# 相关反馈及查询扩展
# Relevance Feedback & Query Expansion

# 主要内容

- **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

- **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。

# 目录

- 动机
- 相关反馈基础
- 相关反馈详细介绍
- 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，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)
  - 第一部分

- **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
  - 第二部分

# 目录

- 动机
- 相关反馈基础
- 相关反馈详细介绍
- 查询扩展

# 相关反馈的基本思想

- 用户提交一个(简短的)查询
- 搜索引擎返回一系列文档
- 用户将部分返回文档标记为**相关**的，将部分文档标记为不相关的
- 搜索引擎根据标记结果计算得到信息需求的一个**新查询表示**。当然希望该表示好于初始的查询表示
- 搜索引擎对新查询进行处理，返回新结果
- 新结果可望(理想上说)有更高的召回率

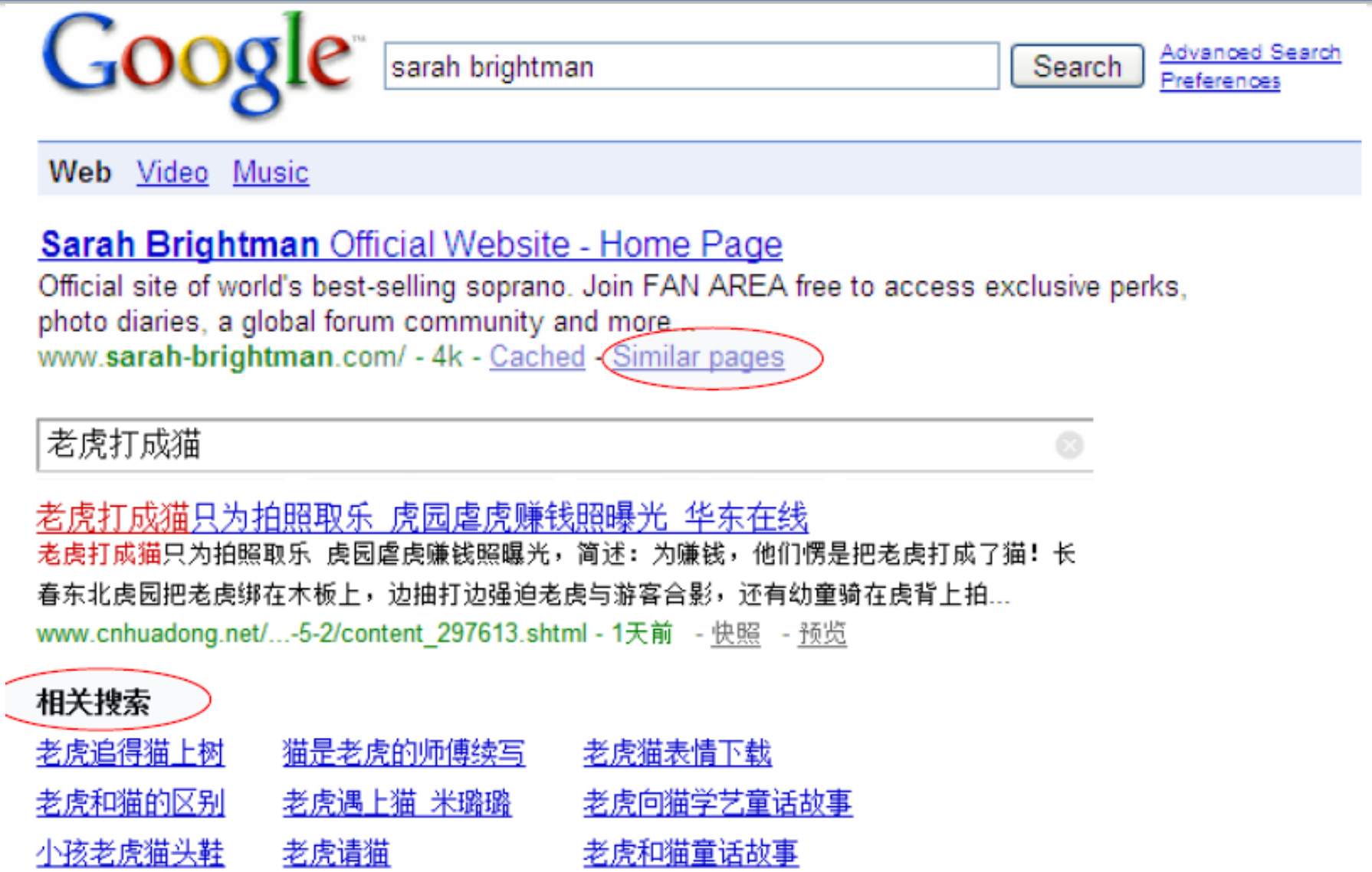
# 相关反馈分类

- 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

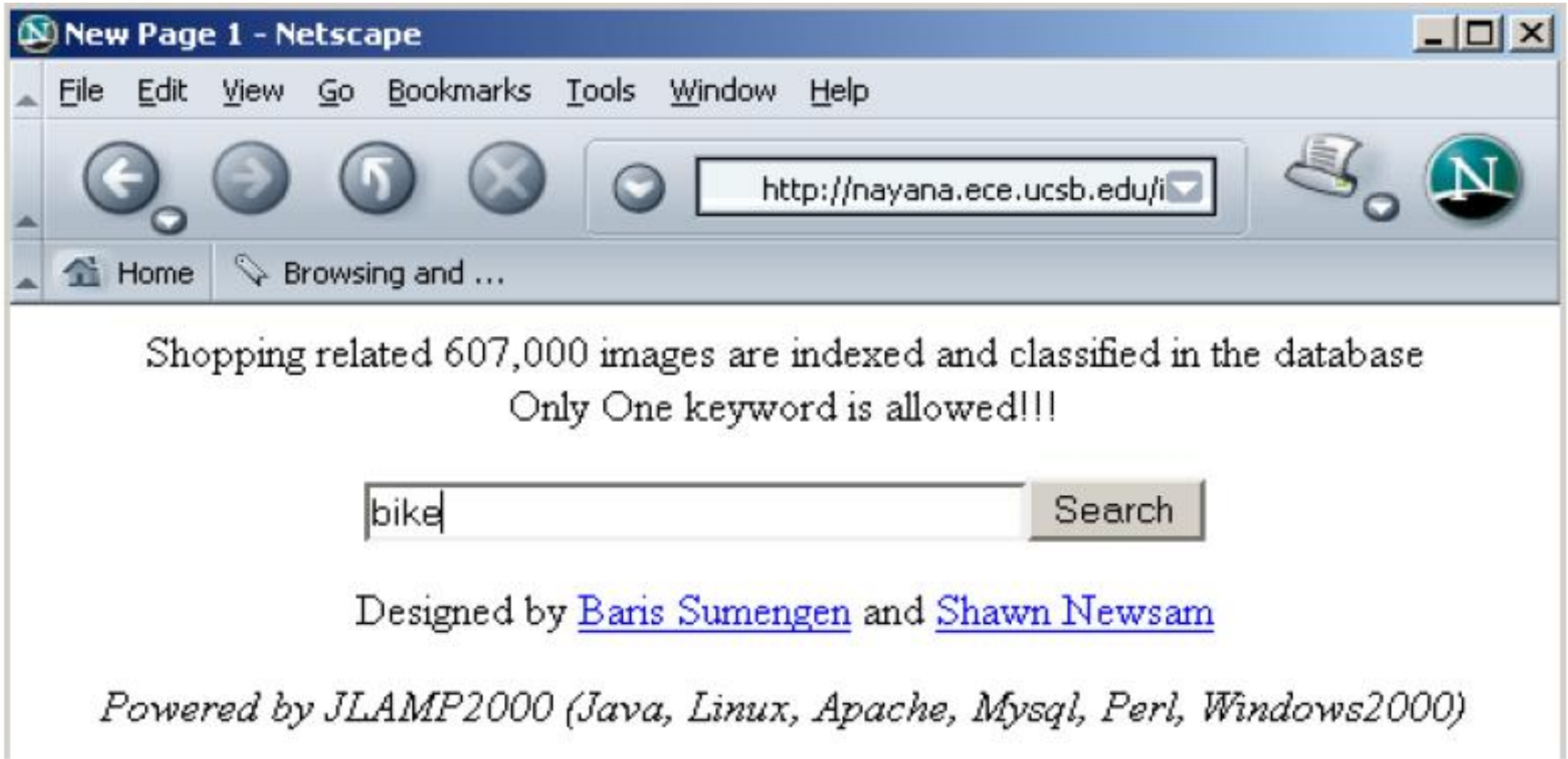
# 相关反馈

- 相关反馈可以循环若干次
- 下面将使用术语ad hoc retrieval来表示那种无相关反馈的常规检索
- 将介绍三个不同的(用户)相关反馈的例子

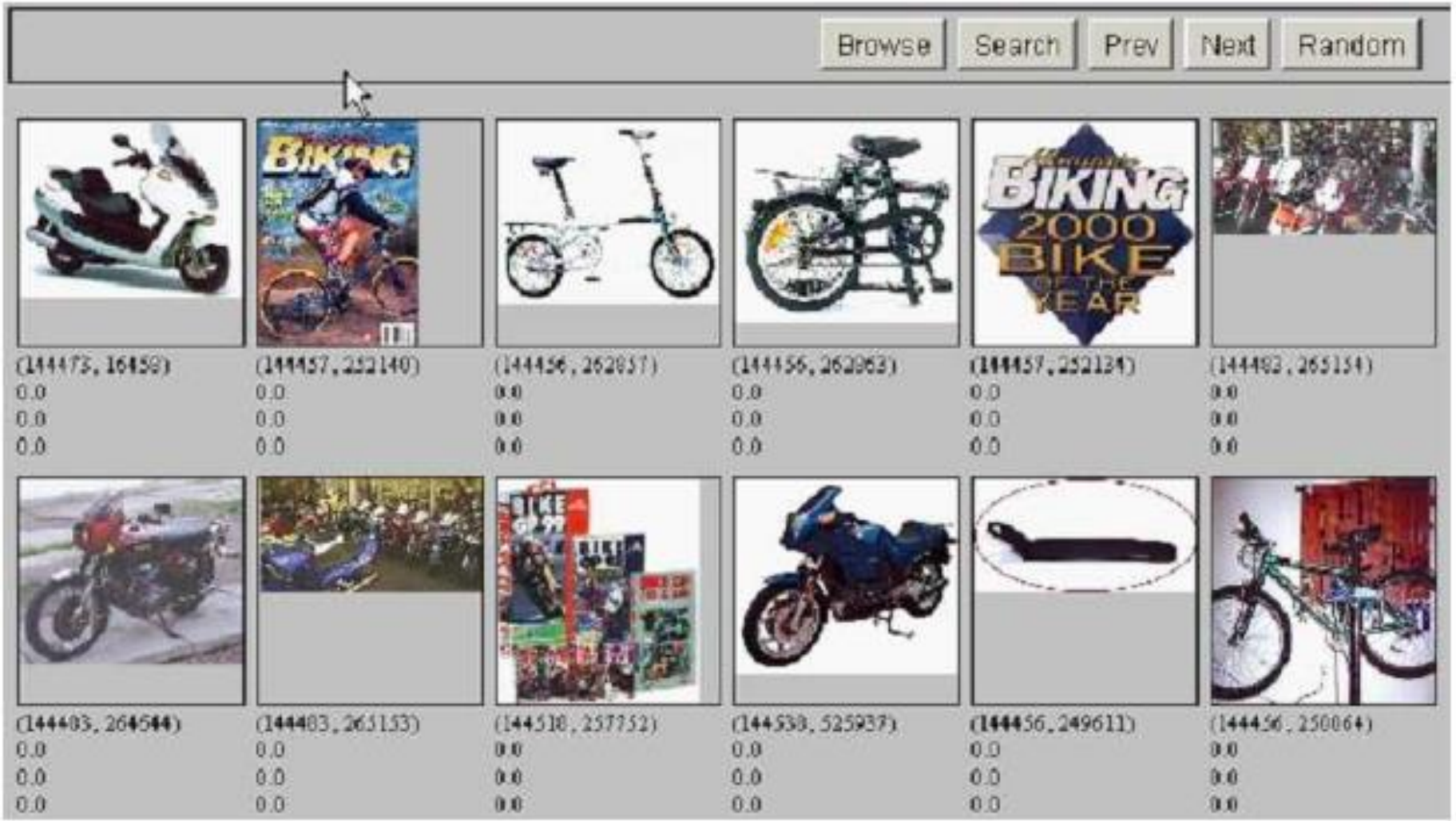
# 例1 类似页面


Google
sarah brightman
Search
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
老虎打成猫
老虎打成猫只为拍照取乐_虎园虐虎赚钱照曝光_华东在线
老虎打成猫只为拍照取乐 虎园虐虎赚钱照曝光，简述：为赚钱，他们愣是把老虎打成了猫！长春东北虎园把老虎绑在木板上，边抽打边强迫老虎与游客合影，还有幼童骑在虎背上拍...
www.cnhuadong.net/...-5-2/content_297613.shtml - 1天前 - 快照 - 预览
相关搜索
老虎追得猫上树
猫是老虎的师傅续写
老虎猫表情下载
老虎和猫的区别
老虎遇上猫_米璐璐
老虎向猫学艺童话故事
小孩老虎猫头鞋
老虎请猫
老虎和猫童话故事

# 例2 图像检索

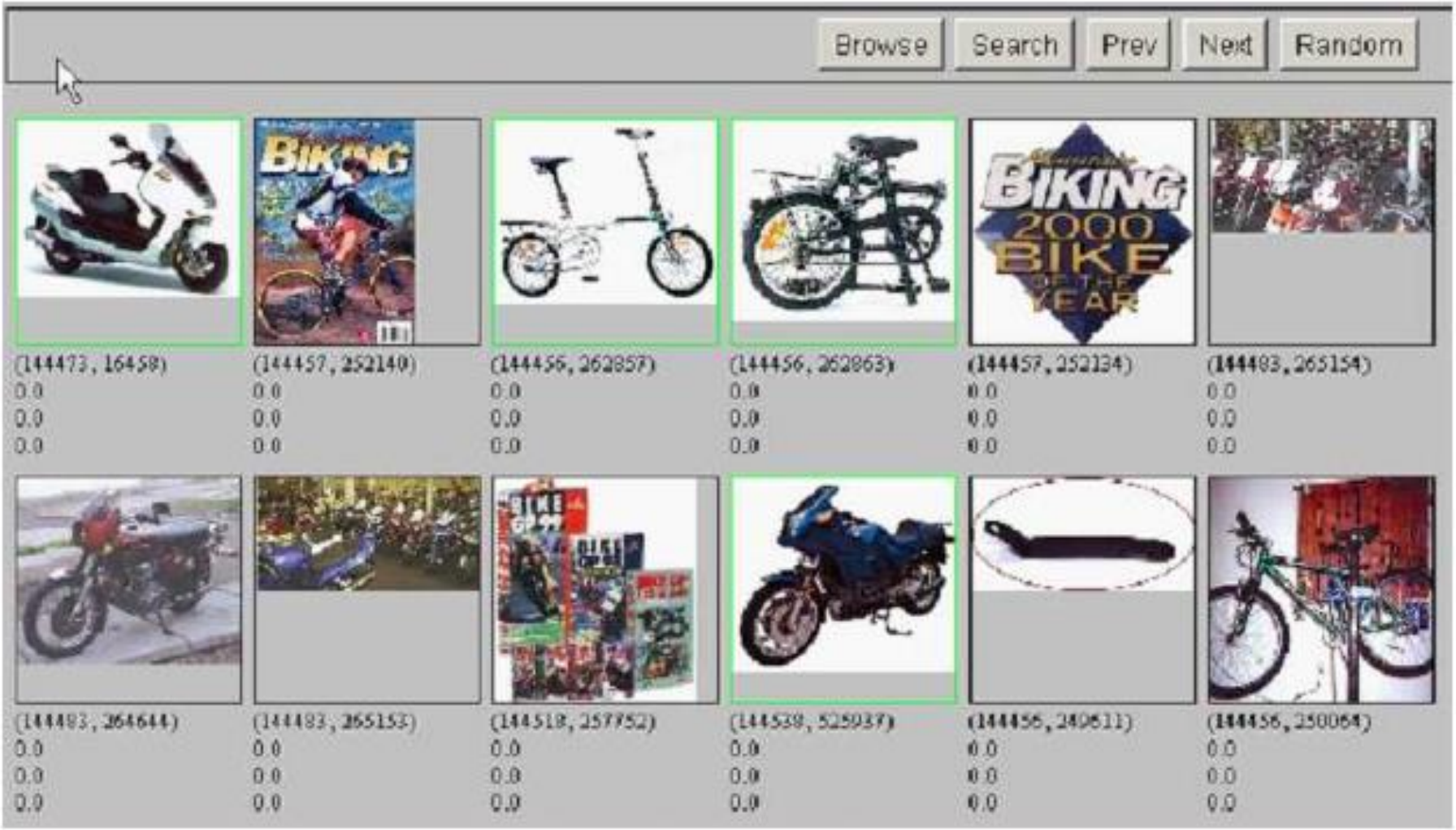

New Page 1 - Netscape
File Edit View Go Bookmarks Tools Window Help
http://nayana.ece.ucsb.edu/i
Home
Browsing and ...
Shopping related 607,000 images are indexed and classified in the database
Only One keyword is allowed!!!
bike
Search
Designed by Baris Sumengen and Shawn Newsam
Powered by JLAMP2000 (Java, Linux, Apache, Mysql, Perl, Windows2000)

# 初始查询的结果


Browse
Search
Prev
Next
Random
BIKING
BIKING 2000 BIKE OF THE YEAR

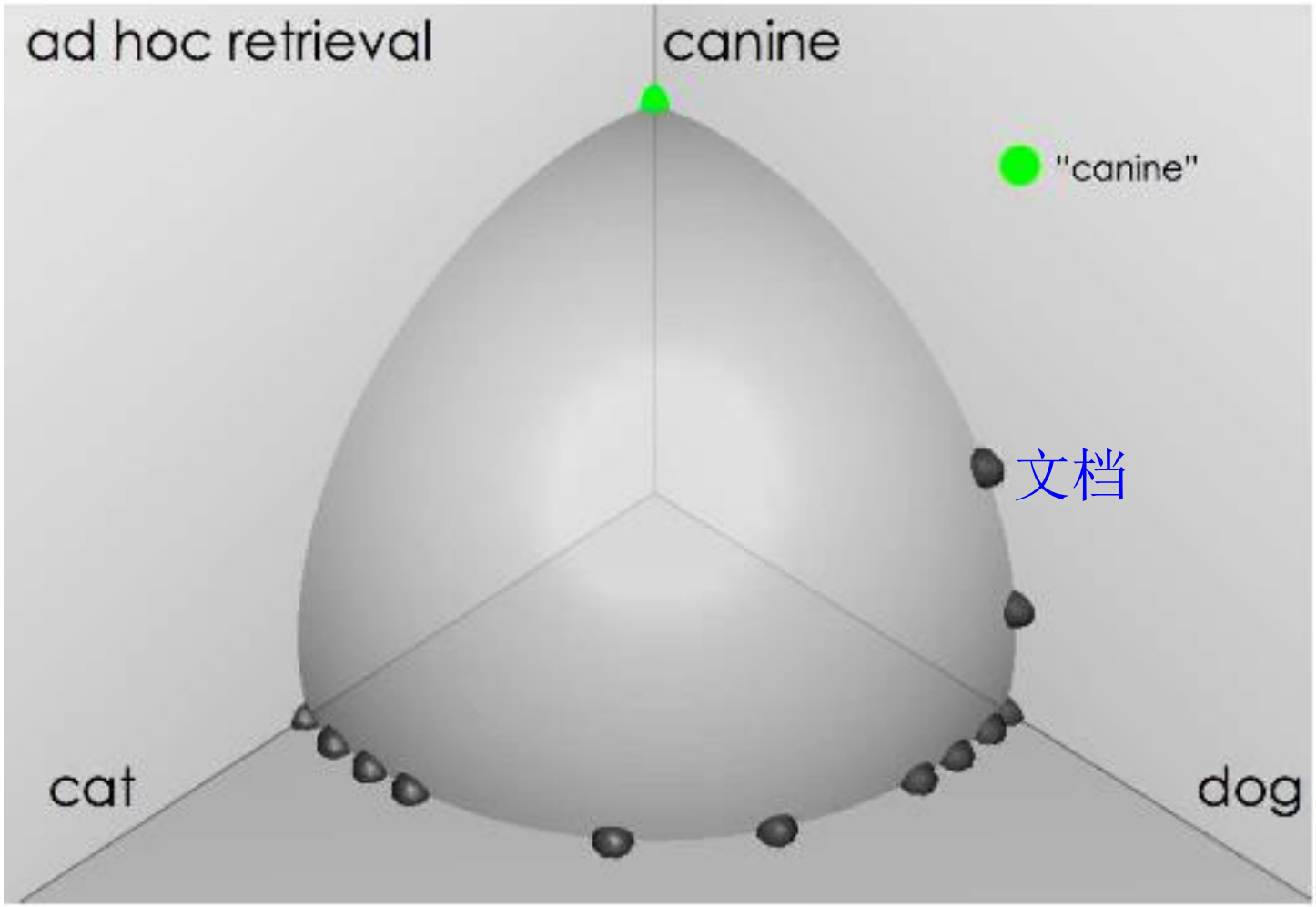
# 用户反馈: 选择相关结果

## 绿框表示用户认为相关的结果



# 相关反馈后再次检索的结果

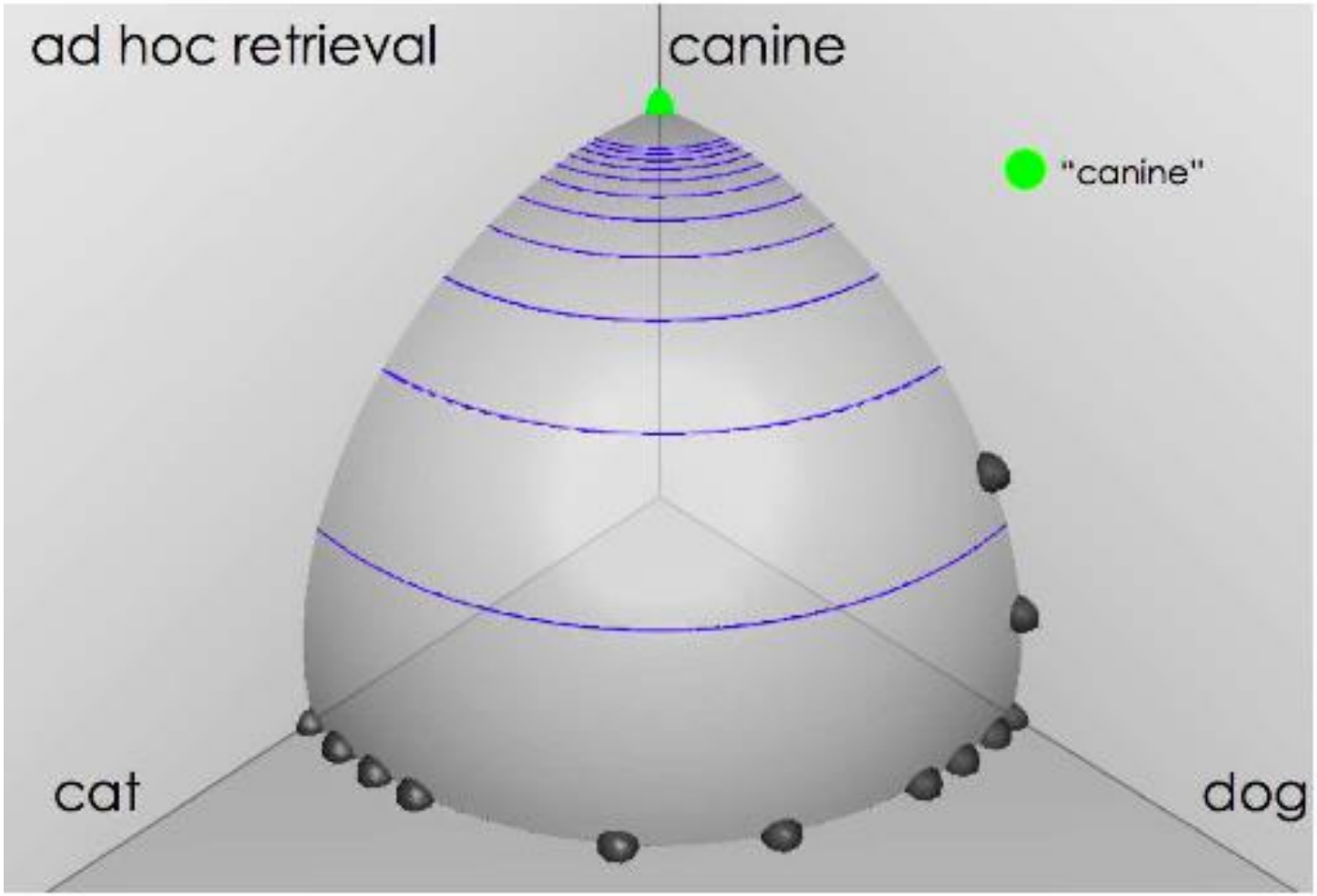
# 例3：向量空间: 查询 “canine” 似犬的




Source:

Fernando Díaz

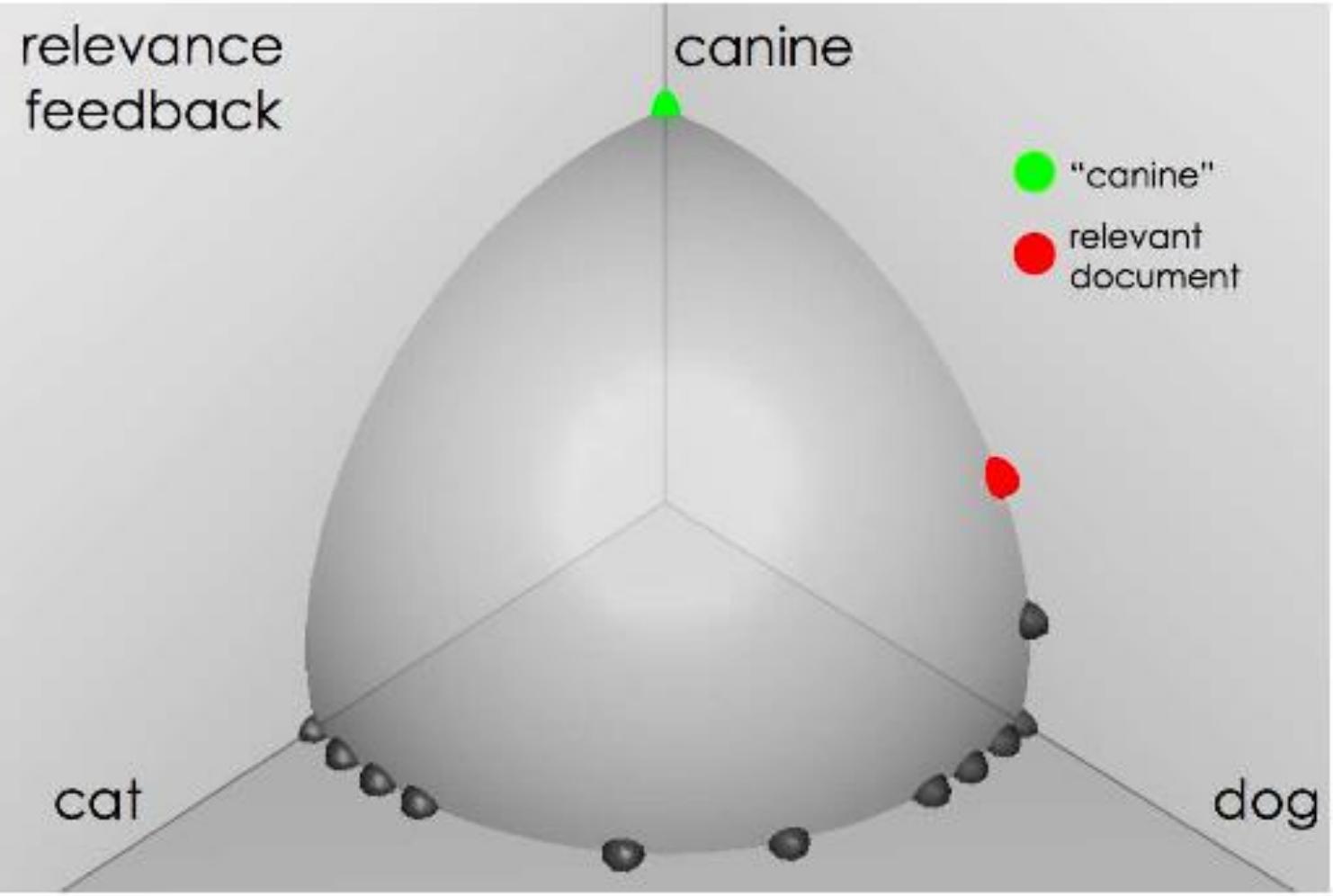
# 文档和查询 “canine” 的相似度




Source:

Fernando Díaz

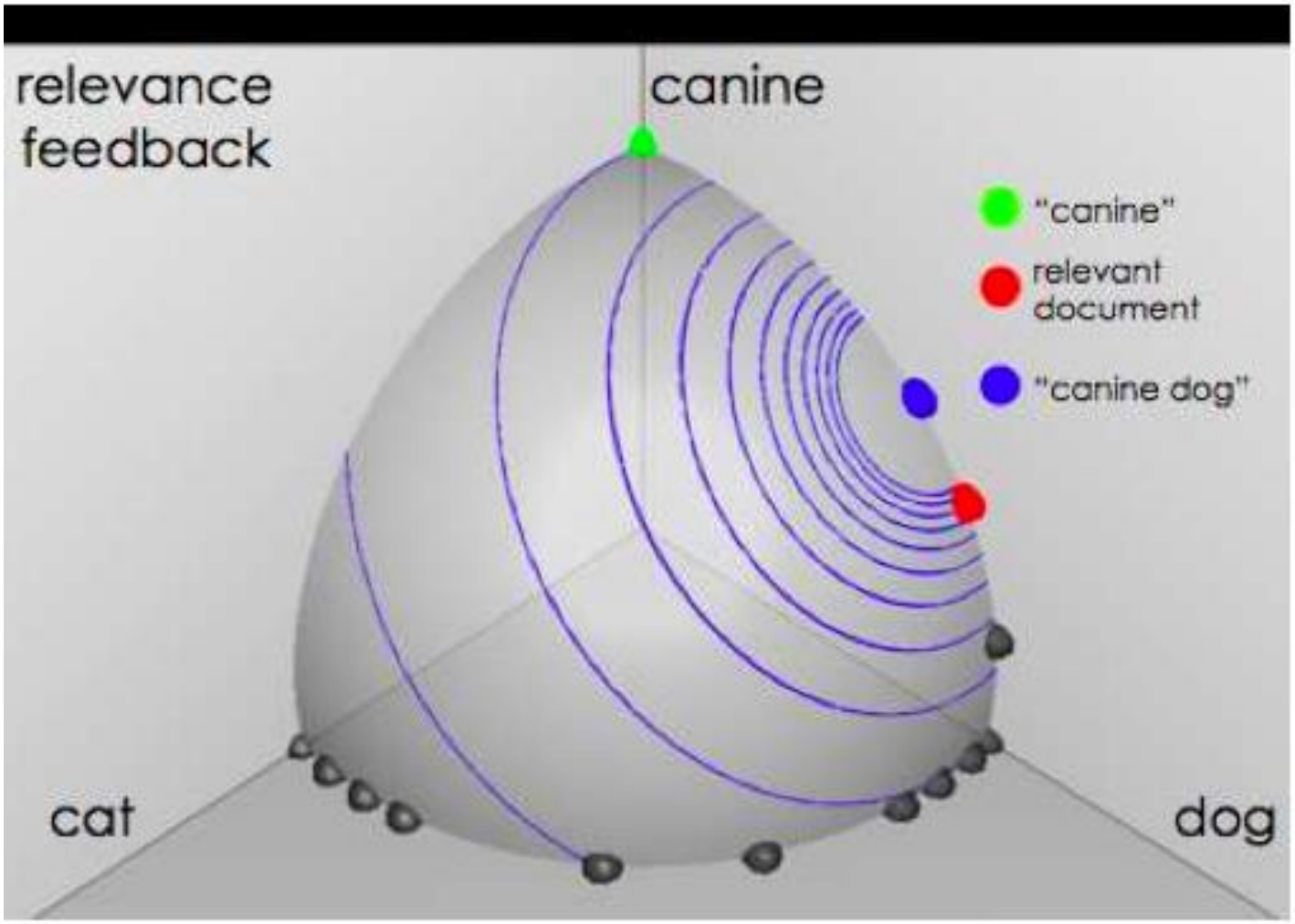
# 用户反馈: 选择一个认为相关的文档




Source:

Fernando Díaz

# 相关反馈后的检索结果




Source:

Fernando Díaz

# 例4: 一个实际的例子

初始查询:

[new space satellite applications] 初始查询的检索结果: (*r* = rank)

| | *r* | | |
|---|---|---|---|
| + | 1 | 0.539 | NASA Hasn't Scrapped Imaging Spectrometer |
| + | 2 | 0.533 | NASA Scratches Environment Gear From Satellite Plan |
| | 3 | 0.528 | Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes |
| | 4 | 0.526 | A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget |
| | 5 | 0.525 | Scientist Who Exposed Global Warming Proposes Satellites for Climate Research |
| | 6 | 0.524 | Report Provides Support for the Critics Of Using Big Satellites to Study Climate |
| | 7 | 0.516 | Arianespace Receives Satellite Launch Pact From Telesat Canada |
| + | 8 | 0.509 | Telecommunications Tale of Two Companies |

用户将一些文档标记为相关 "+".

# 基于相关反馈进行扩展后的查询

| 权重 | 词 | 权重 | 词 |
|---|---|---|---|
| 2.074 | new | 15.106 | space |
| 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 |

查询: [new space satellite applications]

# 基于扩展查询的检索结果

| | *r* | | |
|---|---|---|---|
| * | 1 | 0.513 | NASA Scratches Environment Gear From Satellite Plan |
| * | 2 | 0.500 | NASA Hasn't Scrapped Imaging Spectrometer |
| | 3 | 0.493 | When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own |
| | 4 | 0.493 | NASA Uses 'Warm' Superconductors For Fast Circuit |
| * | 5 | 0.492 | Telecommunications Tale of Two Companies |
| | 6 | 0.491 | Soviets May Adapt Parts of SS-20 Missile For Commercial Use |
| | 7 | 0.490 | Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers |
| | 8 | 0.490 | Rescue of Satellite By Space Agency To Cost $90 Million |

# 目录

- 动机
- 相关反馈基础
- 相关反馈详细介绍
- 查询扩展

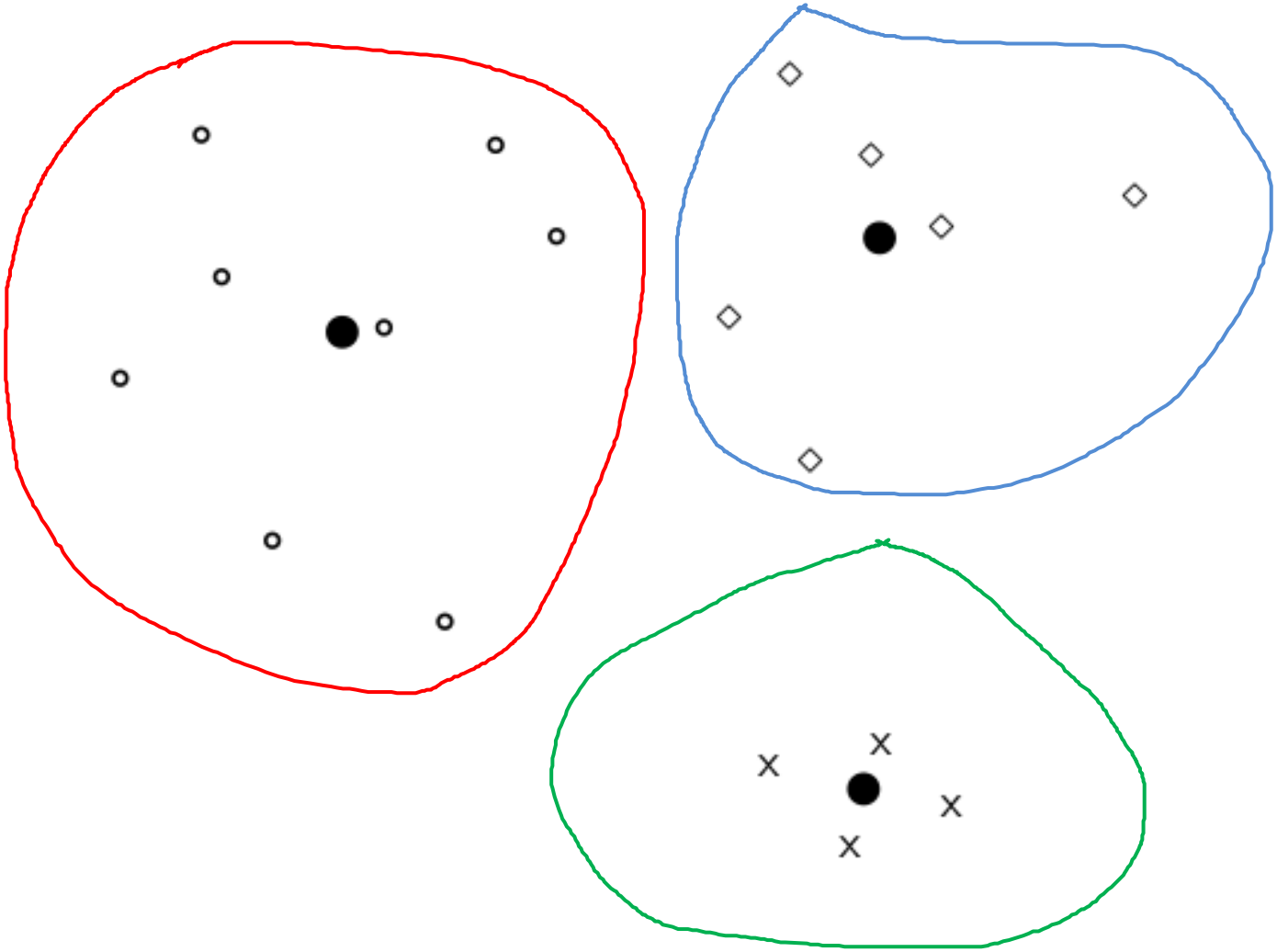
# 相关反馈中的核心概念：质心

- 质心是一系列点的中心
- 前面将文档表示成高维空间中的点
- 因此，可以采用如下方式计算文档的质心

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

其中$C$是一个文档集合，$\vec{d}$ 是文档$d$的向量表示

# 质心的例子



# 相关反馈基本理论

- 基本理论:假定要找一个最优查询向量$\vec{q}$，它与相关文档之间的相似度最大且同时又和不相关文档之间的相似度最小。
- 最优的$\vec{q}$是使下式最大的查询$\vec{q}_{opt}$:

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

$C_r$表示相关文档集，$C_{nr}$表示不相关文档集
$\vec{\mu}(C_r)$ 表示质心

- 上述公式的意图是 $\vec{q}_{opt}$ 是将相关文档与不相关文档区分开的向量

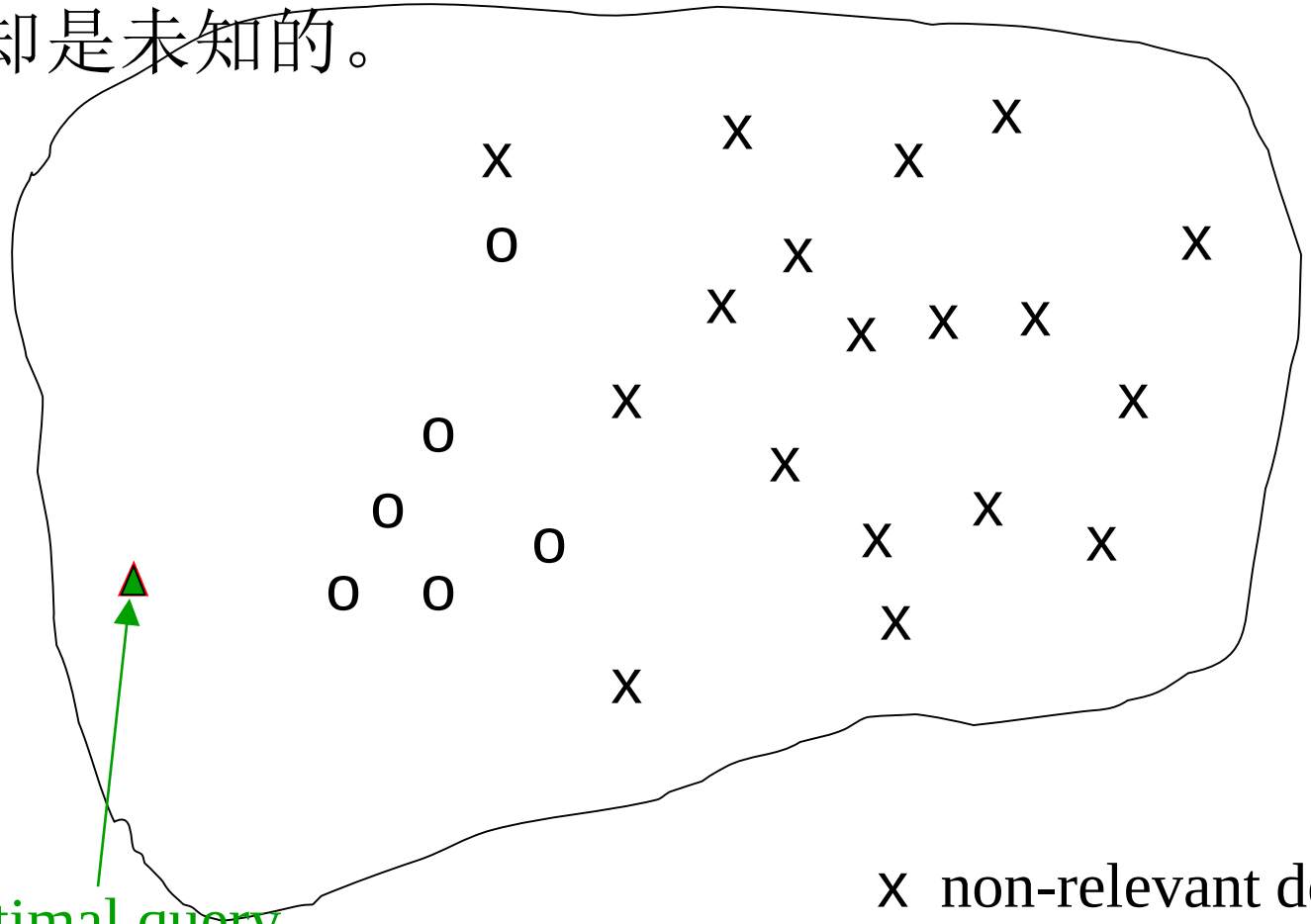
- 当*sim*()函数采用余弦相似度计算时，能够将相关文档与不相关文档区分开的最优查询向量为

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

- 这就是说，最优的查询向量等于相关文档的质心向量和不相关文档的质心向量的差

# The Theoretically Best Query

- 然而，这个发现并没有什么意义，因为检索本来的目的就是要找相关文档，而所有的相关文档集事先却是未知的。



# Rocchio算法

- 加入一些额外的假设，可以将上式改写为:

$$\vec{q}_{opt} = \mu(D_r) + [\mu(D_r) - \mu(D_{nr})]$$

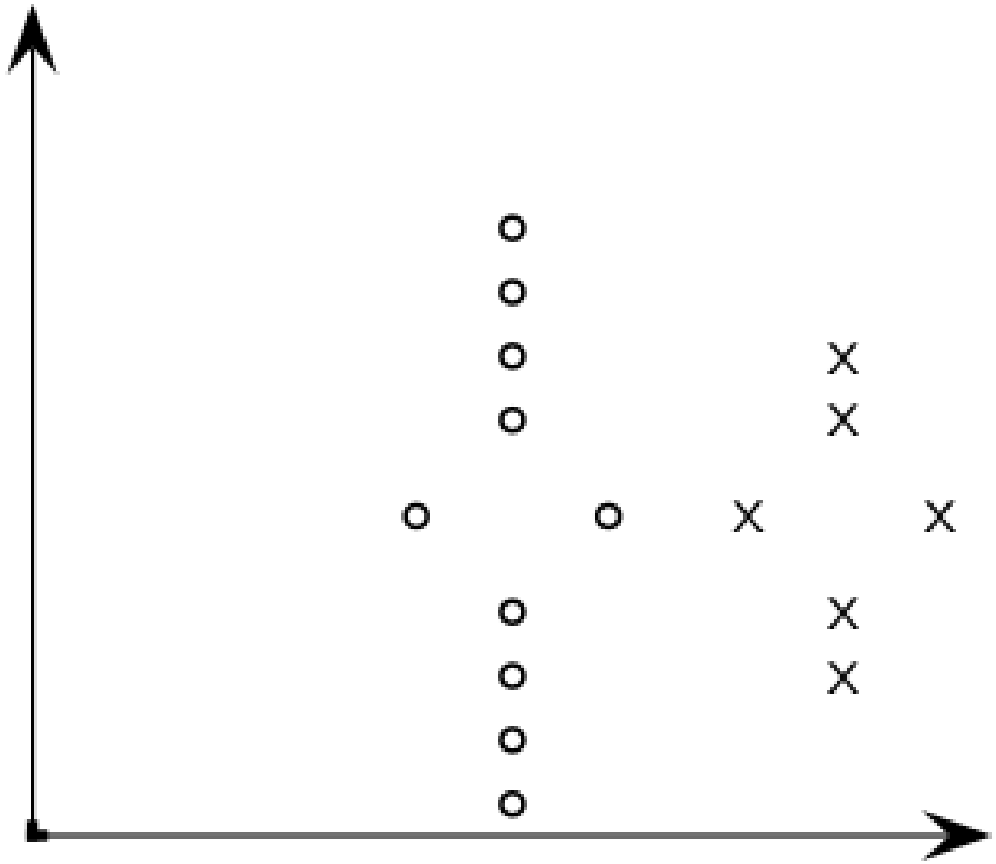
- 最优查询向量为：

$$\vec{q}_{opt} = \mu(D_r) + [\mu(D_r) - \mu(D_{nr})]$$

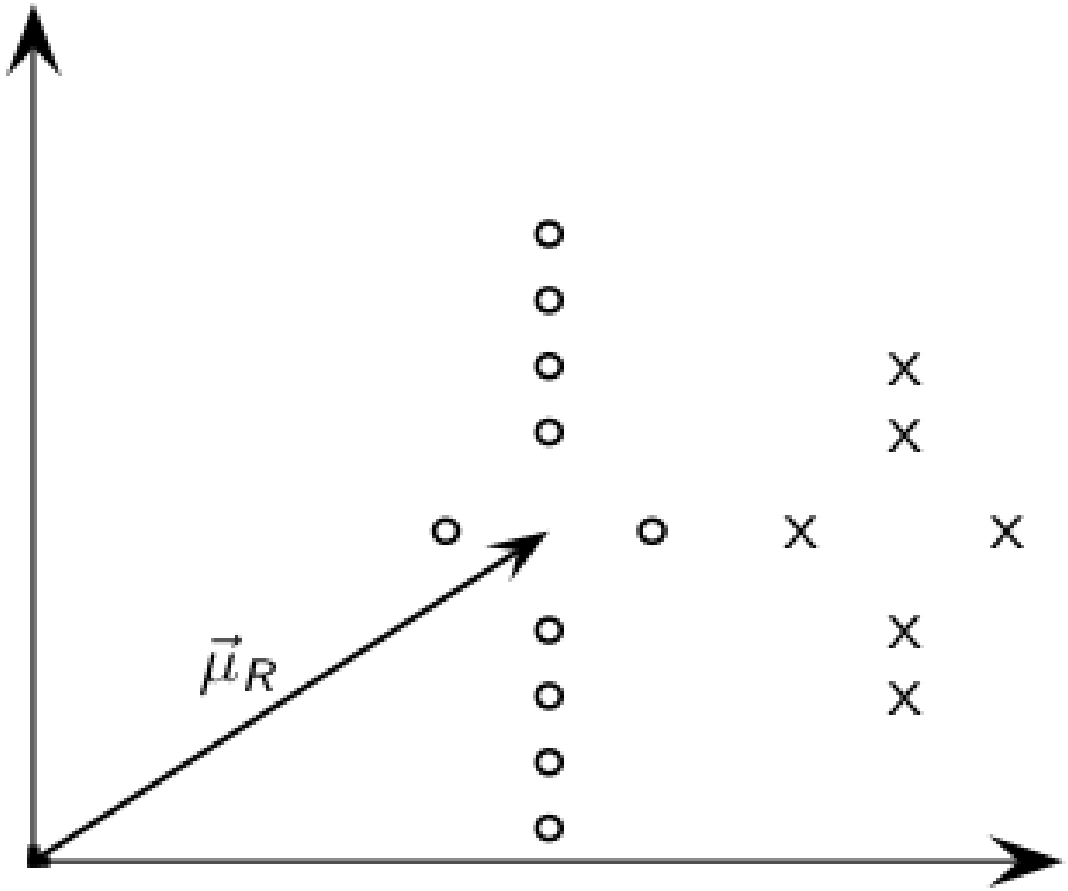
$$= \frac{1}{|D_r|} \sum_{\vec{d}_i \in D_r} \vec{d}_i + [\frac{1}{|D_r|} \sum_{\vec{d}_i \in D_r} \vec{d}_i - \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j]$$

- $D_r$指已知的部分相关文档
- 即将相关文档的质心移动一个量，该量为相关文档质心和不相关文档质心的差异量

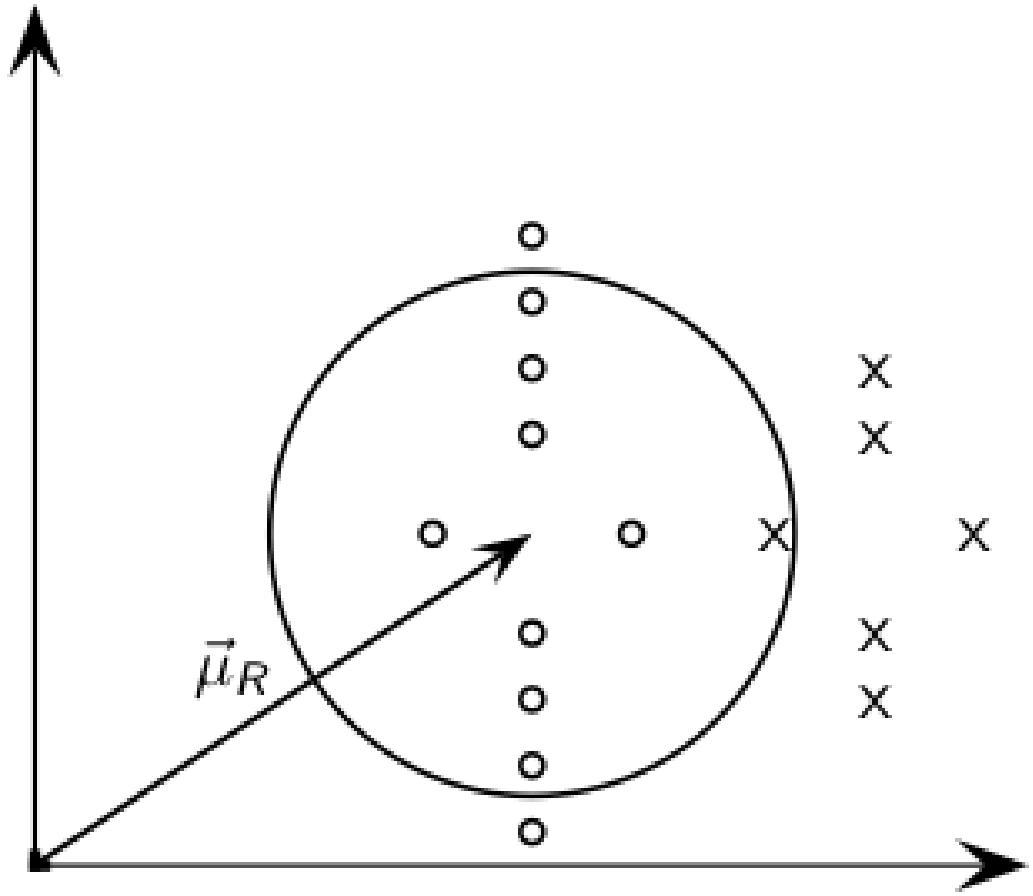
# 计算Rocchio向量



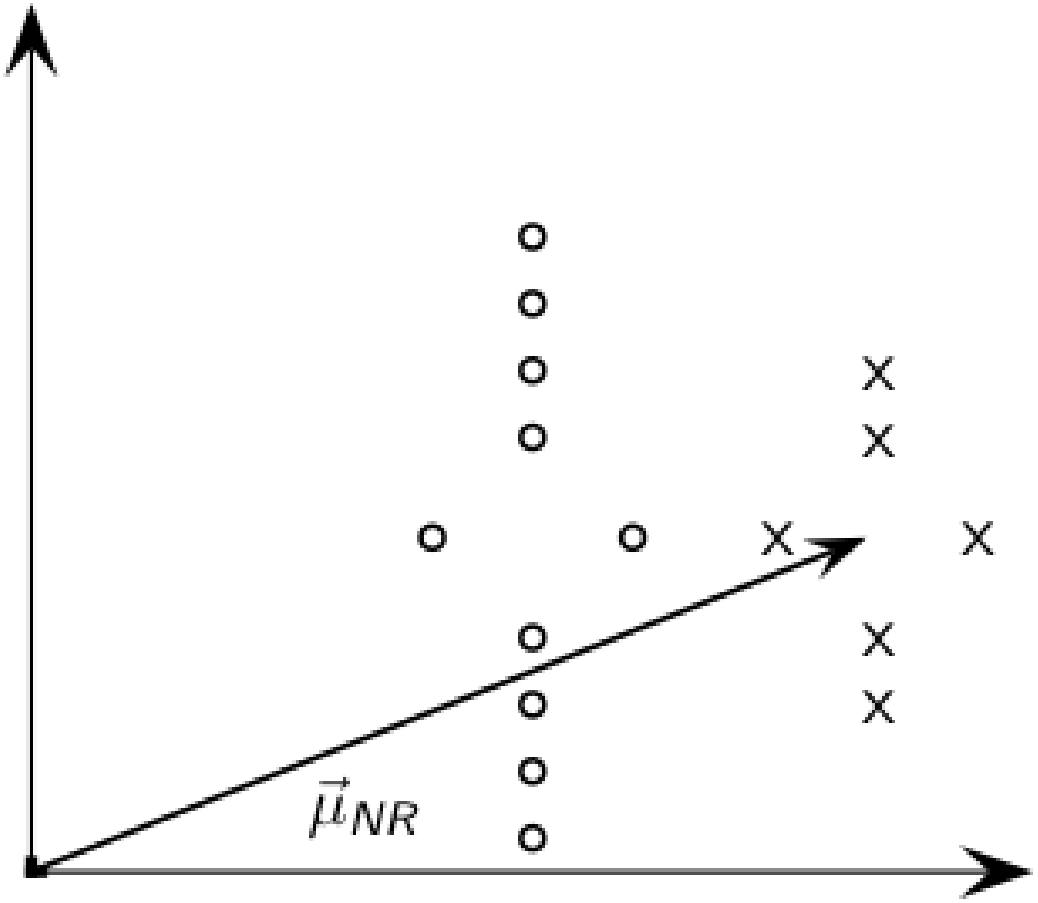
圆形点○: 已知的相关文档
叉叉点×: 已知的不相关文档



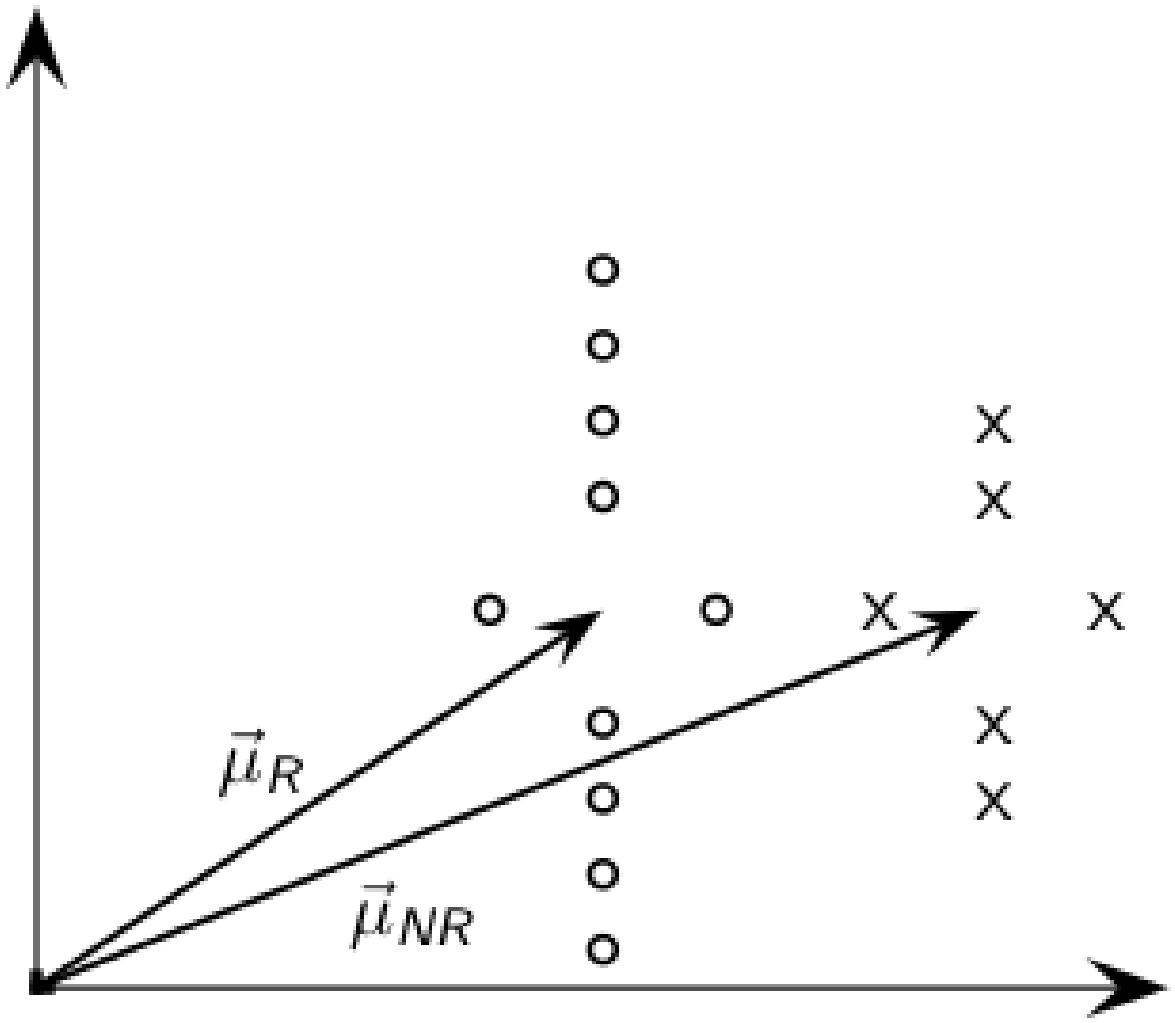

$\vec{\mu}_R$ : 相关文档的质心

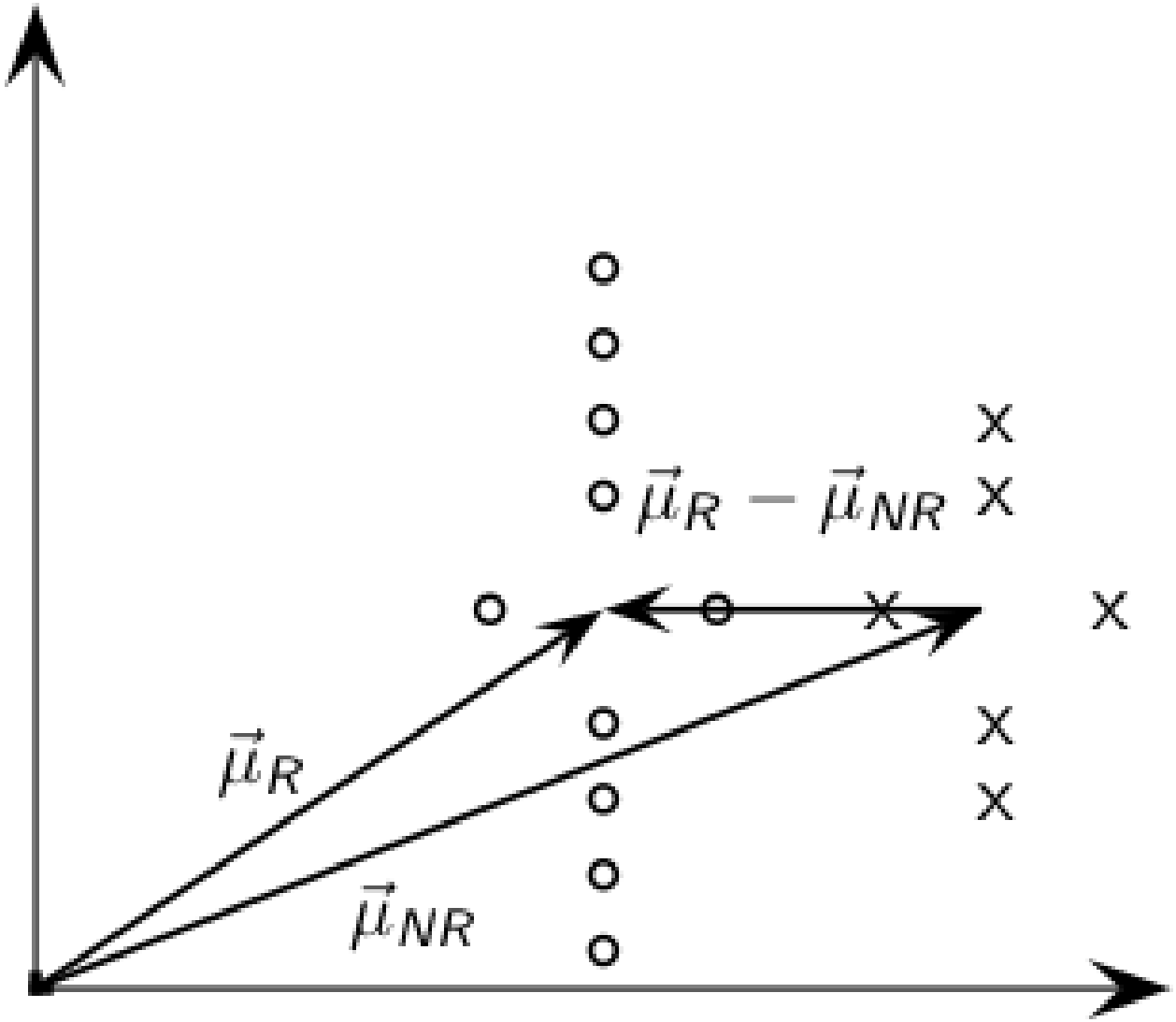



$\vec{\mu}_R$ 不能将相关/不相关文档分开

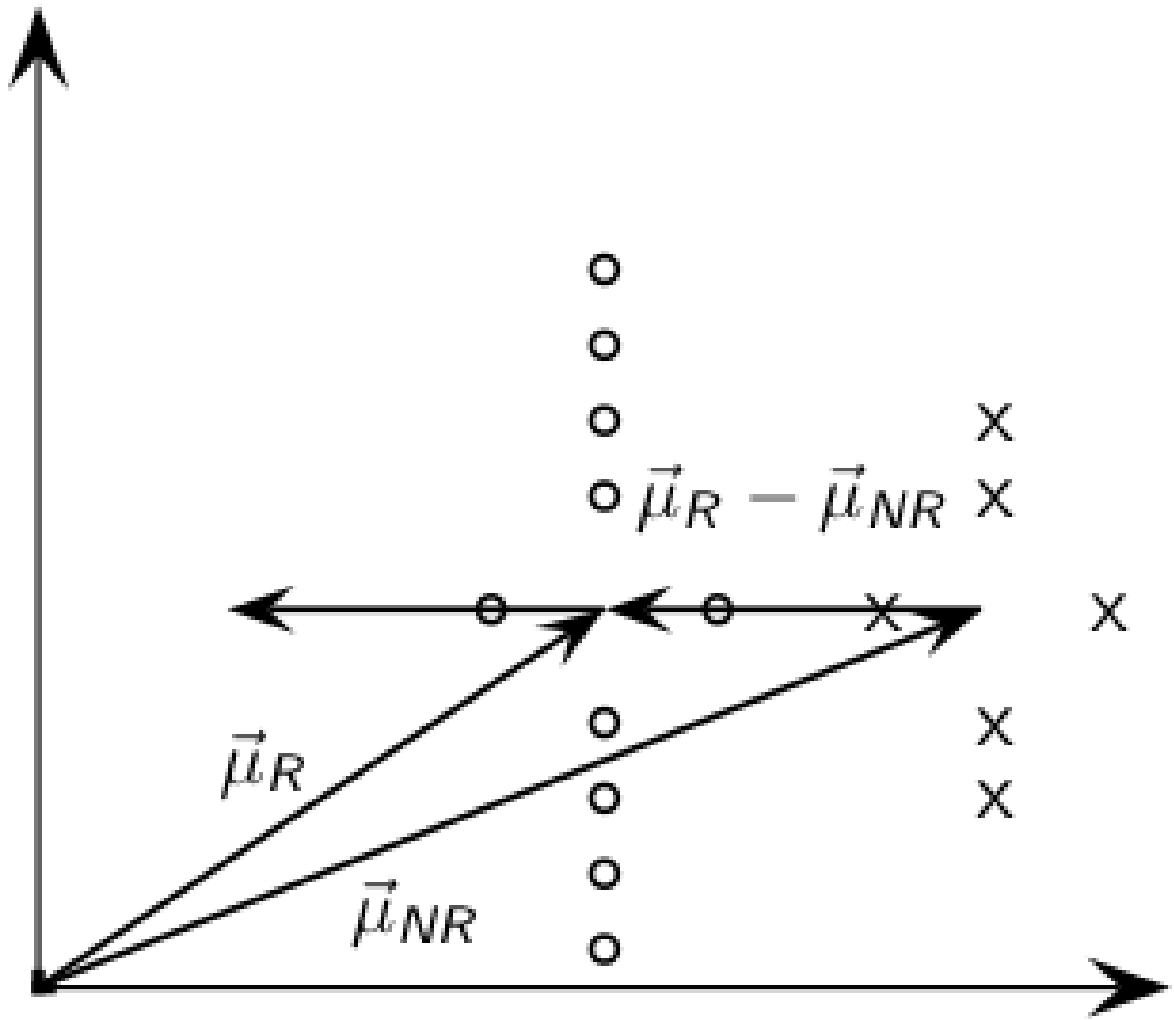



$\vec{\mu}_{NR}$: 不相关文档的质心


$\vec{\mu}_R$
$\vec{\mu}_{NR}$

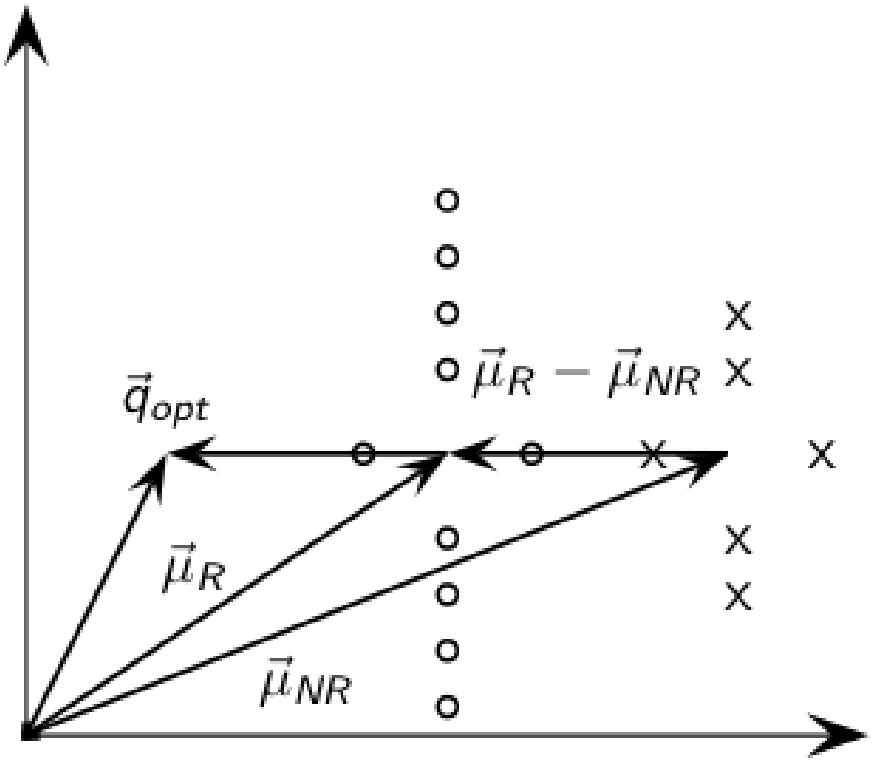



$\vec{\mu}_R$ - $\vec{\mu}_{NR}$: 差异向量

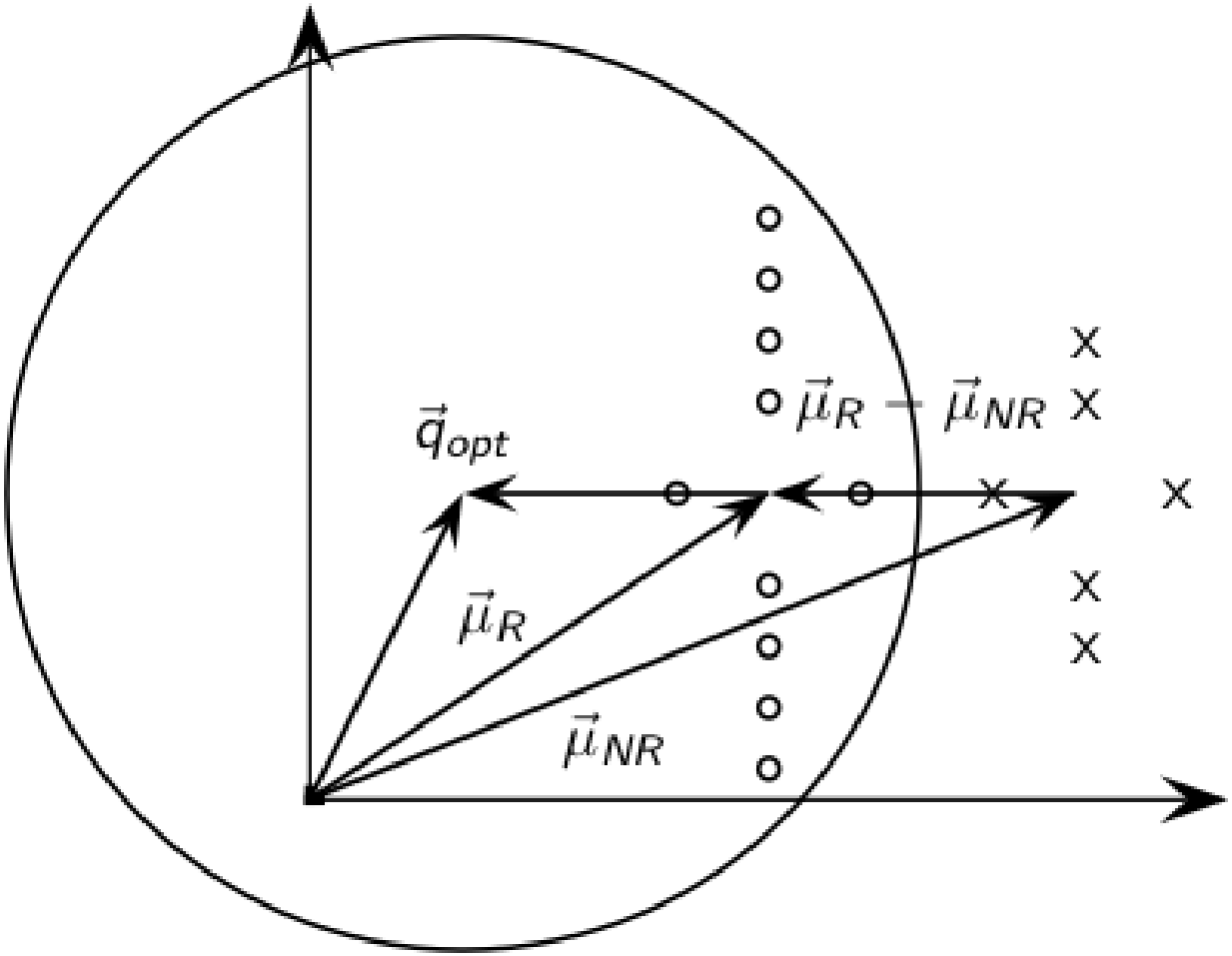



$\vec{\mu}_R$ 加上差异向量

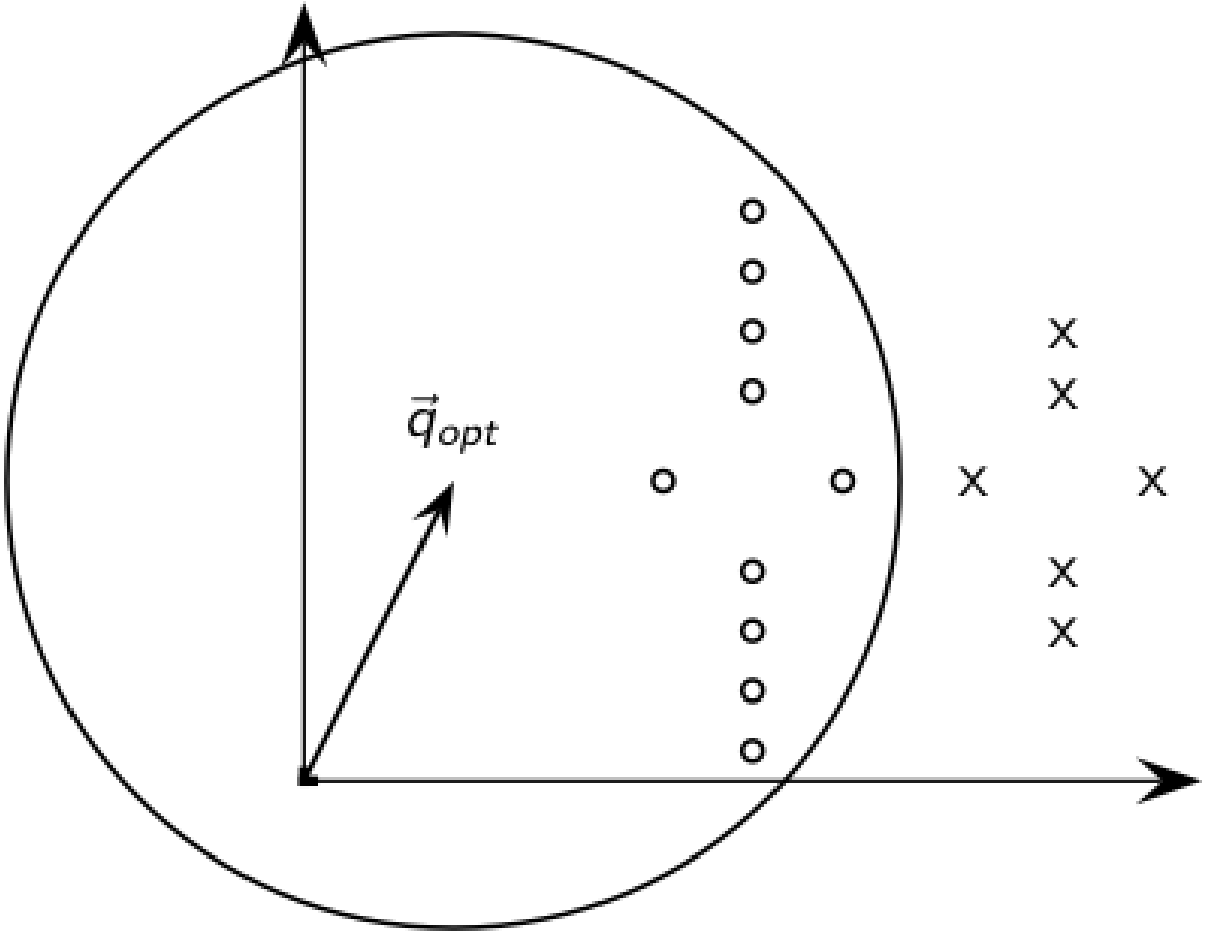
# Rocchio算法图示




得到 $\vec{q}_{opt}$




$\vec{q}_{opt}$ 能够将相关/不相关文档完美地分开




$\vec{q}_{opt}$ 能够将相关/不相关文档完美地分开

# Rocchio 1971 算法 (SMART系统使用)

- 假定有一个用户查询，并知道部分相关文档$D_r$和不相关文档$D_{nr}$的信息，最优查询向量为：

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

$\vec{q}_m$：修改后的查询；$\vec{q}_0$：原始查询；

$D_r$、$D_{nr}$：已知的相关和不相关文档集合

$\alpha, \beta, \gamma$: 权重

- 修改后的新查询从$\vec{q}_0$开始，向着相关文档的质心向量靠近了一段距离，而同时又与不相关文档的质心向量远离了一段距离。
- 即新查询向相关文档靠拢而远离非相关文档

# 正(Positive)反馈 vs. 负(Negative)反馈

- 正反馈价值往往大于负反馈
  - 比如，可以通过设置 $\beta = 0.75$，$\gamma = 0.25$来给正反馈更大的权重
- 很多系统甚至只允许正反馈，即$\gamma=0$

# 相关反馈中的假设

- 假设1：用户对于初始查询有充分的认识，知道使用哪些词项来表达
- 假设2：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
  - 不相关文档的词项分布和相关文档的词项分布不相似
    - 所有文档都紧密聚集在某个prototype周围，形成一个簇
    - 或者：有多个不同的prototype，但是它们之间的词汇具有显著的重合率
    - 相关文档和不相关文档之间的相似度很低

# 不满足假设1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

# 不满足假设2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

# 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

- 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。

# 相关反馈存在的问题

- 开销很大
  - 生成的新查询往往很长
  - 长查询的处理开销很大
- 用户不愿意提供显式的相关反馈
- 很难理解，为什么会返回(应用相关反馈之后)某篇特定文档
- Excite搜索引擎曾经提供完整的相关反馈功能，但是后来废弃了这一功能

# 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，但是省去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。

# 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

# 用户行为种类

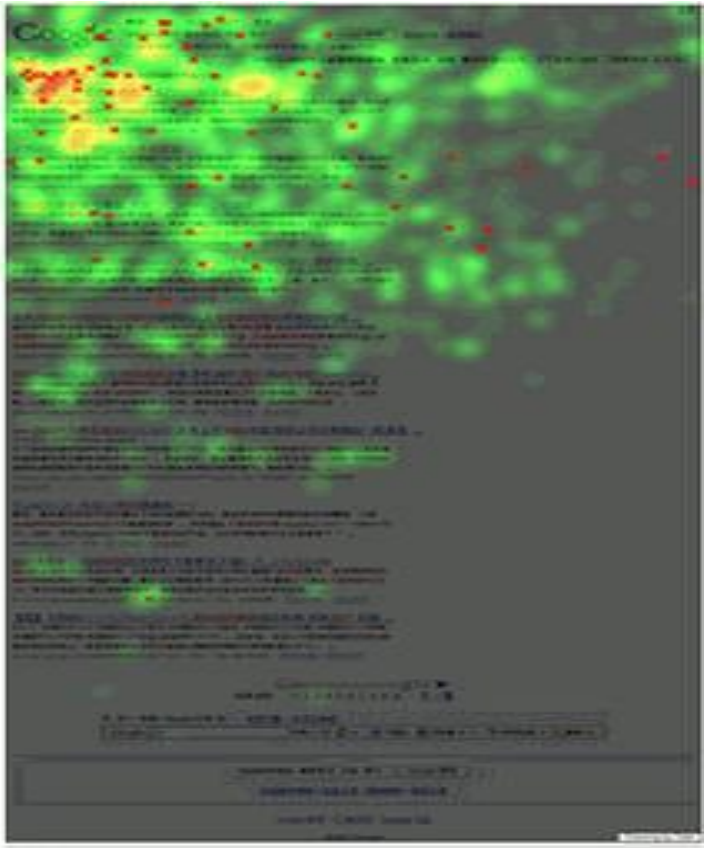
- 鼠标键盘动作
  - 点击链接、加入收藏夹、拷贝粘贴、停留、翻页等等
- 用户眼球动作
  - Eye tracking可以跟踪用户的眼球动作
  - 拉近、拉远、瞟、凝视、往某个方向转

# 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 | http://bbs.cixi.cn/dispbbs.asp?Star=4&boardid=46&id=346721&page=1 |
| 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] |

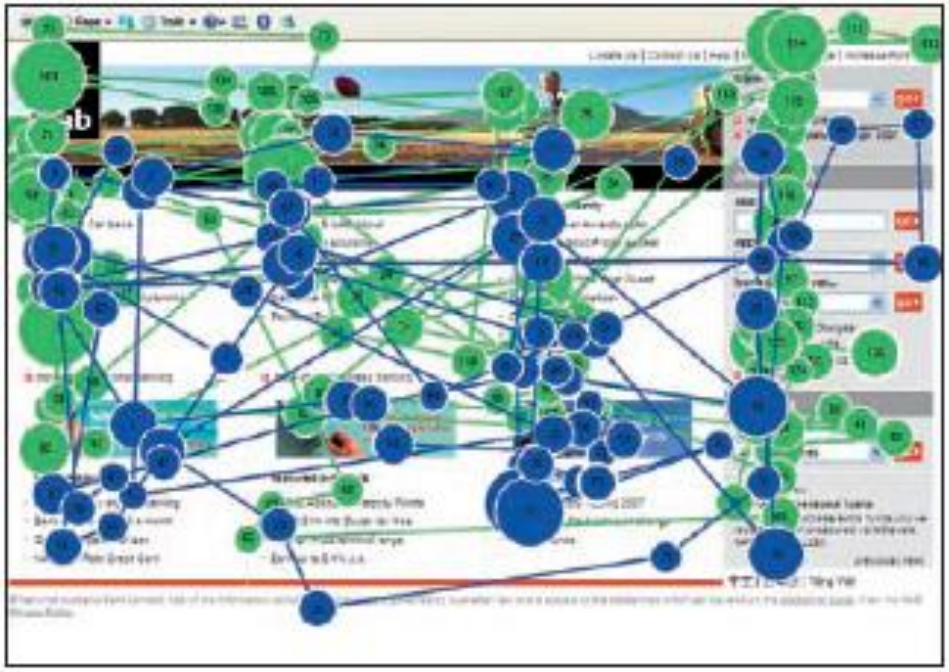
# 眼球动作(通过鼠标轨迹模拟)

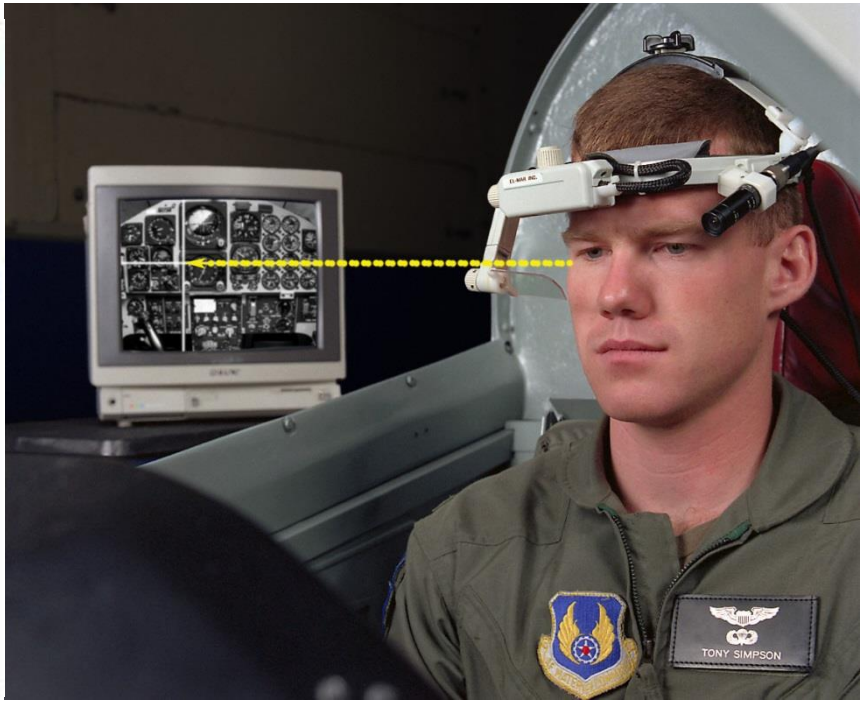
Baidu



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(Pseudo-relevance feedback)

- 伪相关反馈对于真实相关反馈的人工部分进行自动化
- 伪相关反馈算法
  - 对于用户查询返回有序的检索结果
  - 假定前 $k$ 篇文档是相关的
  - 进行相关反馈 (如 Rocchio)
- 平均上效果不错
- 但是对于某些查询而言可能结果很差
- 几次循环之后可能会导致查询漂移(*query drift*)

# 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

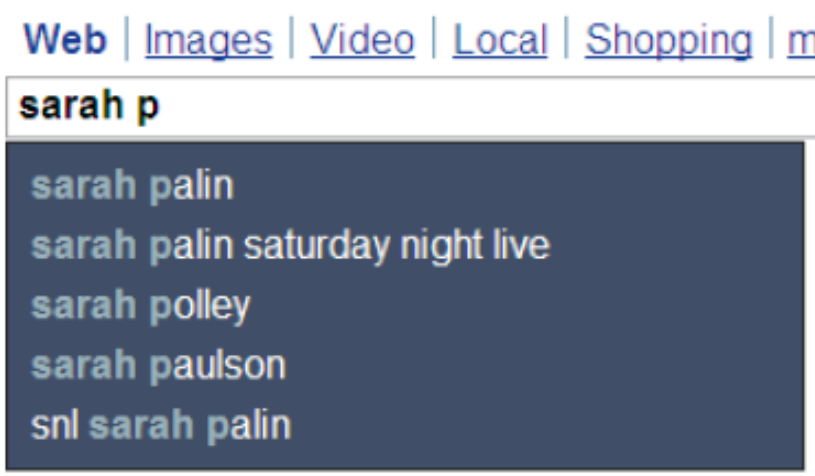
# 目录

- 动机
- 相关反馈基础
- 相关反馈详细介绍
- 查询扩展

# 查询扩展(Query expansion)

- 查询扩展是另一种提高召回率的方法
- 使用“全局查询扩展”来指那些“查询重构(query reformulation)的全局方法”
- 在全局查询扩展中，查询基于一些全局的资源进行修改，这些资源是与查询无关的
- 主要使用的信息: 同义词或近义词
  - 同义词或近义词词典(thesaurus)
  - 两种同(近)义词词典构建方法：人工构建和自动构建

# 查询扩展的例子


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


Sogou搜狗
雅安
搜狗搜索
雅安地震
雅安捐款
雅安地震最新消息
雅安地震图片
雅安地震捐款
雅安地震死亡人数
雅安地图
雅安加油
雅安地震视频
雅安余震
全部时间
一天内
一周内
一月内
一年内
您是不是要找
eye tracking inc
tracking number
17tracking
cosco tracking
A decade ago test shoppers were tethered to bulky computer equipment pushed along behind t
hem in a trolley. Today, wireless goggles do the job. As the cost of eye-tracking gear has...
经济学人 - www.economist.com/...n-persons-gaze - 2013-3-13 - 快照 - 预览

# 用户反馈的类型

- 用户对文档提供反馈
  - 在相关反馈中更普遍
- 用户对词或短语提供反馈
  - 在查询扩展中更普遍

# 查询扩展的类型

- 基于同义词词典
  - 人工构建的同(近)义词词典（人工编辑维护的词典，如 PubMed)
  - 自动导出的同(近)义词词典（比如，基于词语的共现统计信息)
- 基于查询日志挖掘出的查询等价类 (Web上很普遍，比如上面的 “palm” 例子)

# 基于同(近)义词词典的查询扩展

- 对查询中的每个词项*t*，将词典中与*t*语义相关的词扩充到查询中
  - 例子: HOSPITAL → MEDICAL
- 通常会提高召回率
- 可能会显著降低正确率，特别是对那些有歧义的词项
INTEREST RATE → INTEREST RATE FASCINATE
- 广泛应用于特定领域(如科学、工程领域)的搜索引擎中
- 创建并持续维护人工词典的开销非常大
- 人工词典和基于受控词汇表(controlled vocabulary)的标记的效果大体相当

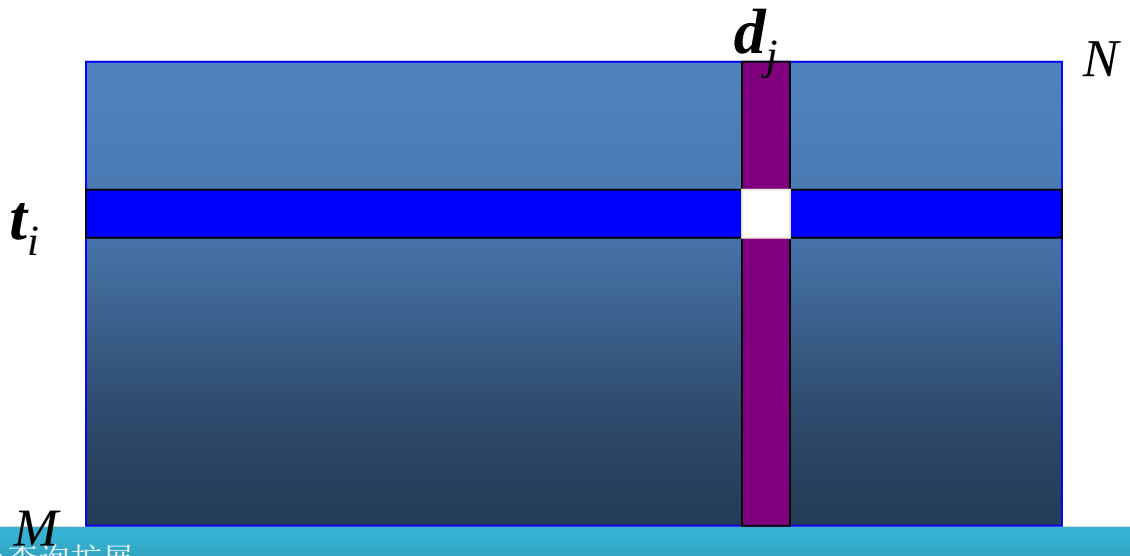
# 基于人工词典的扩展样例: PubMed

# 同(近)义词词典的自动构建

- 通过分析文档集中的词项分布来自动生成同(近)义词词典
- 基本想法：计算词语之间的相似度
  - **定义 1**：如果两个词各自的上下文共现词类似，那么它们类似
    - “car” ≈ “motorcycle”，因为它们都与“road”、“gas”及“license”之类的词共现，因此它们类似
  - **定义 2**：如果两个词同某些一样的词具有某种给定的语法关系的话，那么它们类似
    - 可以“harvest, peel, eat, prepare” apples 和pears, 因此apples 和pears肯定彼此类似
- 共现关系更加鲁棒，而语法关系更加精确

# 基于共现的词典构造

- 简单的方法：通过词典-文档矩阵$A$计算词项-词项的相似度
- 给定$A$，其中$A_{t,d} = (t,d)$ 词项$t$在文档$d$中的(归一化)权重
- 计算$C = AA^T$，其中元素$C_{uv}$表示词项$u$和词项$v$的相似度
- 对每个$t_i$，选择$C$中高权重的词项进行扩展



# 基于共现关系的同(近)义词词典样例

| 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 |

WordSpace demo on web

# 搜索引擎中的查询扩展

- 搜索引擎进行查询扩展主要依赖的资源：查询日志(query log)
  - 例 1: 提交查询 [herbs] (草药)后，用户常常搜索[herbal remedies] (草本疗法)
    - → “herbal remedies” 是 “herb”的潜在扩展查询
  - 例 2: 用户搜索 [flower pix] 时常常点击URL photobucket.com/flower，而用户搜索[flower clipart]常常点击同样的URL
    - → “flower clipart”和“flower pix” 可能互为扩展查询

# 本讲小结

- 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
- 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。