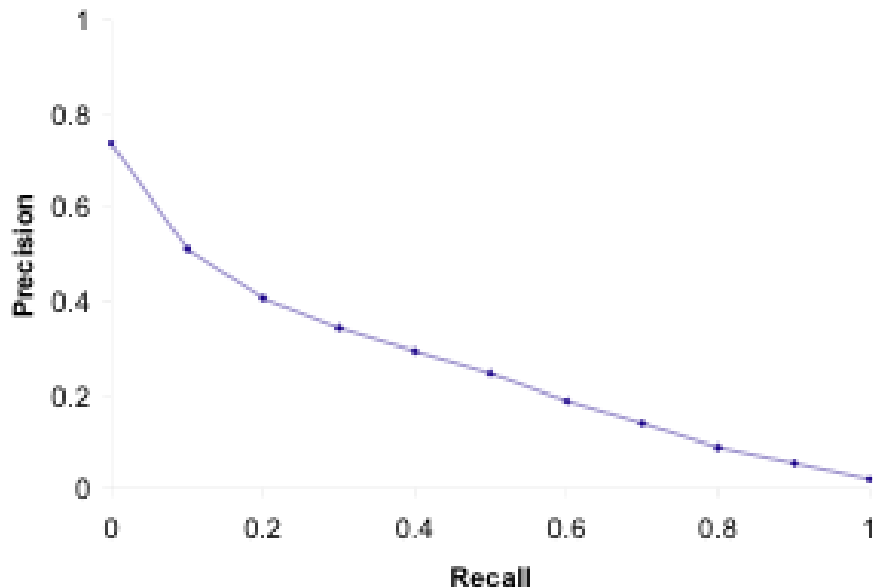


# 正确率-召回率曲线



- 每个点对应top k上的结果 ( $k = 1, 2, 3, 4, \dots$ ).
- 插值 (红色): 将来所有点上的最高结果
- 插值的原理: 如果正确率和召回率都升高, 那么用户可能愿意浏览更多的结果

# 平均的 11-点正确率/召回率曲线



- 计算每个召回率点(0.0, 0.1, 0.2, . . .)上的插值正确率
- 对每个查询都计算一遍
- 在查询上求平均
- 该曲线也是 T R E C 评测上常用的指标之一

# MAP

- 平均正确率 (Average Precision, AP)：对不同召回率点上的正确率进行平均
  - 未插值的AP：某个查询Q共有6个相关结果，某系统排序返回了5篇相关文档，其位置分别是第1，第2，第5，第10，第20位，则
$$AP = (1/1 + 2/2 + 3/5 + 4/10 + 5/20 + 0) / 6$$
- 多个查询的AP的平均值称为系统的MAP (Mean AP)
- MAP是IR领域使用最广泛的指标之一

# R正确率

- Precision@k

- 前k个结果的查准率

- R-Precision

- 检索结果中，在所有相关文档总数位置上的准确率。如某个查询的相关文档总数为 $Re1$ ，返回的结果中前 $Re1$ 个中 $r$ 个是相关文档，则R正确率是 $r/Re1$ 。
  - R正确率能够适应不同的相关文档集的大小
    - 例： $Re1=8$ ； $r=8$ 。此时R正确率是1，但是 $P@20=0.4$
  - 一个完美的系统的R-precision=1

# GMAP

- GMAP (GeometricMAP): TREC2004 Robust 任务引进
- 先看一个例子

系统	Topic	AP	Increase	MAP
系统A	Topic 1	0.02	-	0.113
	Topic 2	0.03	-	
	Topic 3	0.29	-	
系统B	Topic 1	0.08	+300%	0.107
	Topic 2	0.04	+33.3%	
	Topic 3	0.20	-31%	

- 从MAP来看，系统A好于系统B，但是从每个查询来看，3个查询中有2个 Topic B比A有提高，其中一个提高的幅度达到300%

- 几何平均值 
$$GMAP = \sqrt[n]{\prod_{i=1}^n AP_i} = \exp\left(\frac{1}{n} \sum_{i=1}^n \ln AP_i\right)$$

- 上面那个例子  $GMAP_a = 0.056$ ,  $GMAP_b = 0.086$   $GMAP_a < GMAP_b$
- GMAP和MAP各有利弊，可以配合使用，如果存在难Topic时，GMAP更能体现细微差别

# NDCG

- 每个文档不仅仅只有相关和不相关两种情况，而是有相关度级别，比如0, 1, 2, 3。

我们可以假设，对于返回结果：

- 相关度级别越高的结果越多越好
- 相关度级别越高的结果越靠前越好

$$\text{NDCG}(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^k \frac{2^{R(j,m)} - 1}{\log_2(1 + m)}$$

- $R(j, d)$  是评价人员给出的文档 $d$ 对查询 $j$ 的相关性得分， $Z_{j,k}$ 是归一化因子，保证对完美系统NDCG的值为1， $m$ 是返回文档的位置

# 从文档集合如何构建测试集

- 需要
  - 用于测试的查询
  - 相关性的判定
- 用于测试的查询
  - 必须和测试文档集合有密切关系
  - 最好由领域的专家设计
  - 随机的查询并不好
- 相关性的判定
  - 人工判定耗时较长
  - 使用一组人进行判定是否是最好的方式？

# 相关性判定之间的一致性

- Kappa统计量

- 衡量不同人意见的一致性
- 对随机的一致性的简单校正

- $$\text{Kappa} = [ P(A) - P(E) ] / [ 1 - P(E) ]$$

- $P(A)$  - 实际观察到的一致性判断比率

- $P(E)$  - 随机情况下所期望的一致性判断的比率

- $\text{Kappa} = 0$  和随机判断的情况一样, 1 完全一致.

- $k$ 在  $[2/3, 1.0]$ 时, 判定结果是可以接受的

- 如果 $k$ 值比较小, 那么需要对判定方法进行重新设计



# 计算kappa统计量

		Judge 2 Relevance			
		Yes	No	Total	
Judge 1 Relevance	Yes	300	20	320	Observed proportion of the times the judges agreed
	No	10	70	80	
	Total	310	90	400	

$$P(A) = (300 + 70)/400 = 370/400 = 0.925$$

Pooled marginals

$$P(\text{nonrelevant}) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$$

$$P(\text{relevant}) = (320 + 310)/(400 + 400) = 630/800 = 0.7878$$

Probability that the two judges agreed by chance  $P(E) =$

$$P(\text{nonrelevant})^2 + P(\text{relevant})^2 = 0.2125^2 + 0.7878^2 = 0.665$$

Kappa statistic  $\kappa = (P(A) - P(E))/(1 - P(E)) =$

$$(0.925 - 0.665)/(1 - 0.665) = 0.776 \text{ (still in acceptable range)}$$

# 大型搜索引擎的评价

- Web下召回率难以计算
- 搜索引擎常使用top  $k$ 的正确率来度量, 比如,  $k = 10$  . . .
- . . . 或者使用一个考虑返回结果所在位置的指标, 比如正确答案在第一个返回会比第十个返回的系统给予更大的指标
- 搜索引擎也往往使用非相关度指标
  - 比如: 第一个结果的点击率
  - 仅仅基于单个点击使得该指标不太可靠 (比如你可能被检索结果的摘要所误导, 等点进去一看, 实际上是不相关的) . . .
  - 当然, 如果考虑点击历史的整体情况会相当可靠
  - 比如: 一些基于用户行为的指标
  - 比如: A/B 测试

# A/B 测试

- 目标：测试某个独立的创新点
- 先决条件：大型的搜索引擎已经在线上运行
- 很多用户使用老系统
- 将一小部分(如 1%) 流量导向包含了创新点的新系统
- 对新旧系统进行自动评价，并得到某个评价指标，比如第一个结果的点击率
- 于是，可以通过新旧系统的指标对比来判断创新点的效果
- 这也可能是大型搜索引擎最信赖的方法

# 静态摘要

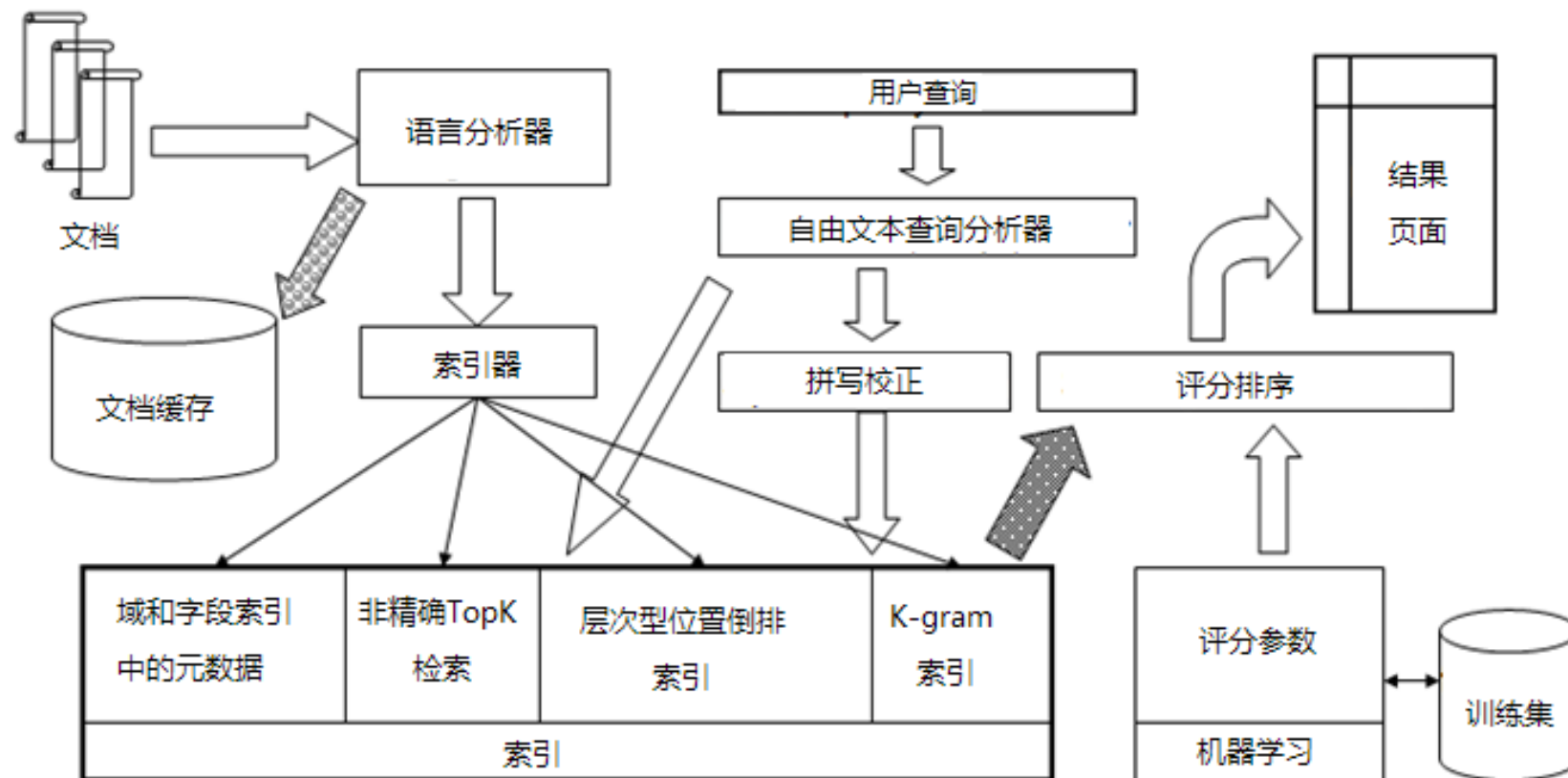
- 在典型的系统中，静态摘要是文档的一个子集
- 最简单的方法：文档的前若干个词汇
  - 在建立索引的时候就缓存好
- 更复杂的方法：从文档中抽取一些关键的句子
  - 用简单的自然语言处理的方法对句子进行打分
  - 用打分最高的几个句子组成摘要
- 最复杂的方法：用自然语言处理的方法合成摘要
  - 在IR系统中几乎不用

# 动态摘要

- 显示文档中包含查询词的一句或者几句文字
  - “KWIC” 片段: Keyword in Context presentation
- 快速在文档中寻找包含查询词的“窗口”（范围）
- 根据查询对文档中上述窗口打分
  - 用多种特征，如窗口的长度，在文档中的位置，等等
  - 用一个打分函数融合多种特征
- 最终将满足条件的窗口显示出来作为摘要

# 回顾：检索系统

能否让查询  
结果更相关？



# 提纲

- ① 上一讲回顾
- ② 动机
- ③ 相关反馈基础
- ④ 相关反馈详细介绍
- ⑤ 查询扩展

# 搜索中提高召回率的方法

- 本讲的主题：两种提高召回率的方法—相关反馈及查询扩展
- 考虑查询 $q$ : [aircraft] ...
- 某篇文档  $d$  包含 “plane”, 但是不包含 “aircraft”
- 显然对于查询 $q$ , 一个简单的IR系统不会返回文档 $d$ , 即使 $d$ 是和 $q$ 最相关的文档
- 我们试图改变这种做法:
- 也就是说, 我们会返回不包含查询词项的相关文档。



# 关于召回率Recall

- 本讲当中会放松召回率的定义，即(在前几页)给**用户返回**更多的相关文档。
- 这可能实际上会**降低召回率**，比如，将jaguar扩展为jaguar(美洲虎；一种汽车品牌)+**panthera**(豹属)
  - 可能会**去掉**一些相关的文档，但是可能**增加前几页**返回给用户的相关文档数

# 提高召回率的方法

- **局部(local)方法**: 对用户查询进行局部的即时的分析
  - 主要的局部方法: 相关反馈(relevance feedback)
- **全局(Global)方法**: 进行一次性的全局分析(比如分析整个文档集)来产生同/近义词词典 (thesaurus)
  - 利用该词典进行查询扩展

# 提纲

- ① 上一讲回顾
- ② 动机
- ③ 相关反馈基础**
- ④ 相关反馈详细介绍
- ⑤ 查询扩展

# 相关反馈的基本思想

- **相关反馈：**用户对初始返回结果的相关性进行反馈
  - 用户提交一个查询
  - 用户将部分结果标记为相关或者不相关
  - 系统根据用户的反馈，对信息需求进行优化，将其表示成更好的形式
  - 相关反馈可以进行多次循环
  - **Idea:如果不能很好地了解文档集合，就很难把自己的信息需求转化成查询，进行多次相关反馈可以有所帮助。**

# 相关反馈分类

- 用户相关反馈或显式相关反馈 (User Feedback or Explicit Feedback): 用户显式参加交互过程
- 隐式相关反馈 (Implicit Feedback): 系统跟踪用户的行为来推测返回文档的相关性, 从而进行反馈。
- 伪相关反馈或盲相关反馈 (Pseudo Feedback or Blind Feedback): 没有用户参与, 系统直接假设返回文档的前 $k$ 篇是相关的, 然后进行反馈。

# 相关反馈

- 下面使用 “**ad hoc retrieval** ” 来指未使用相关反馈的检索。
- 下面来看一下相关反馈的例子。

# 例1：类似页面

[Advanced Search](#)  
[Preferences](#)[Web](#) [Video](#) [Music](#)

## [Sarah Brightman Official Website - Home Page](#)

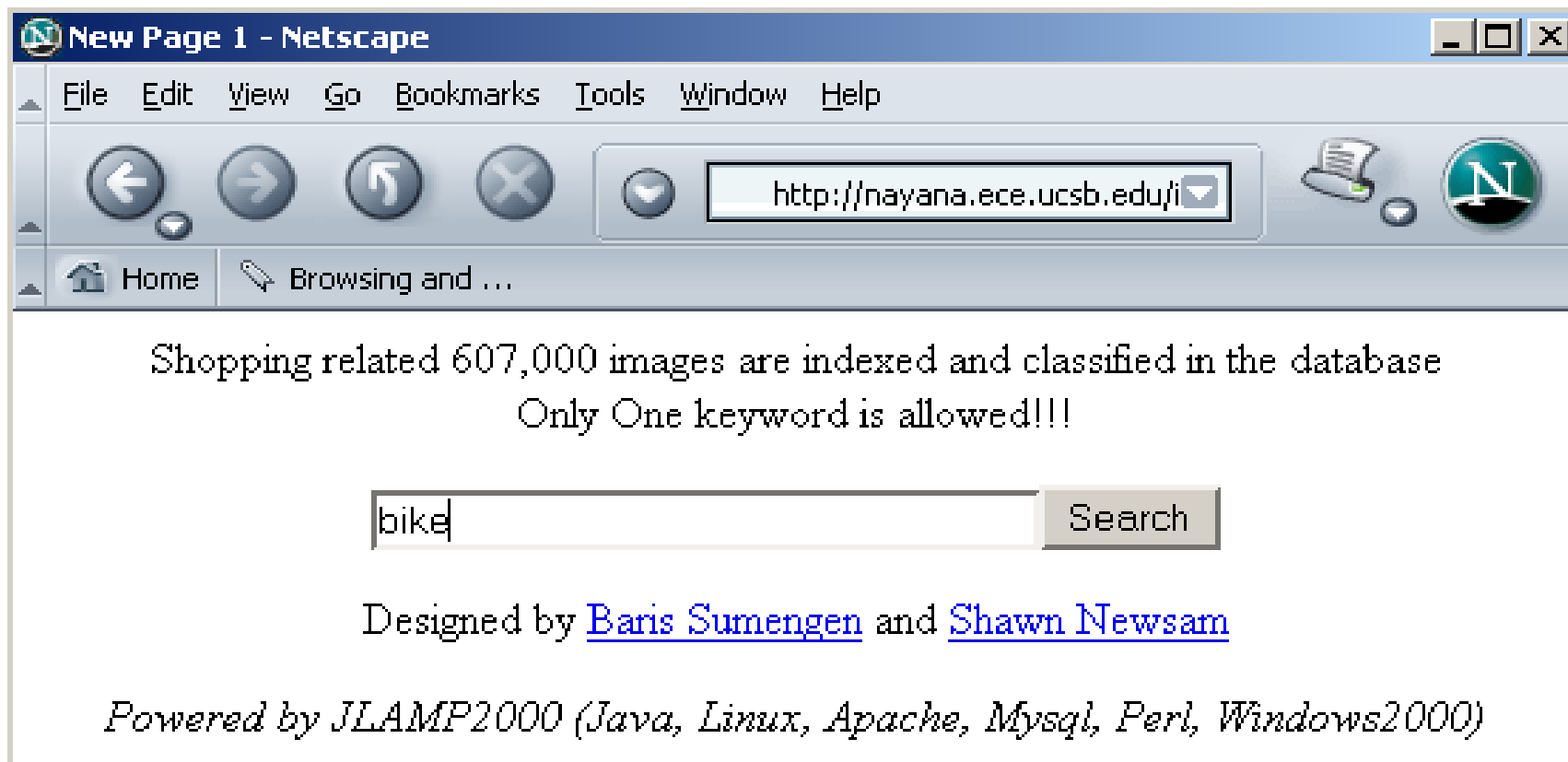
Official site of world's best-selling soprano. Join FAN AREA free to access exclusive perks, photo diaries, a global forum community and more...

www.sarah-brightman.com/ - 4k - [Cached](#) [Similar pages](#)

## 例2:

- Image search engine

<http://nayana.ece.ucsb.edu/imsearch/imsearch.html>

















# 首次查询的返回结果

Interface showing search results for a query, displaying a grid of images and their associated coordinates and scores.













Navigation buttons: Browse, Search, Prev, Next, Random

					
(144473, 16458)	(144457, 252140)	(144456, 262857)	(144456, 262863)	(144457, 252134)	(144483, 265154)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
					
(144483, 264644)	(144483, 265153)	(144518, 257752)	(144538, 525937)	(144456, 249611)	(144456, 250064)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0













# 相关反馈

绿框表示用户认为相关的结果

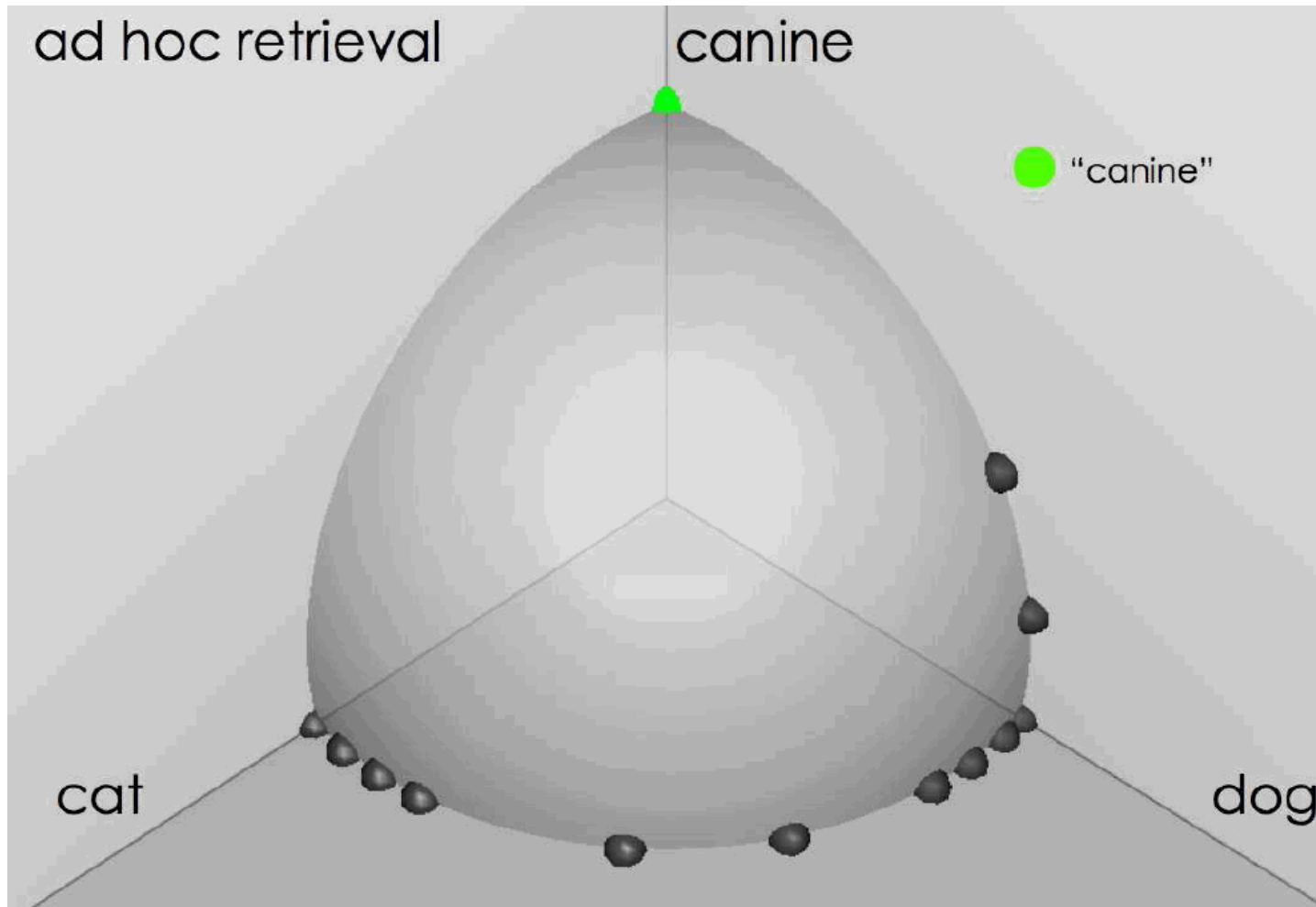
Interface showing a grid of 12 images related to bicycles and motorcycles, with a mouse cursor pointing at the first image. The interface includes navigation buttons: Browse, Search, Prev, Next, Random.

Image 1	Image 2	Image 3	Image 4	Image 5	Image 6
					
(144473, 16458)	(144457, 252140)	(144456, 262857)	(144456, 262863)	(144457, 252134)	(144483, 265154)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
					
(144483, 264644)	(144483, 265153)	(144518, 257752)	(144538, 525937)	(144456, 249611)	(144456, 250064)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0

# 相关反馈后再次检索的结果

<a href="#">Browse</a> <a href="#">Search</a> <a href="#">Prev</a> <a href="#">Next</a> <a href="#">Random</a>					
					
(144538, 523493) 0.54182 0.231944 0.309876	(144538, 523835) 0.56319296 0.267304 0.295889	(144538, 523529) 0.584279 0.280881 0.303398	(144456, 253569) 0.64501 0.351395 0.293615	(144456, 253568) 0.650275 0.411745 0.23853	(144538, 523799) 0.66709197 0.358033 0.309059
					
(144473, 16249) 0.6721 0.393922 0.278178	(144456, 249634) 0.675018 0.4639 0.211118	(144456, 253693) 0.676901 0.47645 0.200451	(144473, 16328) 0.700339 0.309002 0.391337	(144483, 265264) 0.70170796 0.36176 0.339948	(144478, 512410) 0.70297 0.469111 0.233859

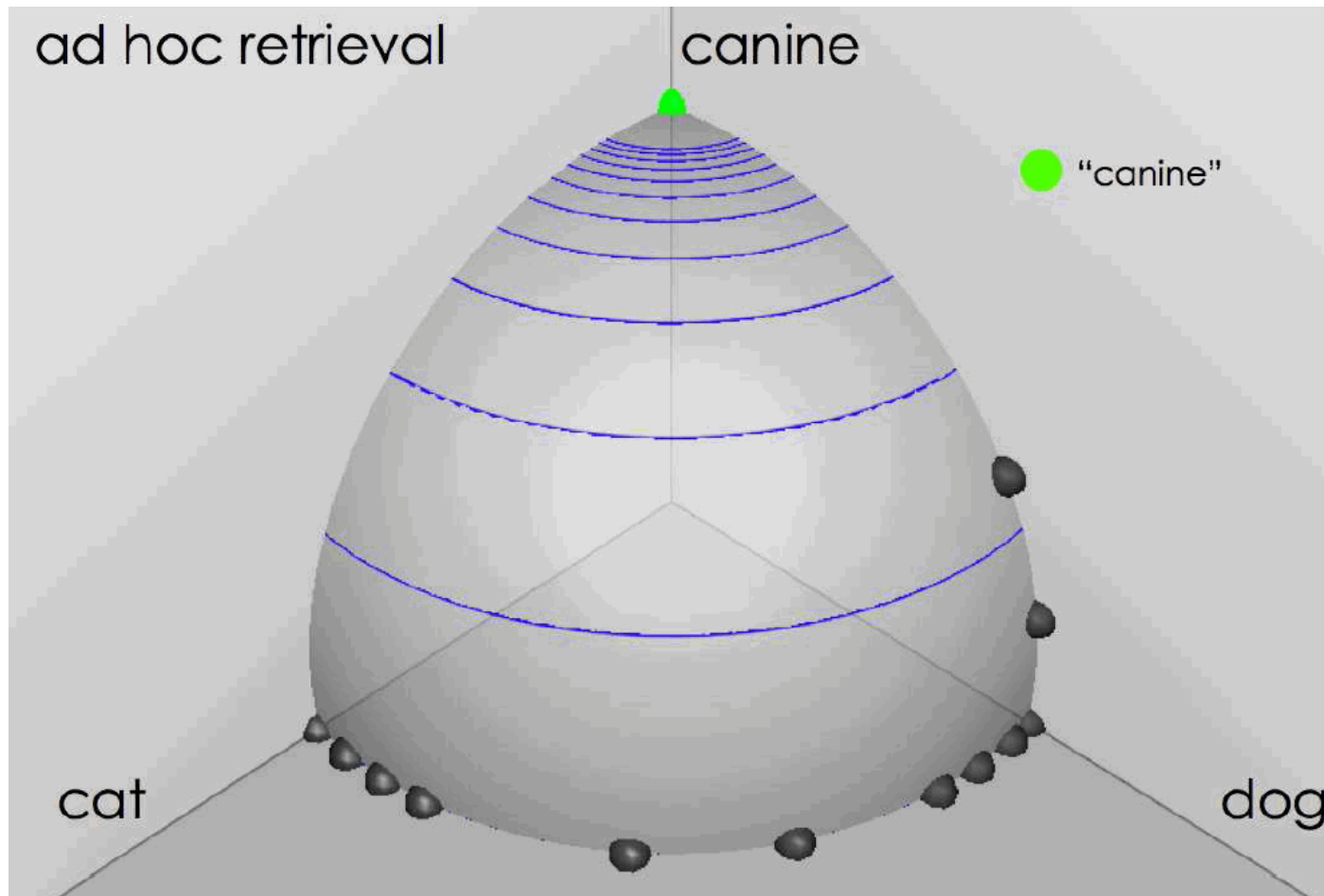
# 例3：向量空间的例子：查询 “canine”



Source:

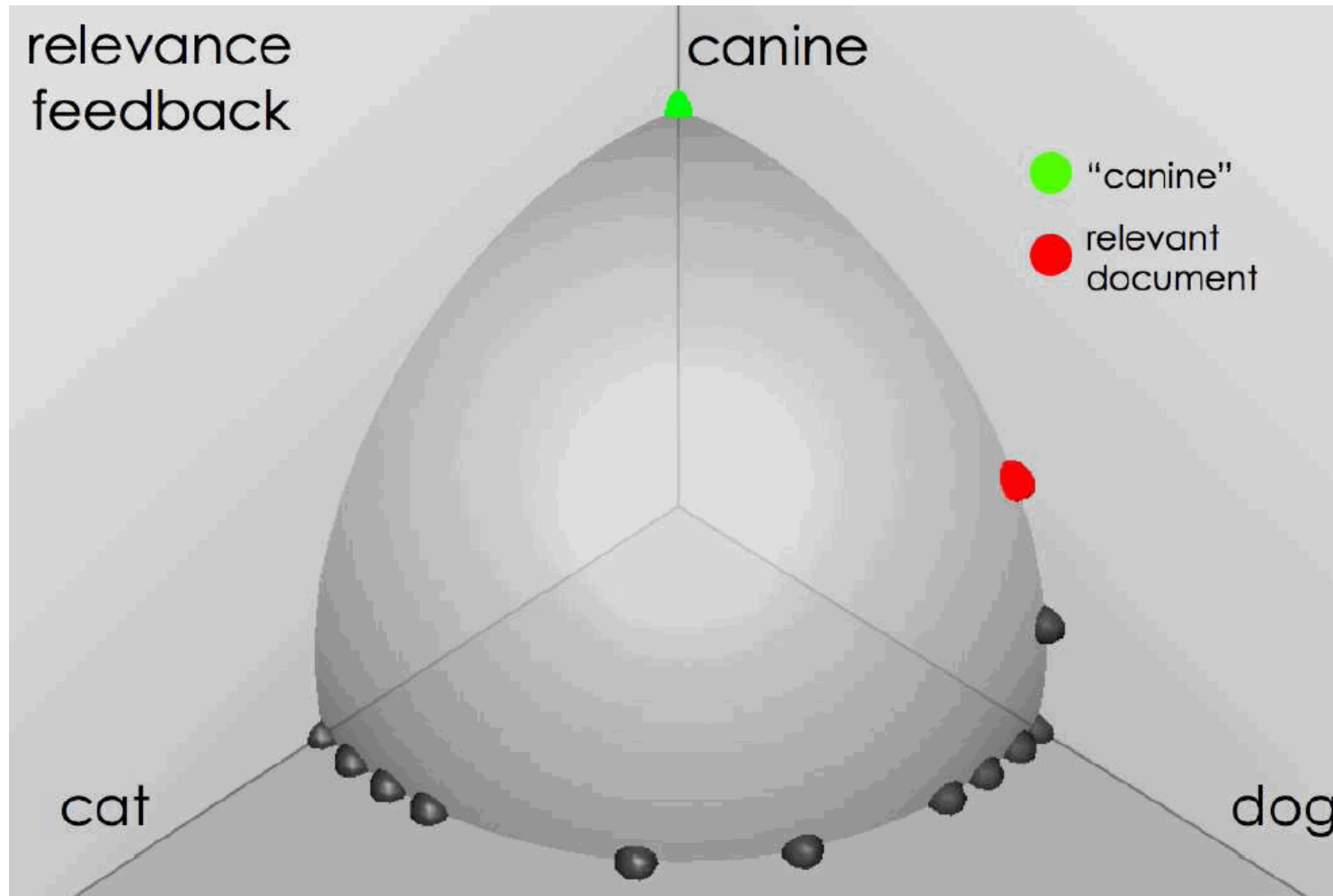
Fernando Díaz

# 文档和查询 “canine” 的相似度



Source:  
Fernando Díaz

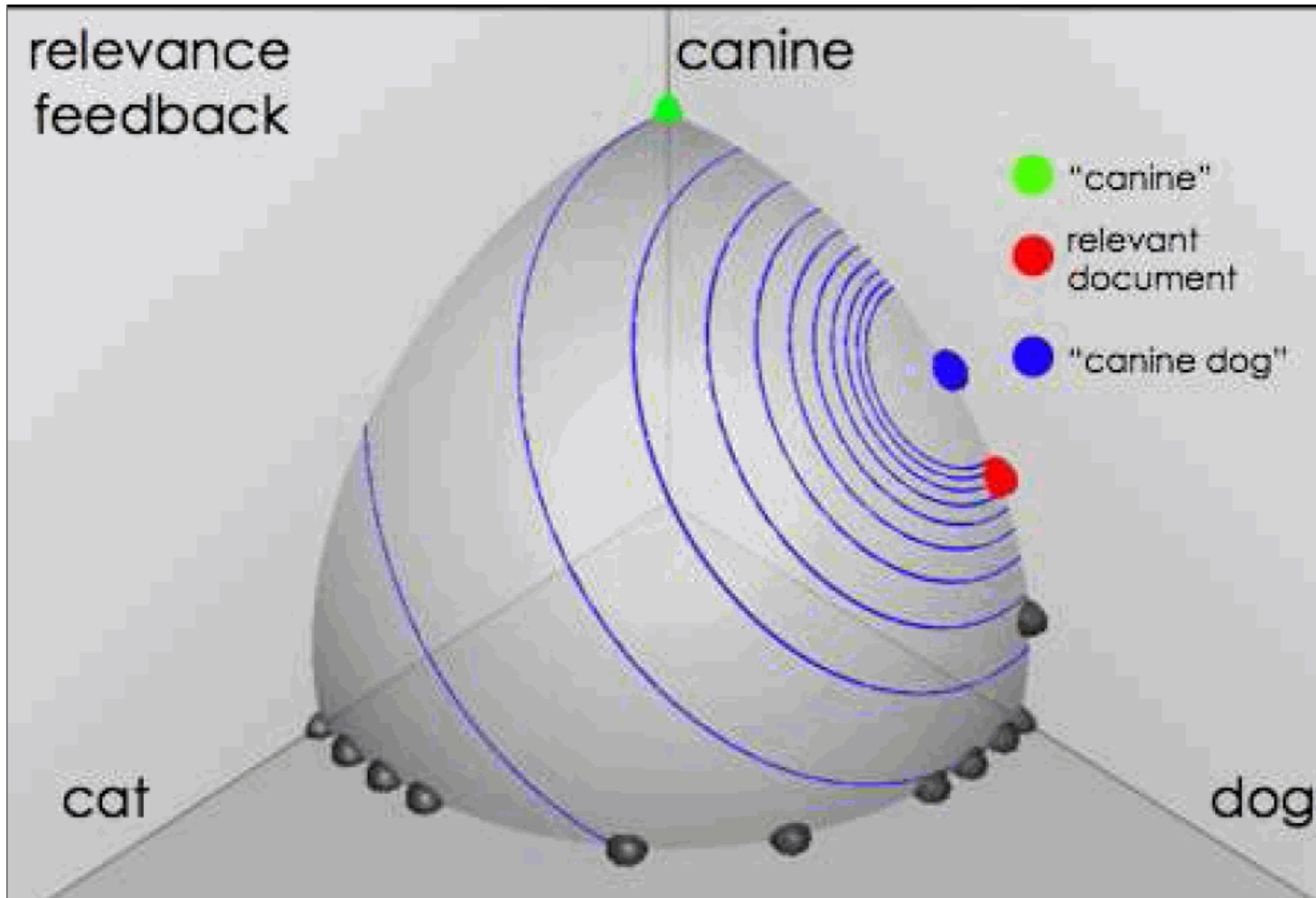
# 用户的反馈：选择一个认为相关的文档



Source:  
Fernando Díaz



# 查询扩展后的结果



Source:  
Fernando Díaz

## 例4：一个实际的例子

### ● 初始查询: *New space satellite applications*

- + 1. 0.539, 08/13/91, [NASA Hasn't Scrapped Imaging Spectrometer](#)
- + 2. 0.533, 07/09/91, [NASA Scratches Environment Gear From Satellite Plan](#)
- 3. 0.528, 04/04/90, [Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes](#)
- 4. 0.526, 09/09/91, [A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget](#)
- 5. 0.525, 07/24/90, [Scientist Who Exposed Global Warming Proposes Satellites for Climate Research](#)
- 6. 0.524, 08/22/90, [Report Provides Support for the Critics Of Using Big Satellites to Study Climate](#)
- 7. 0.516, 04/13/87, [Arianespace Receives Satellite Launch Pact From Telesat Canada](#)
- + 8. 0.509, 12/02/87, [Telecommunications Tale of Two Companies](#)

### ● 用户使用“+”标记相关的文档



## 相关反馈之后，扩展了的查询

<b>2.074</b>	<b>new</b>	<b>15.106</b>	<b>space</b>
30.816	satellite	5.660	application
5.991	nasa	5.196	eos
4.196	launch	3.972	aster
3.516	instrument	3.446	arianespace
3.004	bundespost	2.806	ss
2.790	rocket	2.053	scientist
2.003	broadcast	1.172	earth
0.836	oil	0.646	measure

# 扩展查询的结果

- 2 1. 0.513, 07/09/91, [NASA Scratches Environment Gear From Satellite Plan](#)
- 1 2. 0.500, 08/13/91, [NASA Hasn't Scrapped Imaging Spectrometer](#)
- 3. 0.493, 08/07/89, [When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own](#)
- 4. 0.493, 07/31/89, [NASA Uses 'Warm' Superconductors For Fast Circuit](#)
- 8 5. 0.492, 12/02/87, [Telecommunications Tale of Two Companies](#)
- 6. 0.491, 07/09/91, [Soviets May Adapt Parts of SS-20 Missile For Commercial Use](#)
- 7. 0.490, 07/12/88, [Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers](#)
- 8. 0.490, 06/14/90, [Rescue of Satellite By Space Agency To Cost \\$90 Million](#)

# 提纲

- ① 上一讲回顾
- ② 动机
- ③ 相关反馈基础
- ④ 相关反馈详细介绍
- ⑤ 查询扩展

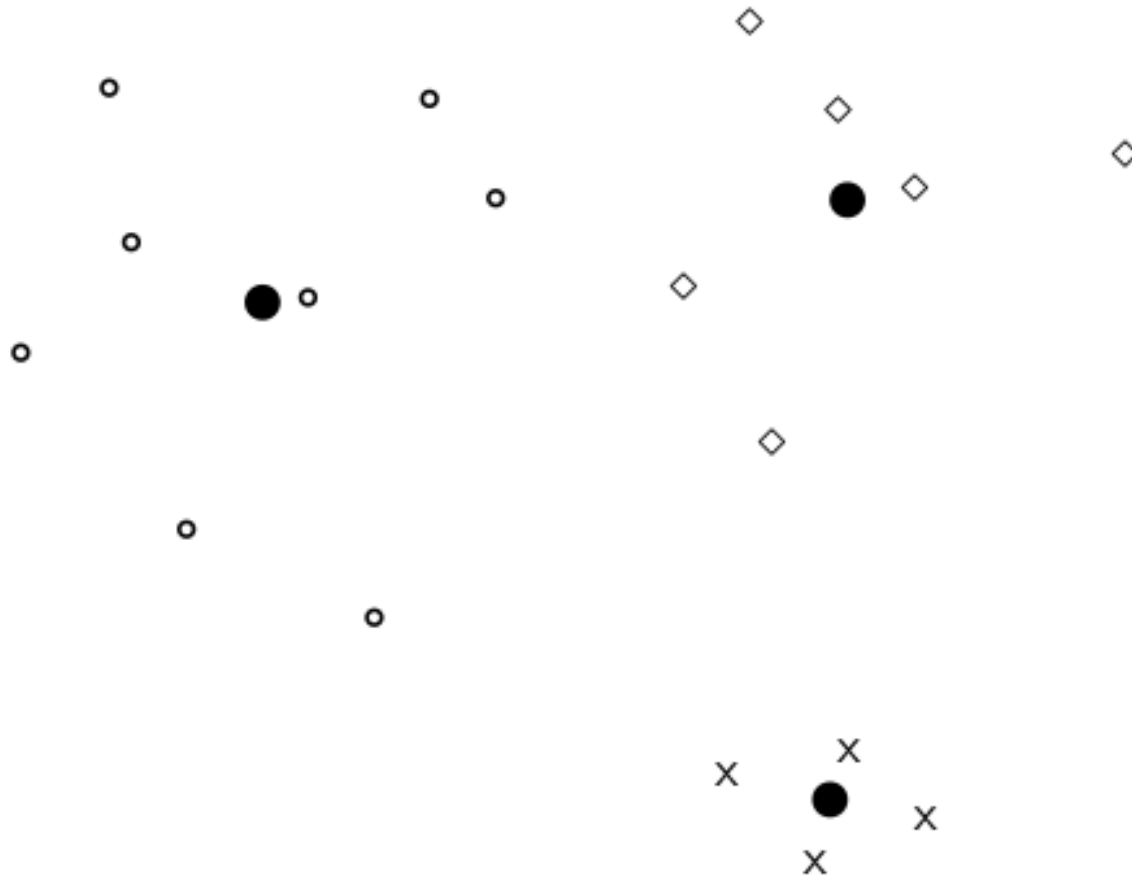
# 相关反馈中的核心概念：质心

- 质心是一系列点的质量的中心
- 回顾一下：我们将文档看作高维空间中的点
- 我们可以采用如下方式计算文档的质心

$$\vec{\mu}(D) = \frac{1}{|D|} \sum_{d \in D} \vec{v}(d)$$

其中  $D$  是一个文档集合， $\vec{v}(d) = \vec{d}$  是文档  $d$  的向量表示

# 质心的例子



# 罗基奥 (Rocchio) 算法

- Rocchio 算法使用向量空间模型来收集相关反馈
- Rocchio 算法试图寻找一个查询  $\vec{q}_{opt}$ ，使得：

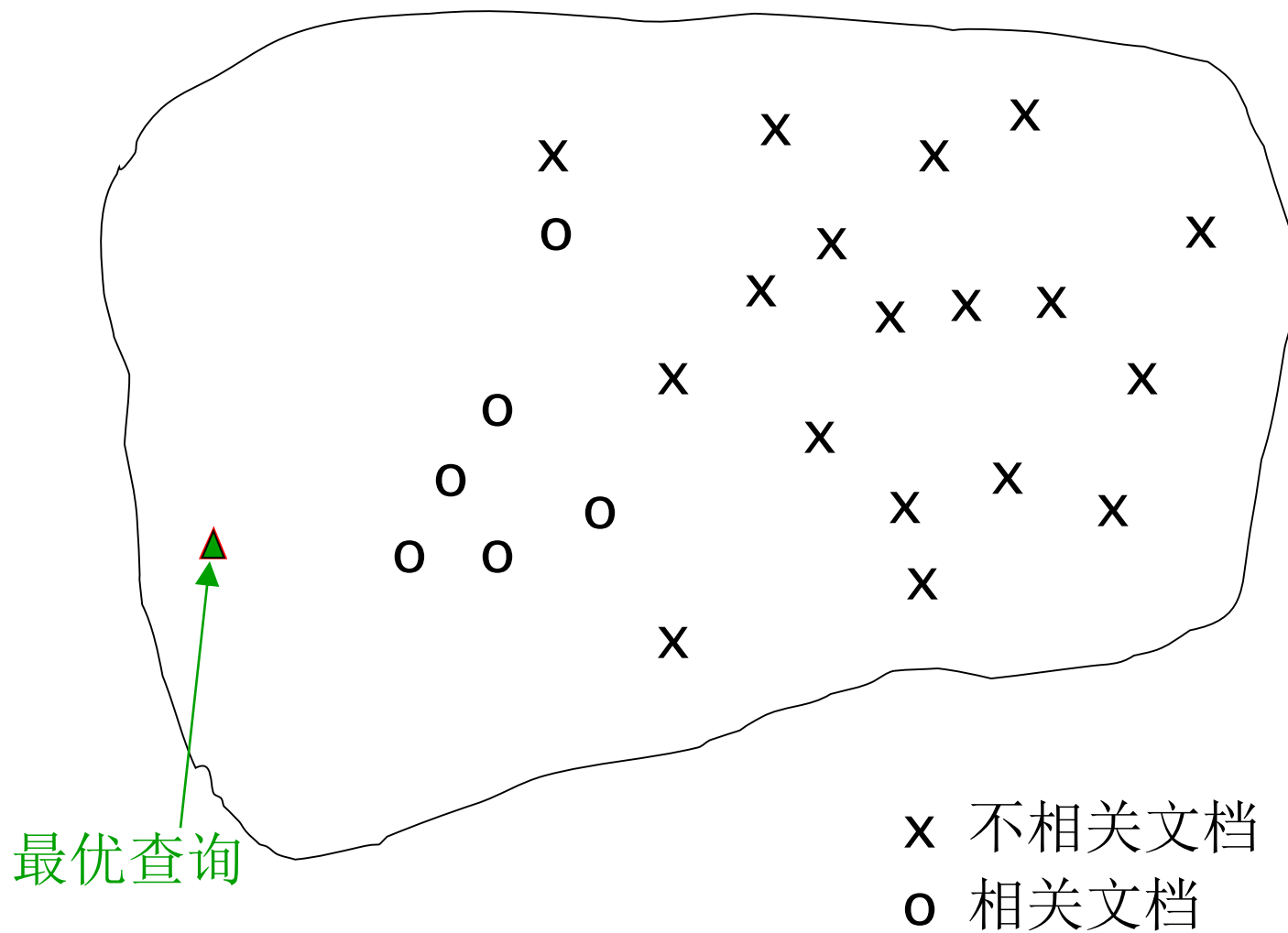
$$\vec{q}_{opt} = \arg \max_{\vec{q}} [\cos(\vec{q}, \vec{\mu}(C_r)) - \cos(\vec{q}, \vec{\mu}(C_{nr}))]$$

- 试图将相关文档和不相关文档分开

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\vec{d}_j \in C_r} \vec{d}_j - \frac{1}{|C_{nr}|} \sum_{\vec{d}_j \notin C_r} \vec{d}_j$$

- 问题是，我们并不知道哪些文档是真正相关的

# 理论上的最好的查询



# Rocchio 1971 算法 (SMART)

- 实际使用:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

- $D_r$  = 已知相关文档的向量集合
- $D_{nr}$  = 已知的不相关文档的向量集合
  - 注意和文档集合  $D_r$  和  $D_{nr}$  不同
- $q_m$  = 优化过的查询向量;  $q_0$  = 原始的查询向量;  $\alpha, \beta, \gamma$ : 权重 (手工或者根据经验设定)
- 新的查询向量向相关文档向量移动, 远离不相关文档向量

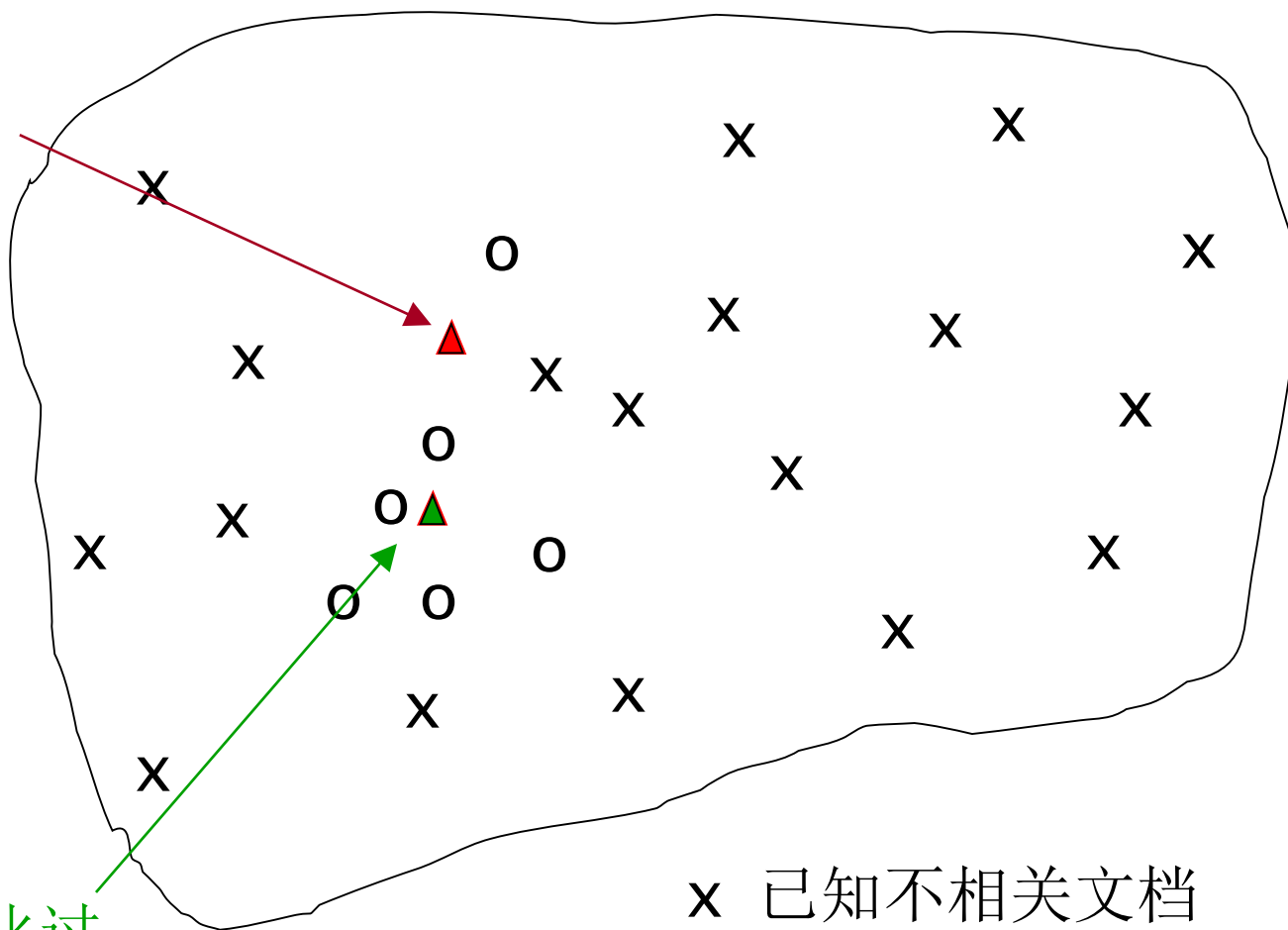


# 需要注意的细节

- $\alpha$  和  $\beta/\gamma$  的权衡: 如果很多文档已经评价了相关度, 那么  $\beta/\gamma$  应该大一些.
- 查询向量的某些权值可能为负数
  - 忽略负的权值

# 对初始查询的相关反馈

原始  
查询



优化过的  
查询

x 已知不相关文档

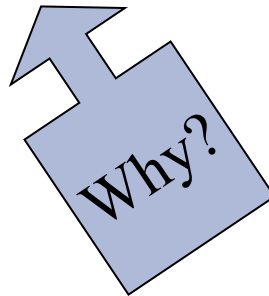
o 已知相关文档

# 向量空间中的相关反馈

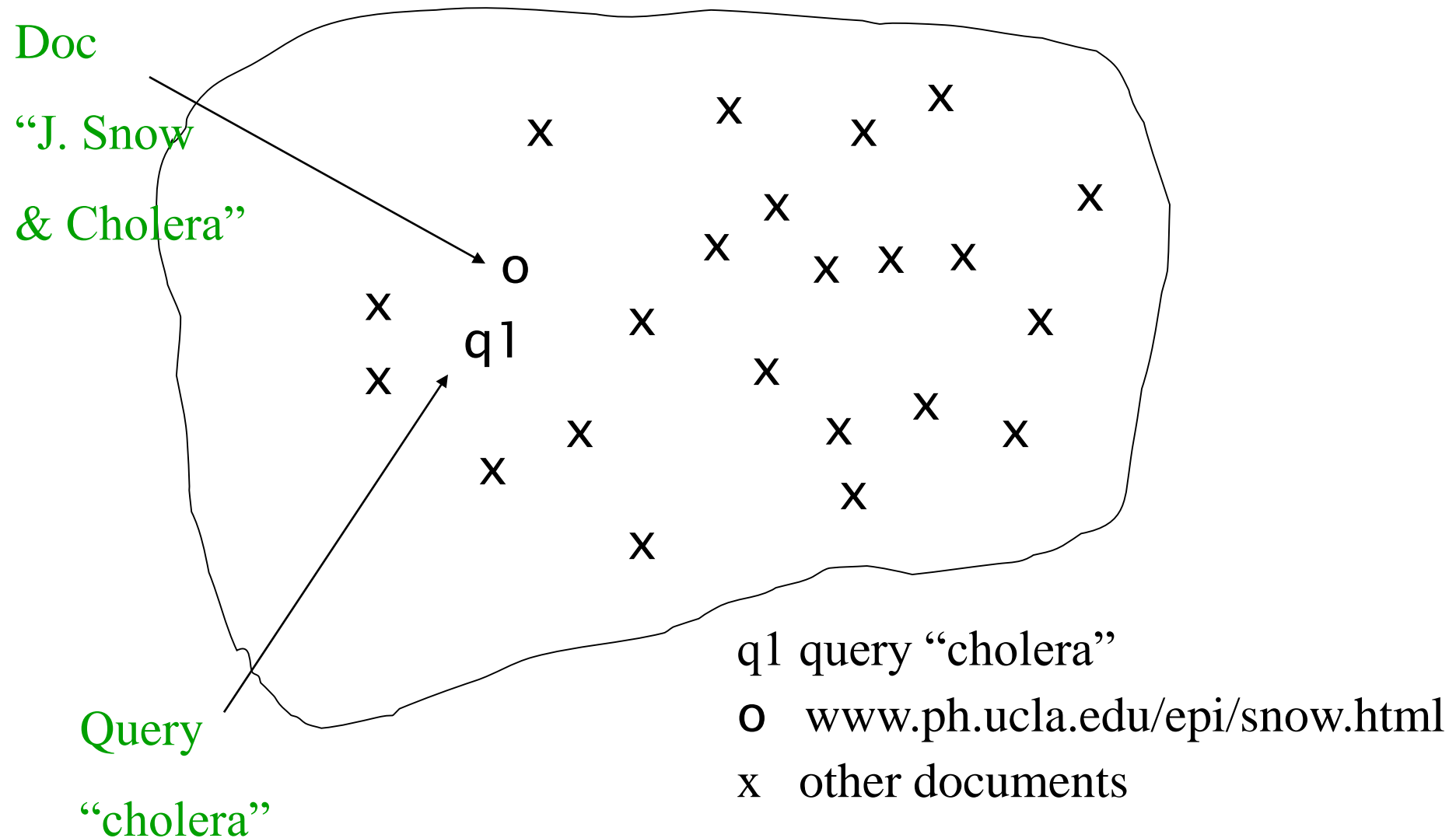
- 根据相关反馈修改查询，然后使用向量空间模型
- 仅使用标记过的文档.
- 相关反馈可以提高查全率和查准率
- 当查全率很重要的时候，相关反馈是最有用的提高查全率的方法
  - 期望用户查看结果后，愿意花时间来反馈，或者多次反馈

# 正反馈 vs 负反馈

- 正反馈比负反馈更有价值(即设置 $\gamma < \beta$ ; e.g.  $\gamma = 0.25$ ,  $\beta = 0.75$ ).
- 许多系统只使用正反馈( $\gamma=0$ ).



# 向量空间可能与直觉不符



# 高维向量空间

- 查询“cholera” 和 “john snow” 在向量空间中离得很远
- 文档“John Snow and Cholera” 如何和两个查询离得都近
- 我们处理2、3维的直觉方法在高维空间不行，比如维数  $>10,000$  的时候.
- 三维的情况：如果一个文档和许多查询离得很近，那么这些查询中的某几个查询相互离得很近.
- 上面的情况在高维空间不成立.

# 相关反馈的假设

- A1: 用户对于初始查询有充分的认识.
- A2: 相关文档的原型有一种良好的形式.
  - 相关文档的词项分布相似
  - 不相关的文档的词项分布和相关文档的词项分布不相似
    - 所有相关文档都聚集在某个原型（**prototype**）周围，形成一个簇.
    - 或者: 有不同的原型, 但是它们的词汇有很大重合.
    - 相关文档和不相关文档的相似度很小

# 不满足A1假设的情况

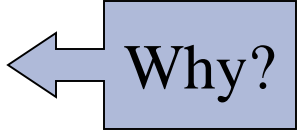
- 用户没有足够的知识来建立一个初始的查询.
- 比如:
  - 拼写错误(小田田布兰妮).
  - 跨语言的搜索(hígado).
  - 用户的词汇和文档集合里的词汇不吻合
    - 硬盘/磁碟



# 不满足A2假设的情况

- 相关文档聚成几个不同的簇.
- 这种情况可能发生的情形:
  - 文档子集使用不同的词汇，如Burma/Myanmar（缅甸）
  - 某个查询的答案本身就需要不同类的文档来组成，如  
**Pop stars that worked at Burger King**
- 通用概念需要由多个具体概念体现
- 文档当中精心编辑的内容往往可以解决上述的问题

# 相关反馈的问题

- 长的查询对典型的IR系统是低效的.  Why?
  - 将结果返回给用户的耗时较长.
  - 检索系统的消耗大.
  - 能部分解决这个问题方法:
    - 相关反馈时只对重要的查询词项重新计算权值
    - 比如按照词频，取前20
- 用户一般不太情愿提供明确的反馈
- 在使用了相关反馈之后，可能很难解释某个文档为什么会被返回

# 相关反馈策略的评价

- 使用初始查询  $q_0$ ，然后计算“查准率-查全率”曲线
- 使用相关反馈后修改的查询  $q_m$ ，然后计算“查准率-查全率”曲线
- 方法一、在整个文档集合上评价
  - 有显著的改善, 但是有作弊的嫌疑
  - 部分原因是会把已知的相关文档排的很前
  - 需要用用户没有看到的文档集合来评价
- 方法二、使用剩余的文档集合来评价(总的文档集合减去评价过相关性的文档)
  - 评价结果往往比初始查询的结果差
  - 但是这种方法更现实
  - 可以用来有效地比较不同相关反馈方法之间的相对效果

# 相关反馈策略的评价

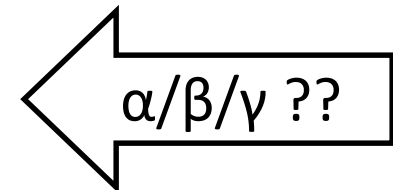
- 方法三、使用两个文档集合
  - 在第一个文档集合上使用初始查询 $q_0$ ，并进行相关反馈
  - 在第二个文档集合上使用初始查询 $q_0$ 和修改过的查询 $q_m$ 进行评价
- 从经验上说，一轮相关相关反馈很有用. 两轮相关反馈的效果就不那么明显.

# 评价的误区

- 评价不同相关反馈方法的效用的时候，必须考虑消耗时间的要素.
- 代替相关反馈的方法：用户修改并重新提交查询.
- 相对于判断文档的相关性，用户可能更愿意修改并重新提交查询.
- 没有证据能表明相关反馈占用了用户的时间就能给用户带来最大的效用.

# Web上的相关反馈

- 一些搜索引擎提供“相似或相关网页”的功能(这是一种简单形式的相关反馈)
  - Google (link-based)
  - Altavista
  - Stanford WebBase
- 有些搜索引擎没有，因为很难向普通用户解释清楚什么是相关反馈：
  - Alltheweb
  - bing
  - Yahoo
- Excite搜索引擎最初有相关反馈, 但由于没有人用，所以就取消了



# Excite搜索引擎的相关反馈

Spink et al. 2000

- 只有大概4% 的查询会话使用了相关反馈
  - 这些都是通过查询结果后面的“More like this” 链接来实现的
- 70%的用户只浏览了第一页的结果

# 间接相关反馈

- 可以使用间接的资源进行相关反馈。比如 DirectHit 搜索引擎
- DirectHit将用户点击频率高的文档排在前面.
  - 点击的多的页面被认为是相关的.
  - 从用户的点击记录中挖掘信息，进行相关反馈
- 这种方法是全局的，并不依赖特定用户或查询.
  - 这是点击流挖掘（clickstream mining）的典型应用场景
- 现在这是通过机器学习产生排序的一部分



# 隐式相关反馈

- 通过观察用户对当前检索结果采取的行为来给出对检索结果的相关性判定。
- 判定不一定很准确，但是省却了用户的显式参与过程。
- 对用户非当前检索行为或非检索相关行为的分析也可以用于提高检索的效果，这些是个性化信息检索(Personalized IR)的主要研究内容，并非本节的主要内容。

# 用户行为种类

- 鼠标键盘动作：
  - 点击链接、加入收藏夹、拷贝粘贴、停留、翻页等等
- 用户眼球动作
  - Eye tracking可以跟踪用户的眼球动作
  - 拉近、拉远、瞟、凝视、往某个方向转

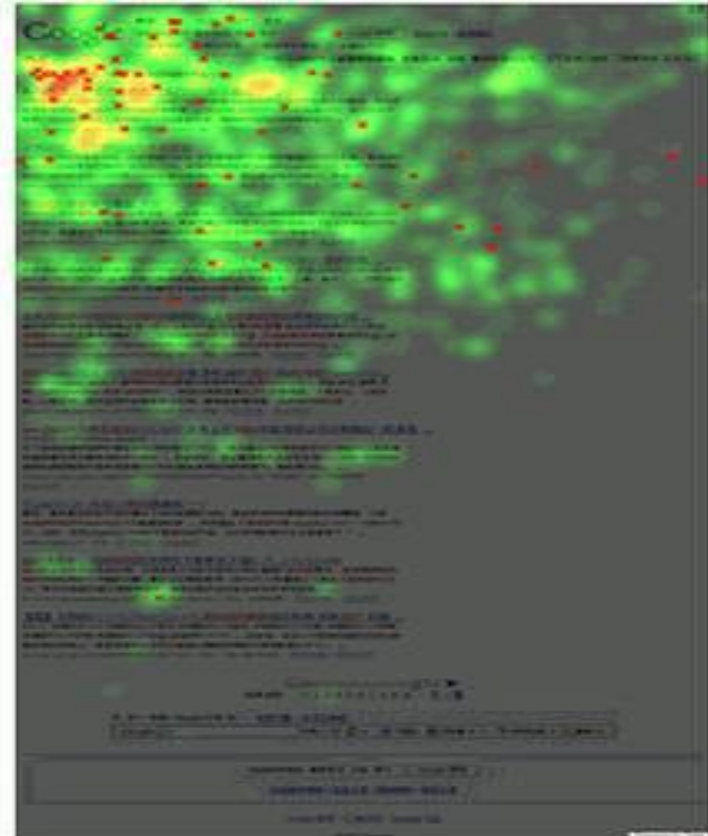
# 点击行为 (Click through behavior)

FIELD	VALUE
User ID	1162742023015
Time stamp	06/Nov/2006:00:01:35
Query terms	嫁给警察的理由
URL	<a href="http://bbs.cixi.cn/dispbbs.asp?Star=4&amp;boardid=46&amp;id=346721&amp;page=1">http://bbs.cixi.cn/dispbbs.asp?Star=4&amp;boardid=46&amp;id=346721&amp;page=1</a>
Page number	1
Rank	7
Anchor text	姑娘们，你们愿意<FONT color=#cc0033>嫁给警察</FONT>吗？ [慈溪社区]

# 眼球动作 (通过鼠标轨迹模拟)

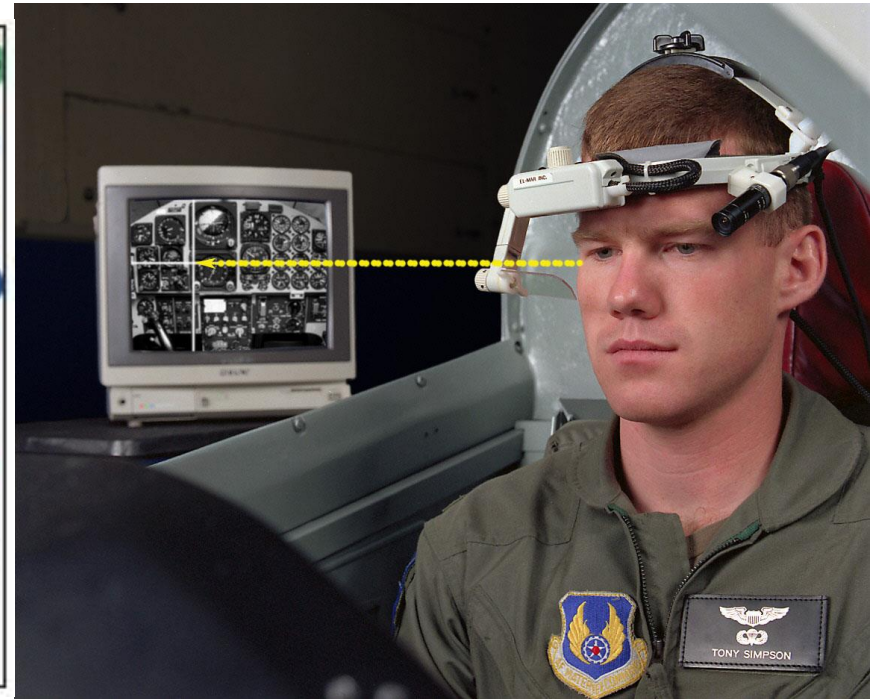
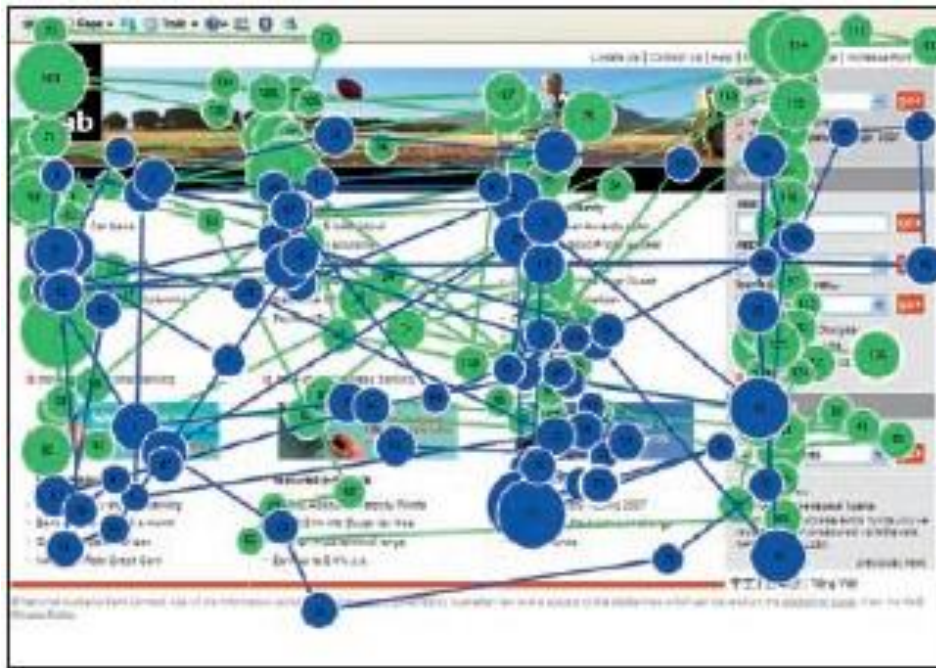


Baidu



Google

# 关于Eye tracking



# 隐式相关反馈小结

- 优点：
  - 不需要用户显式参与，减轻用户负担
  - 用户行为某种程度上反映用户的兴趣，具有可行性
- 缺点：
  - 对行为分析有较高要求
  - 准确度不一定能保证
  - 某些情况下需要增加额外设备

# 伪相关反馈

- 伪相关反馈将相关反馈部分的人工操作自动化.
- 伪相关反馈的算法:
  - 根据用户的查询, 检索出结果列表
  - 假设列表中前 $k$ 个结果是相关的.
  - 进行相关反馈(e.g., Rocchio)
- 平均效果很好
- 但对于某些查询可能错的很严重.
- 几步迭代后就可能出现严重的偏移.
- 为什么?

# TREC4上的伪相关反馈实验

- 使用Cornell大学的SMART系统
- 50个查询，每个查询基于前100个结果进行反馈 (因此所有的反馈文档数目是5000):

检索方法	相关文档数目
Inc.ltc	3210
Inc.ltc-PsRF	3634
Lnu.ltu	3709
Lnu.ltu-PsRF	4350

- 比较了两种长度归一化机制 (L vs. I) 以及反馈不反馈后的结果 (PsRF).
- 实验中的伪相关反馈方法对查询只增加了20个词项 (Rocchio将增加更多的词项)
- 上述结果表明，伪相关反馈在平均意义上说是有效的方法



# 伪相关反馈小结

- 优点：
  - 不用考虑用户的因素，处理简单
  - 很多实验也取得了较好效果
- 缺点：
  - 没有通过用户判断，所以准确率难以保证
  - 不是所有的查询都会提高效果

# 提纲

- ① 上一讲回顾
- ② 动机
- ③ 相关反馈基础
- ④ 相关反馈详细介绍
- ⑤ 查询扩展

# 查询扩展

- 在相关反馈中，用户针对文档的相关或者不相关给出额外的输入，这些输入将被用来重新计算查询词项的权值
- 在查询扩展中，用户针对词汇或短语的好坏给出额外的输入

# 查询提示

[Web](#) | [Images](#) | [Video](#) | [Local](#) | [Shopping](#) | [more](#) ▼

sarah p

Search

[Options](#) ▼

YAHOO!

sarah palin

sarah palin saturday night live

sarah polley

sarah paulson

snl sarah palin

# 如何扩展用户的查询?

- 利用人工编纂的同义词辞典
  - E.g. MedLine: physician, 同义词: doc, doctor, MD, medico
  - 这些同义词可以作为查询
- 全局的分析: (static; of all documents in collection)
  - 同义词辞典的自动生成
    - 统计词汇的共现 (co-occurrence)
  - 利用对查询日志的挖掘进行优化
    - Web中最常用的
- 局部的分析: (动态的)
  - 分析查询的结果文档集合

# 人工编纂的范例

The screenshot shows the PubMed interface. At the top, there are logos for NCBI, PubMed, and the National Library of Medicine (NLM). Below the logos is a navigation bar with tabs for PubMed, Nucleotide, Protein, Genome, Structure, PopSet, and Taxonomy. The main search area has a search bar with the text 'cancer' and buttons for 'Go' and 'Clear'. Below the search bar are links for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. On the left side, there is a sidebar with links for 'About Entrez', 'Text Version', 'Entrez PubMed', 'Overview', 'Help | FAQ', 'Tutorial', 'New/Noteworthy', 'E-Utilities', 'PubMed Services', 'Journals Database', 'MeSH Browser', 'Single Citation', and 'MetaBox'. The main content area displays the 'PubMed Query:' and a text box containing the query string: `("neoplasms"[MeSH Terms] OR cancer[Text Word])`. At the bottom of the text box are buttons for 'Search' and 'URL'.

NCBI

PubMed

National Library of Medicine NLM

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy

Search PubMed for cancer Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Text Version

Entrez PubMed

Overview

Help | FAQ

Tutorial

New/Noteworthy

E-Utilities

PubMed Services

Journals Database

MeSH Browser

Single Citation

MetaBox

PubMed Query:

`("neoplasms"[MeSH Terms] OR cancer[Text Word])`

Search URL

# 利用同义词辞典进行查询扩展

- 对查询词汇  $t$ , 使用辞典中的同义词或者词汇进行扩展
  - feline → feline cat
- 相对于原始的查询词汇, 可以给扩展的词汇分配更小的权重.
- 通常可以增加查全率
- 在科研和工程领域广泛应用
- 可能会明显地降低查准率, 特别是对于含混不清的查询词汇.
  - “interest rate” → “interest rate fascinate evaluate”
- 人工编纂同义词辞典需要很大代价

# 同义词辞典的自动生成

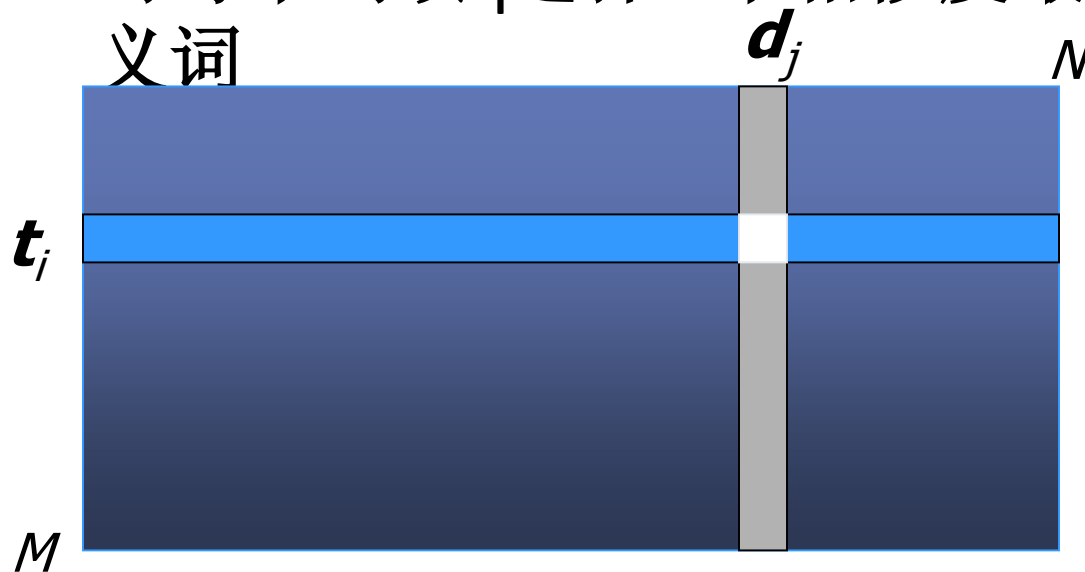
- 通过分析文档集合，可以得到两个词的相似性
- 定义1：如果两个词经常和同样的词同时出现，那么这两个词相似.
- 定义2：如果两个词经常和同样的词在某种语法关系里出现，那么这两个词相似.
- 你可以“削，吃，收获”“苹果，梨”，那么“苹果”和“梨”就相似.
- 简单采用词的共现更鲁棒，但采用语法关系更准确.

← 为什么?



# 共现同义词辞典

- 最简单的计算词和词之间相似性的方法是计算  $C = AA^T$ ,  $A$  是词项-文档矩阵.
- $w_{i,j} = (t_i, d_j)$  的归一化的权值, 使得  $A$  中的行向量大小为1
- 对每个词项  $t_i$  选择  $C$  中相似度最高的词项作为同义词



如果  $A$  是一个词项-文档出现矩阵 (0/1 矩阵),  $C$  中的元素是怎么的?

# 自动生成同义词辞典的例子

词语	同(近)义词
absolutely	absurd whatsoever totally exactly nothing
bottomed	dip copper drops topped slide trimmed
captivating	shimmer stunningly superbly plucky witty
doghouse	dog porch crawling beside downstairs
makeup	repellent lotion glossy sunscreen skin gel
mediating	reconciliation negotiate case conciliation
keeping	hoping bring wiping could some would
lithographs	drawings Picasso Dali sculptures Gauguin
pathogens	toxins bacteria organisms bacterial parasite
senses	grasp psyche truly clumsy naive innate

# 自动生成同义词词典的讨论

- 词项关联的质量是一个问题.
- 有歧义的查询词可能会引入统计上相关, 但意思不相关的词.
  - “Apple computer” → “Apple red fruit computer”
- 问题:
  - 假正率 (False positives): 不相似的词别认为相似
  - 假负率 (False negatives): 相似的词被认为不相似
- 由于扩展的查询词和原查询词很相关, 扩展的查询也未必能得到更多的相关文档.

# 总结

- 交互式相关反馈 (Interactive relevance feedback): 在初始检索结果的基础上, 通过用户交互指定哪些文档相关或不相关, 然后改进检索的结果
- 最著名的相关反馈方法: Rocchio 相关反馈
- 伪相关反馈
- 查询扩展 (Query expansion): 通过在查询中加入同义或者相关的词项来提高检索结果
  - 相关词项的来源: 人工编辑的同义词词典、自动构造的同义词词典、查询日志等等

# 课后练习

习题9-3

习题9-4

习题9-7

*谢谢大家!*