

信息组织与检索

第9讲：相关反馈和查询扩展

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 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
- 信息检索评测语料及会议
- 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, tp)	伪正例(false positives, fp)
未返回(not retrieved)	伪反例(false negatives, fn)	真反例(true negatives, tn)

$P = TP / (TP + FP)$ 系统返回结果里面有多少比例是正确的

$R = TP / (TP + FN)$ 文档库里面有多少比例相关文档被找到

正确率和召回率相结合的指标： F 值

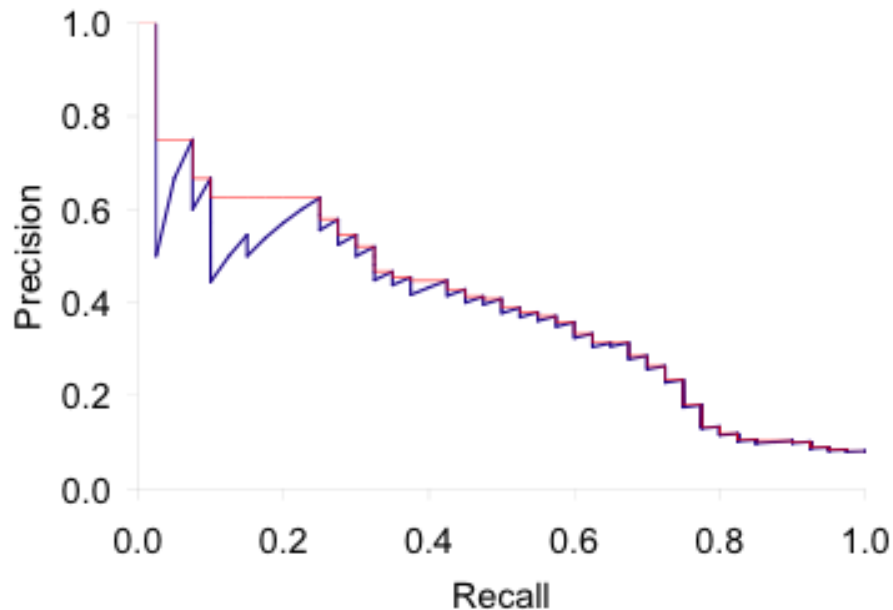
- F 允许正确率和召回率的折中

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

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

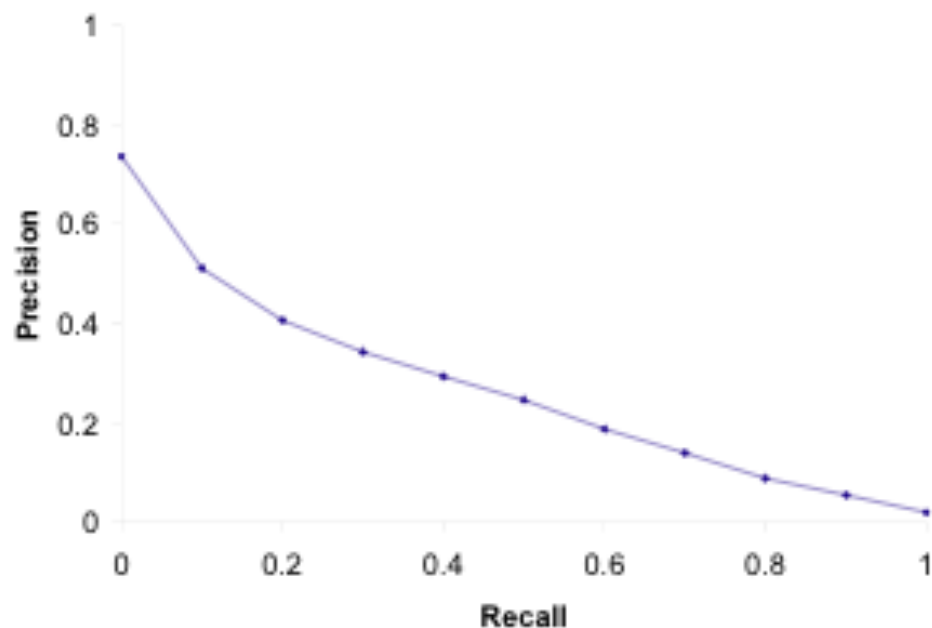
$$\frac{1}{F} = \frac{1}{2} \left(\frac{1}{P} + \frac{1}{R} \right)$$

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

标准的评价会议: TREC

- TREC = Text Retrieval Conference (TREC)
- 美国标准技术研究所 (NIST) 组织
- TREC 实际上包含了对多个任务的评测
- 最出名的任务: TREC Ad Hoc 任务, 1992 到 1999 年前 8 届会议中的标准任务
- TREC disk 包含 189 百万 篇文档, 主要是新闻报道, 有 450 个信息需求
- 由于人工标注的代价太大, 所有没有完整的相关性判定
- 然而, NIST 采用了一种所谓 结果缓冲(pooling) 的办法来进行人工标注, 首先将所有参测系统的前 k 个结果放到一个缓冲池(pool), 然后仅对缓冲池的文档进行标注, 并认为所有的相关文档均来自该缓冲池中。

动态摘要

- 给出一个或者多个“窗口”内的结果(snippet)，这些窗口包含了查询词项的多次出现
- 出现查询短语的snippet优先
- 在一个小窗口内出现查询词项的snippet优先
- 最终将所有snippet都显示出来作为摘要

本讲内容

- 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。

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] ... 或者 [电脑]
- 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。

计算机_百度搜索

https://www.baidu.com/s?wd=计算机&rsv_spt=1&rsv_iqid=0x9c2d1e3c000816ea&issp=1&f=8&rs

搜索

拖拽上传

百度一下

百度首页 消息 设置 rainarch

为您推荐: 计算机二级 计算机编程入门 计算机在线应用 计算机学习网站

2016年3月山东计算机二级成绩查询时间 搜狐

3小时前

两市资金净流出314亿 出逃金融及计算机等... 中金在线

4小时前

国金计算机行业周报第7期:政策加速落地... 中证网

11小时前

CSEC/Code.org请愿:希望国会为计算机科学... 网易数码

13小时前

计算机基础知识教程-我要自学网



教程程度: 初级 所需基础: 零基础 交流提问: [点击进入] 适合人群: 电脑初学者 相关素材: 暂无素材资料 课程光盘: [点击链接]标题...
www.51zxw.net/li...asp... V2 - 百度快照

太平洋电脑网 专业IT门户网站



太平洋电脑网是专业IT门户网站,为用户和经销商提供IT资讯和行情报价,涉及电脑,手机,数码产品,软件等.
www.pconline.com.cn/ - 百度快照 - 74%好评

计算机系统 百度百科



2014年4月14日 - 计算机系统由计算机硬件和软件两部分组成。硬件包括中央处理机、存储器和外部设备等;软件是计算机的运行程序和相应的文档。计算机系统具有接收和存储信息、按程序快速...
baike.baidu.com/link?u... V3 - 百度快照

电脑维修电脑, 仅售180元, 价值45



团购: 龙芯电脑主板维修 电话: 15502126759

机构: 龙芯电脑维修 地址: 天目西路547号逸升..

特色: 超值 | 免预约 | 随时退

百度糯米-我的生活

推广链接

搜索 Web 和 Windows



22:04

2016/4/27

计算机 VS 电脑

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

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

- 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。当然我们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

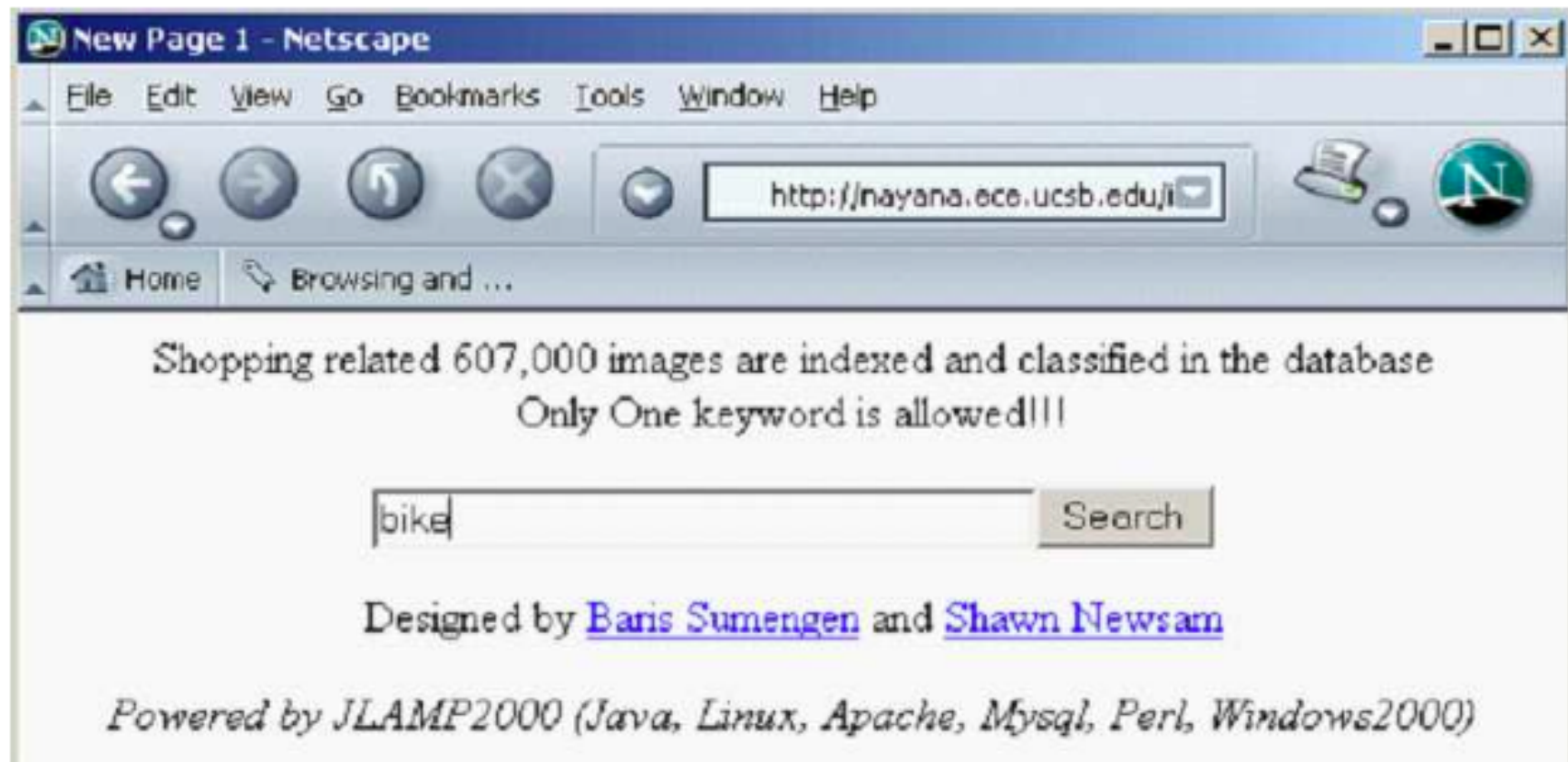
相关反馈分类

- 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。

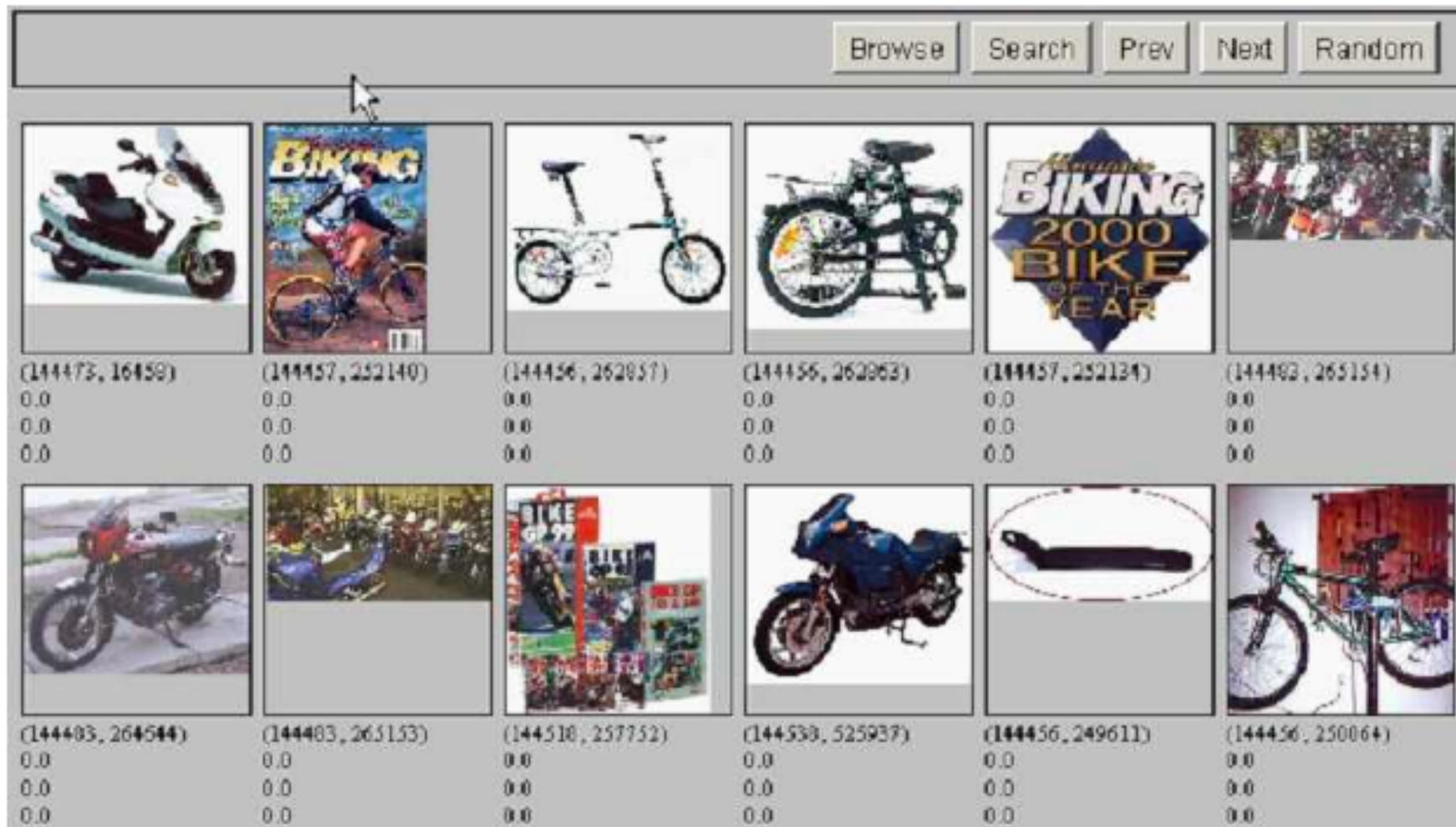
相关反馈

- 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

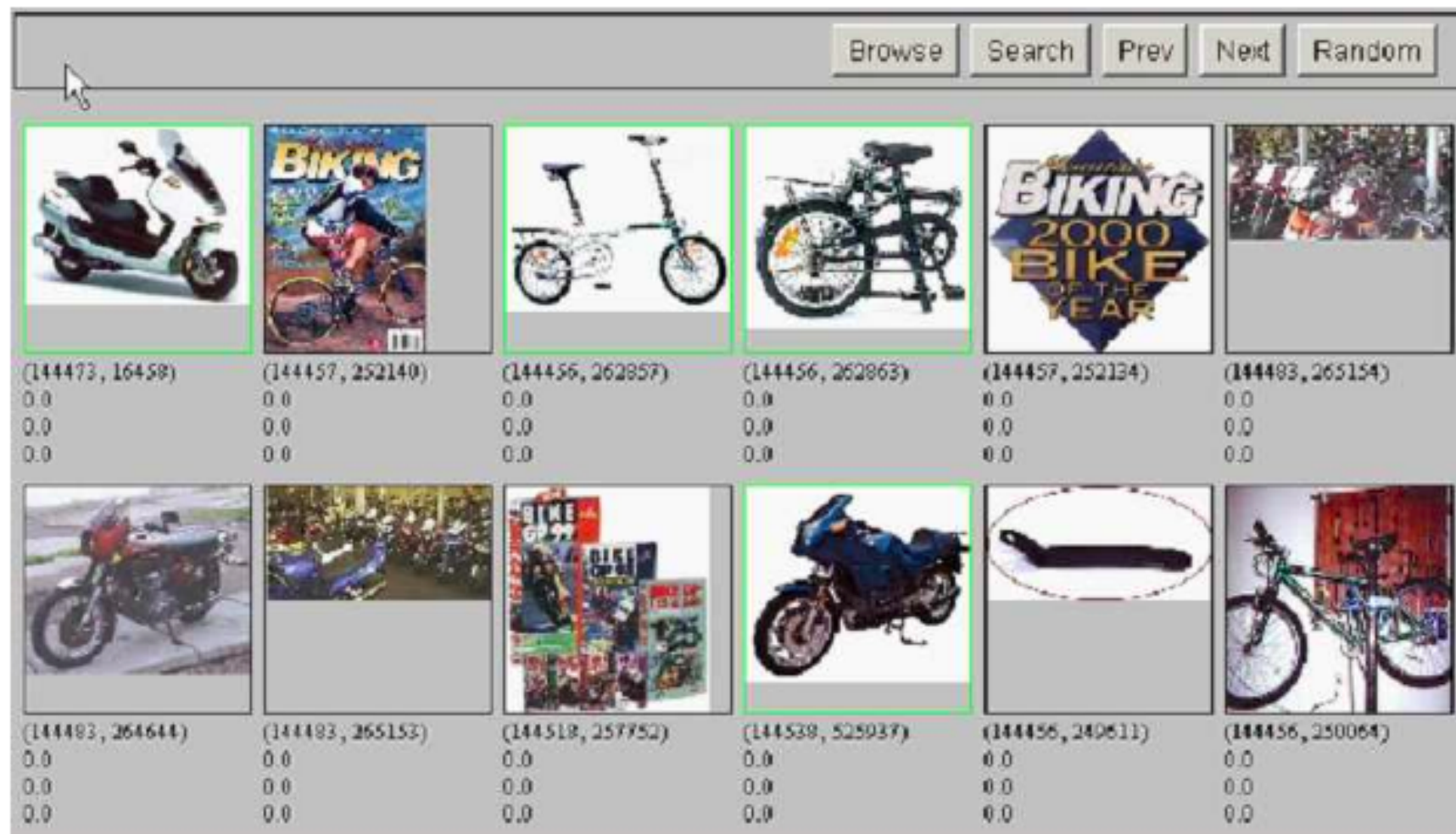
例1



初始查询的结果









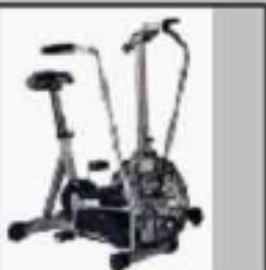





用户反馈: 选择相关结果

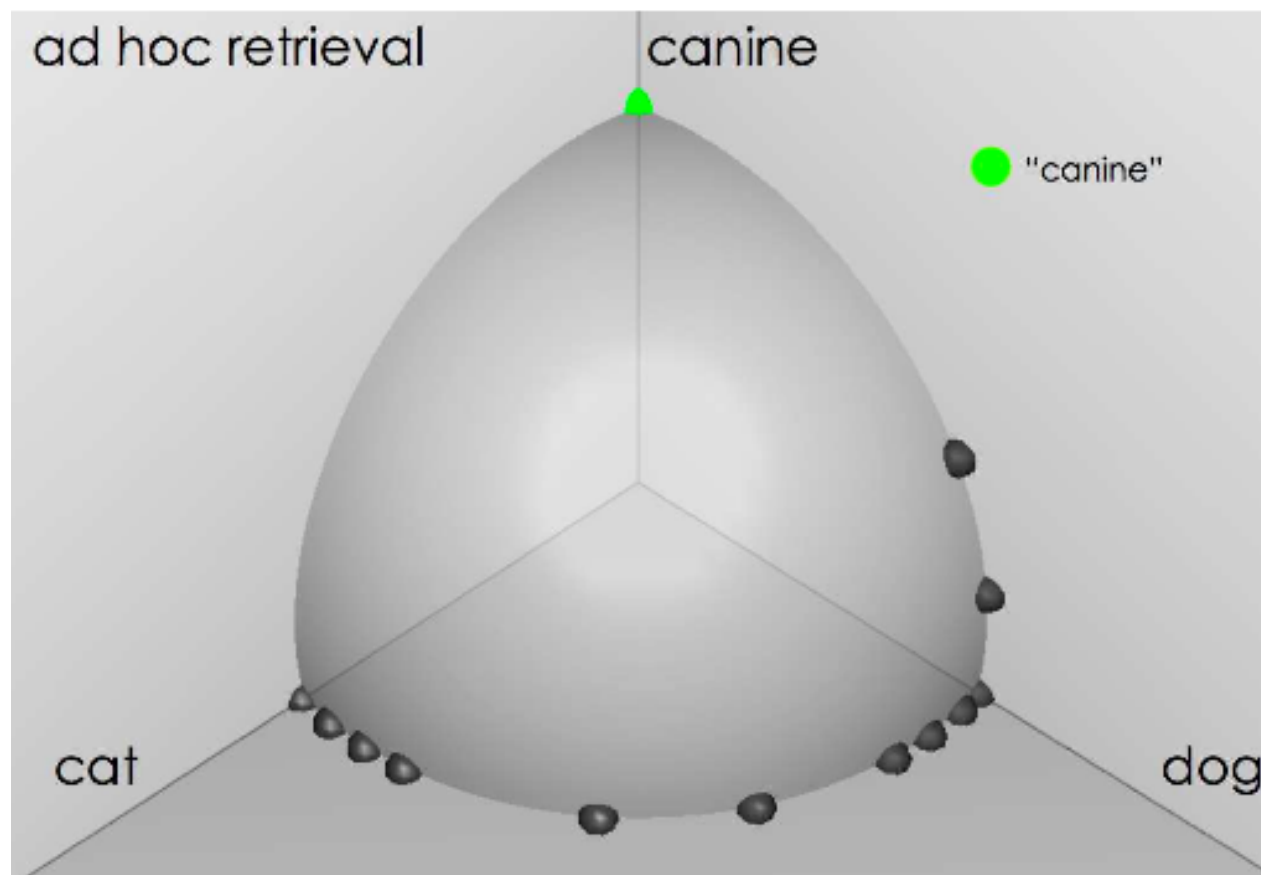


相关反馈后再次检索的结果

[Browse](#) [Search](#) [Prev](#) [Next](#) [Random](#)

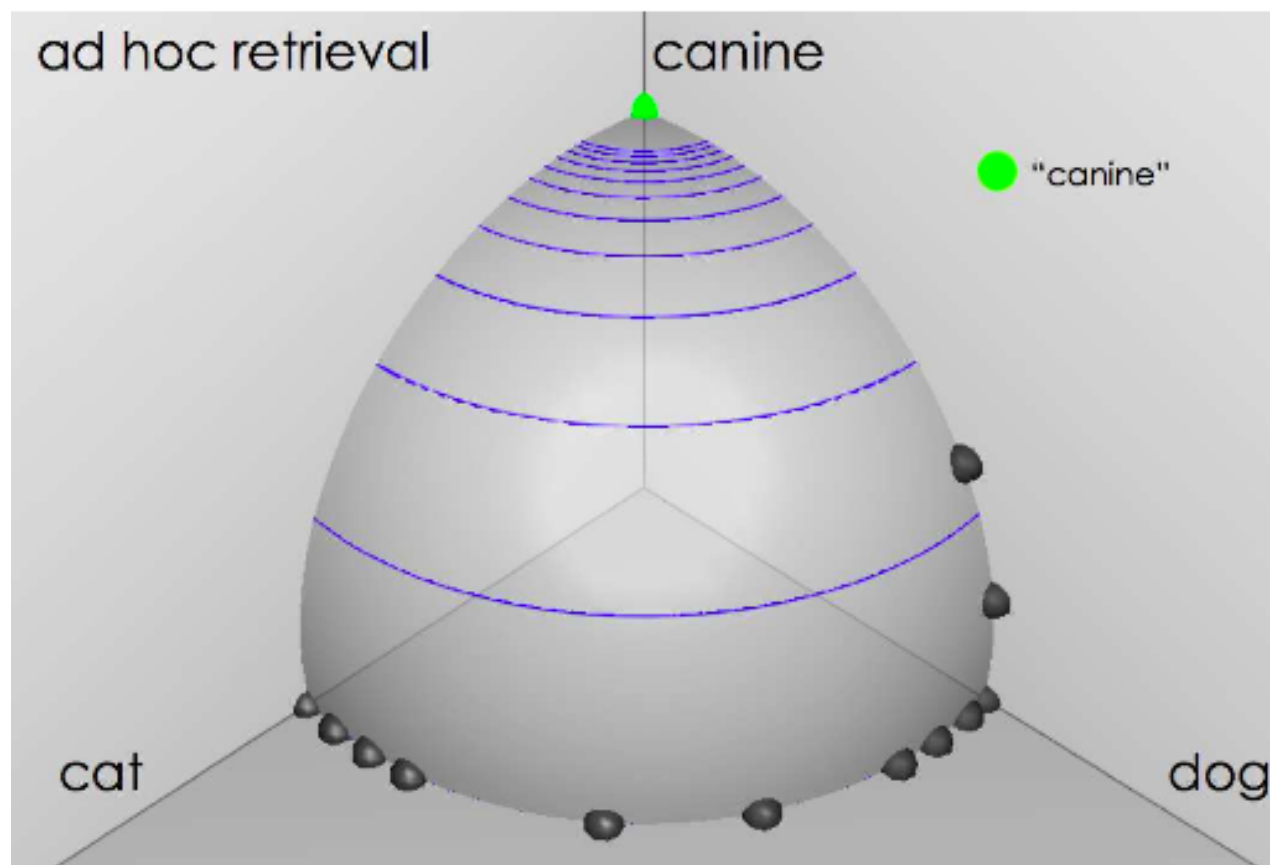
					
(144538, 523493) 0.54182 0.231944 0.309876	(144538, 523835) 0.56319296 0.267364 0.295889	(144538, 523529) 0.584279 0.280881 0.303398	(144456, 253569) 0.64501 0.351395 0.293615	(144456, 253568) 0.650275 0.411745 0.23833	(144538, 523799) 0.66709197 0.358033 0.309059
					
(144473, 16249) 0.6721 0.393922 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

向量空间的例子: 查询 “canine/犬” (1)



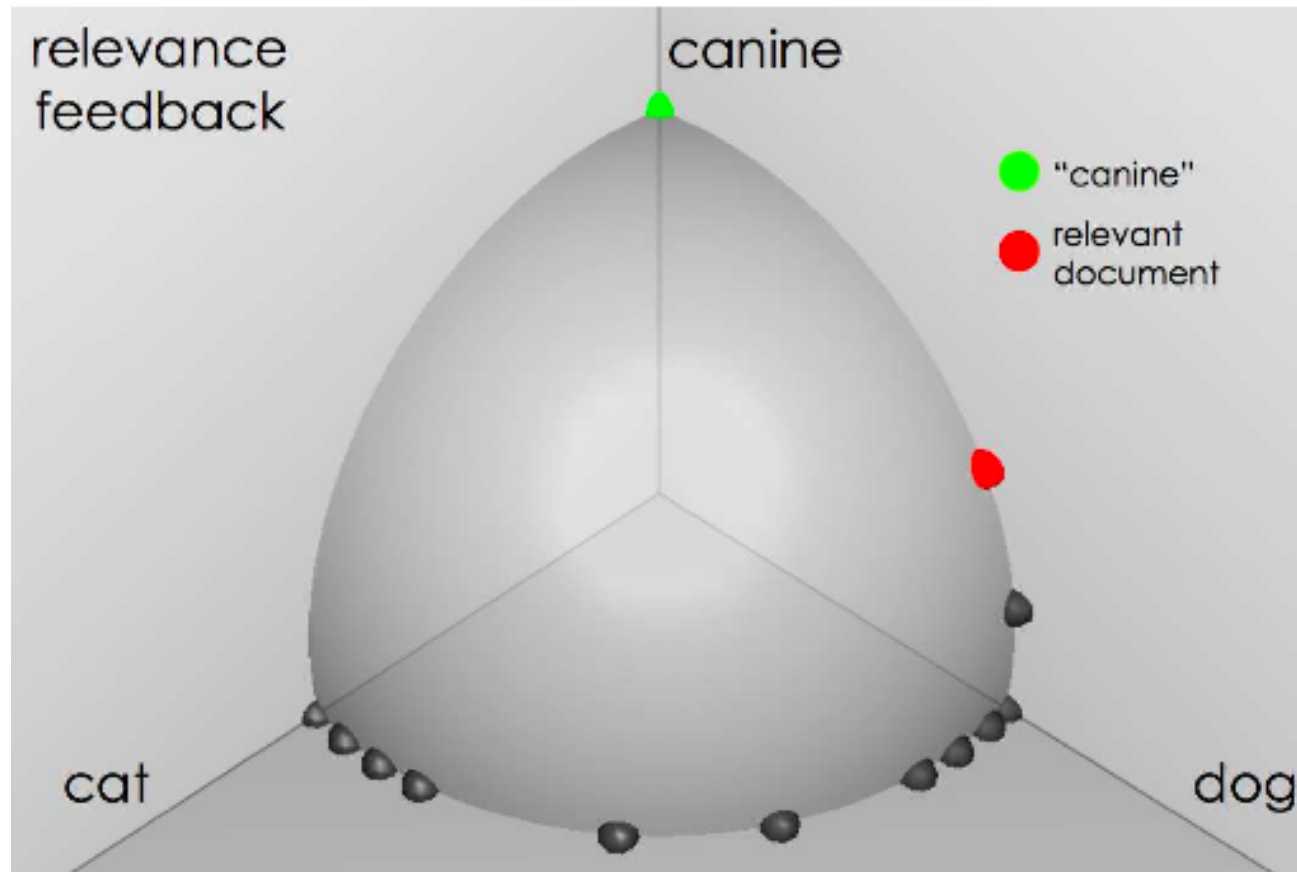
Source:
Fernando Díaz

文档和查询“canine”的相似度



Source:
Fernando Díaz

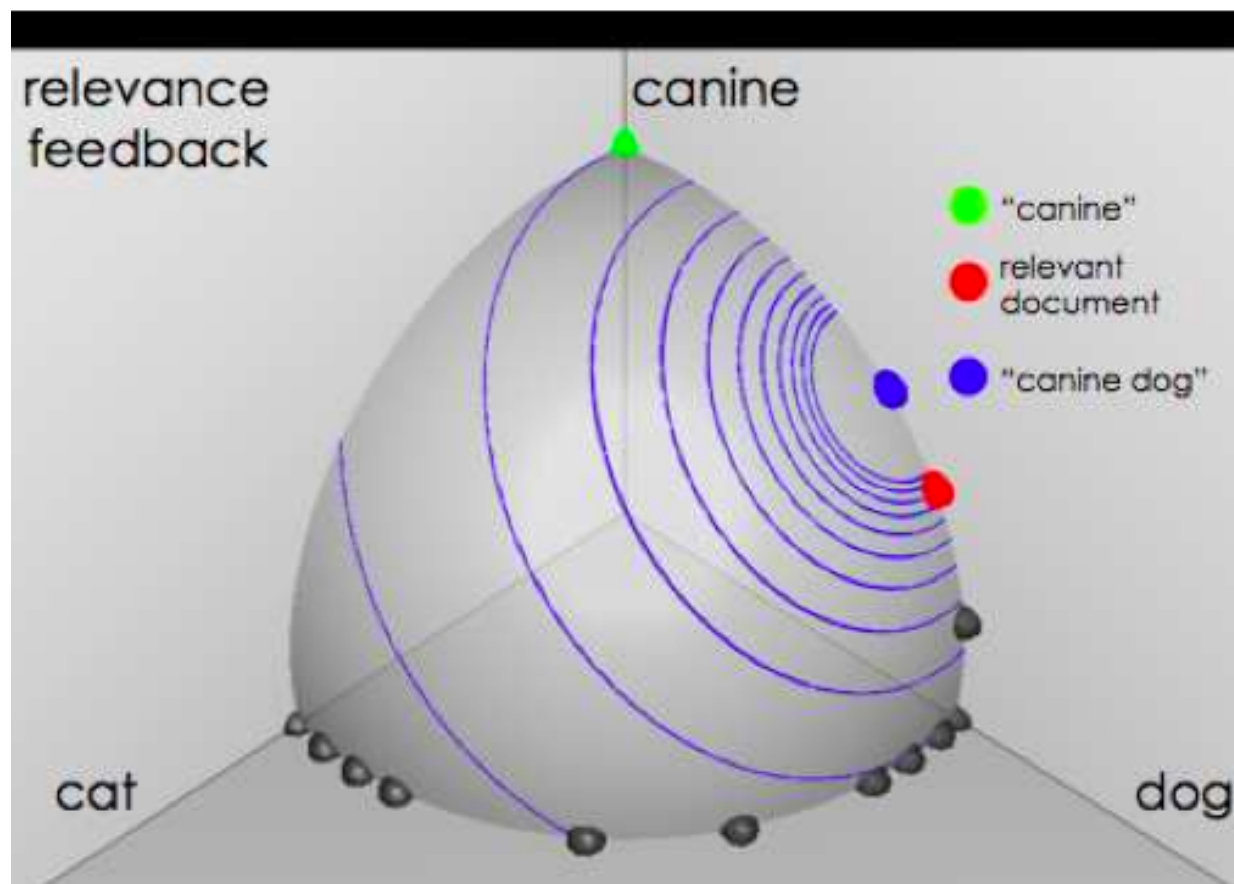
用户反馈: 选择相关文档



Source:
Fernando Díaz

相关反馈后的检索结果

Source:
Fernando Díaz



例3: 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 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

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

	<i>r</i>	
*	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
	8	0.490 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

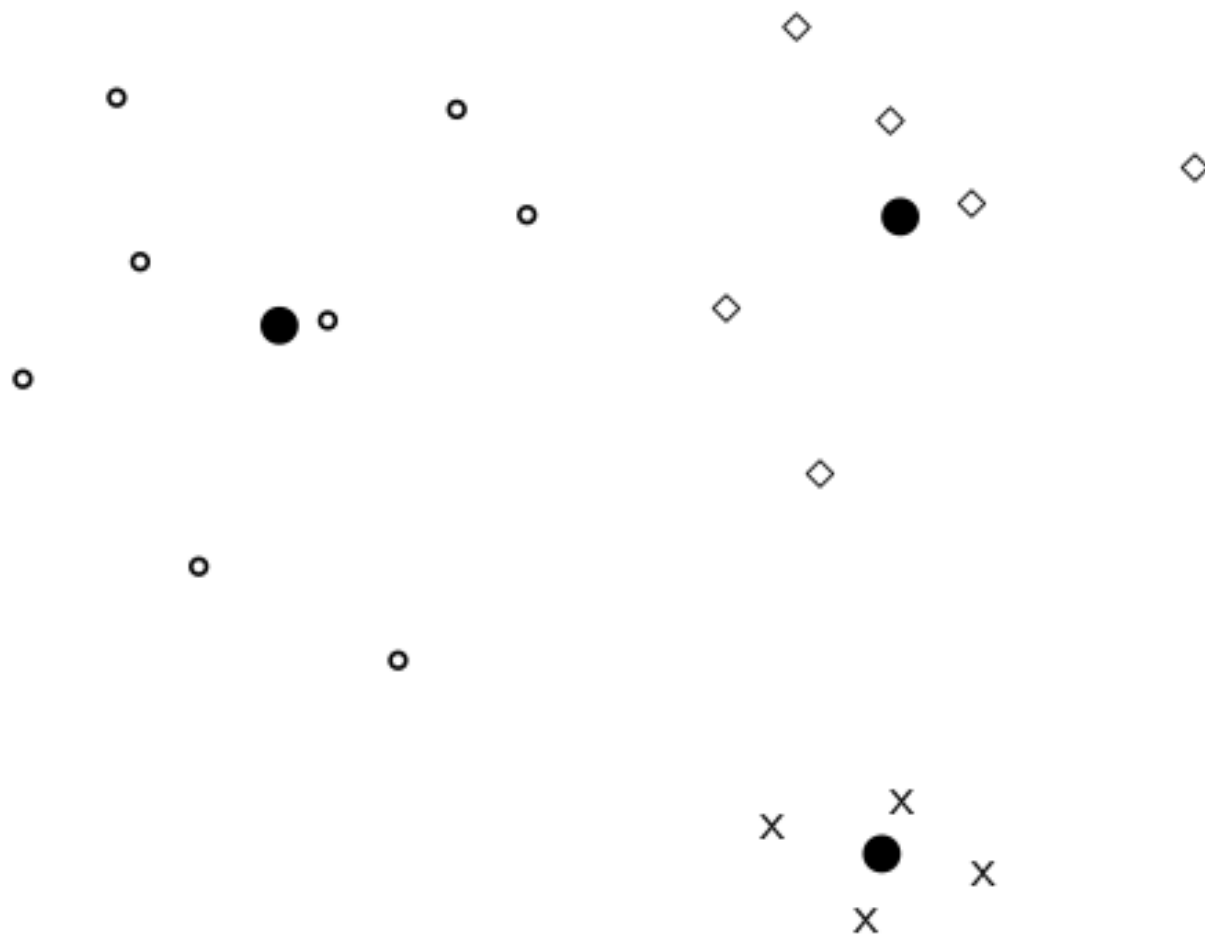
相关反馈中的核心概念：质心

- 质心指的是一系列点的中心
- 前面我们将文档表示成高维空间中的点
- 因此，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 的的向量表示

质心的例子



Rocchio算法

- Rocchio算法是向量空间模型中相关反馈的实现方式
- Rocchio算法选择使下式最大的查询 \vec{q}_{opt}

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

D_r : 相关文档集; D_{nr} : 不相关文档集

- 上述公式的意图 \vec{q}_{opt} 是将相关文档和不相关文档分得最开的向量。
- 加入一些额外的假设，可以将上式改写为：

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

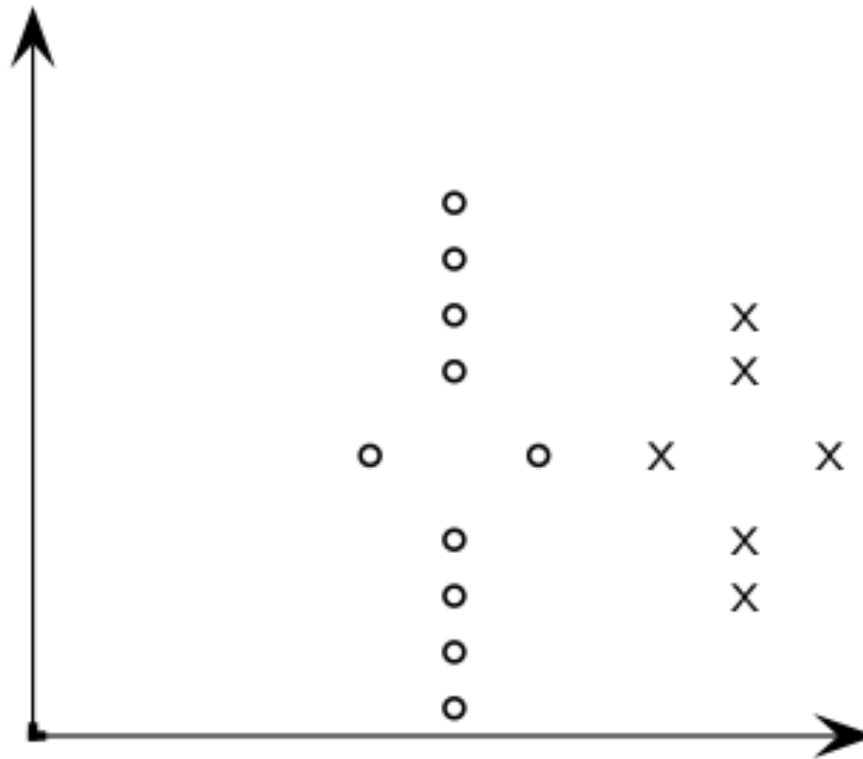
Rocchio算法

- 最优查询向量为：

$$\begin{aligned}\vec{q}_{opt} &= \mu(D_r) + [\mu(D_r) - \mu(D_{nr})] \\ &= \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j + \left[\frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j \right]\end{aligned}$$

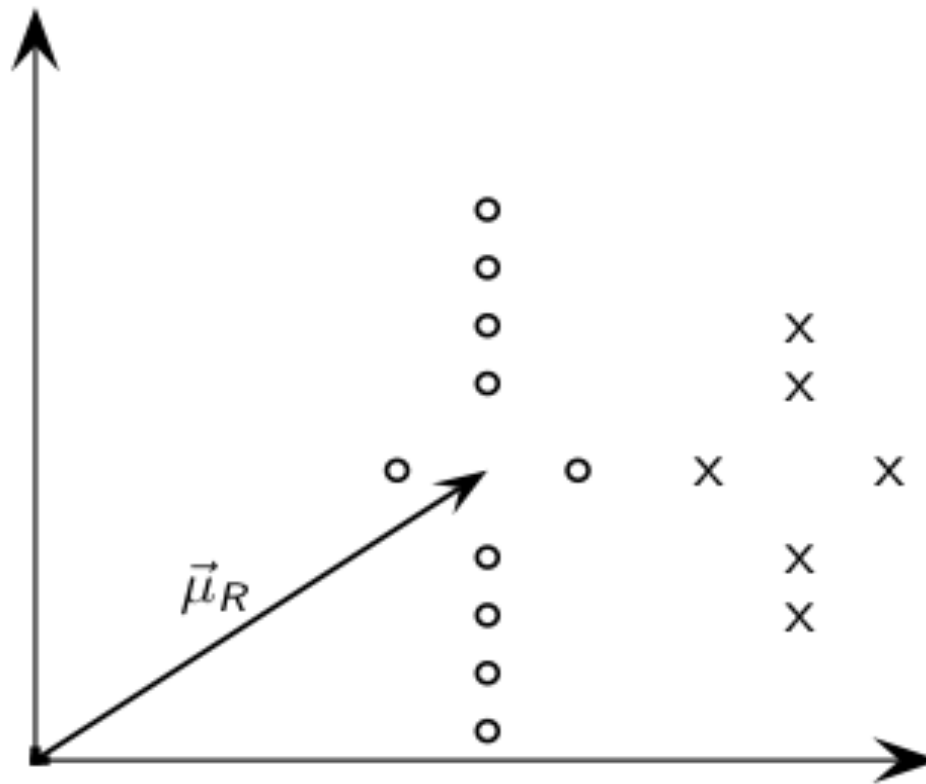
- 即将相关文档的质心移动一个量，该量为相关文档质心和不相关文档的差异量

计算Rocchio向量



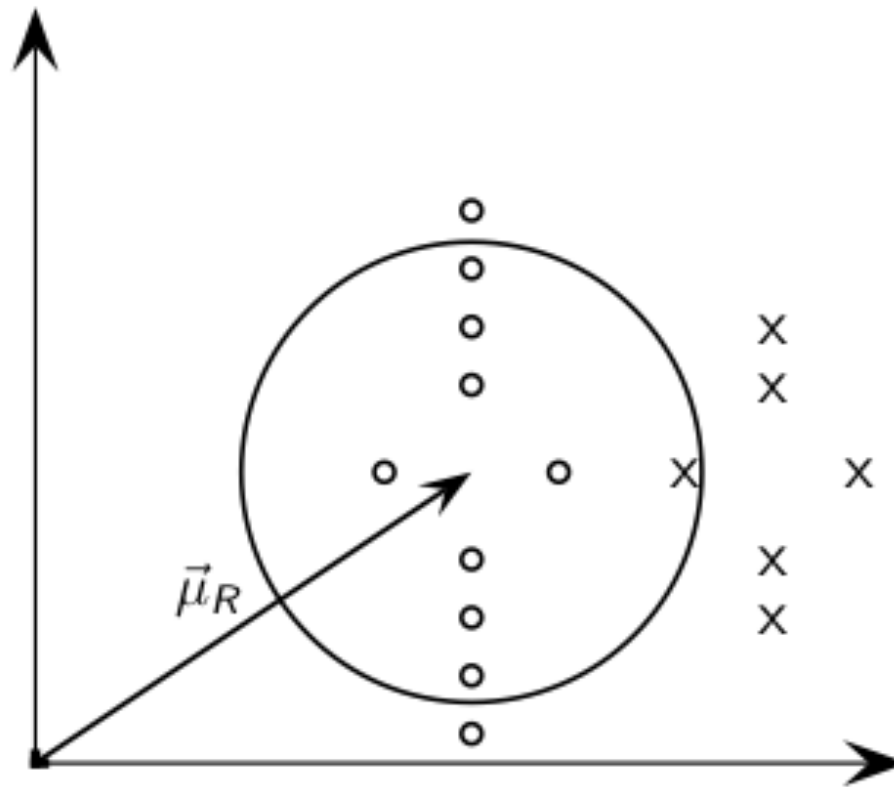
圆形点: 相关文档, 叉叉点: 不相关文档

Rocchio算法图示



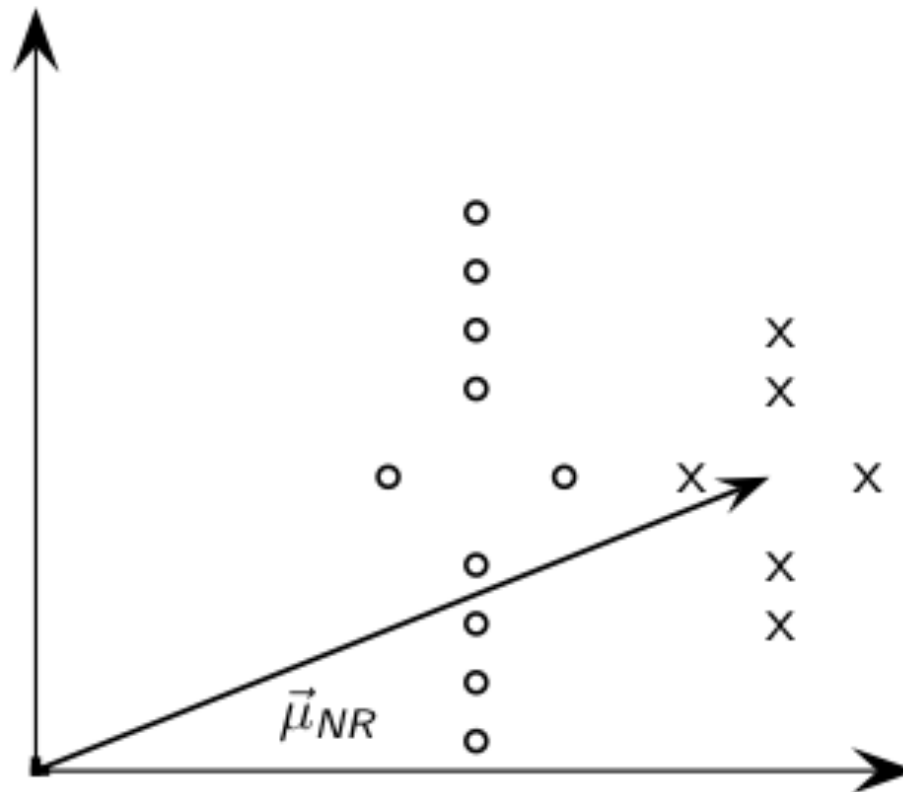
$\vec{\mu}_R$: 相关文档的质心

Rocchio算法图示



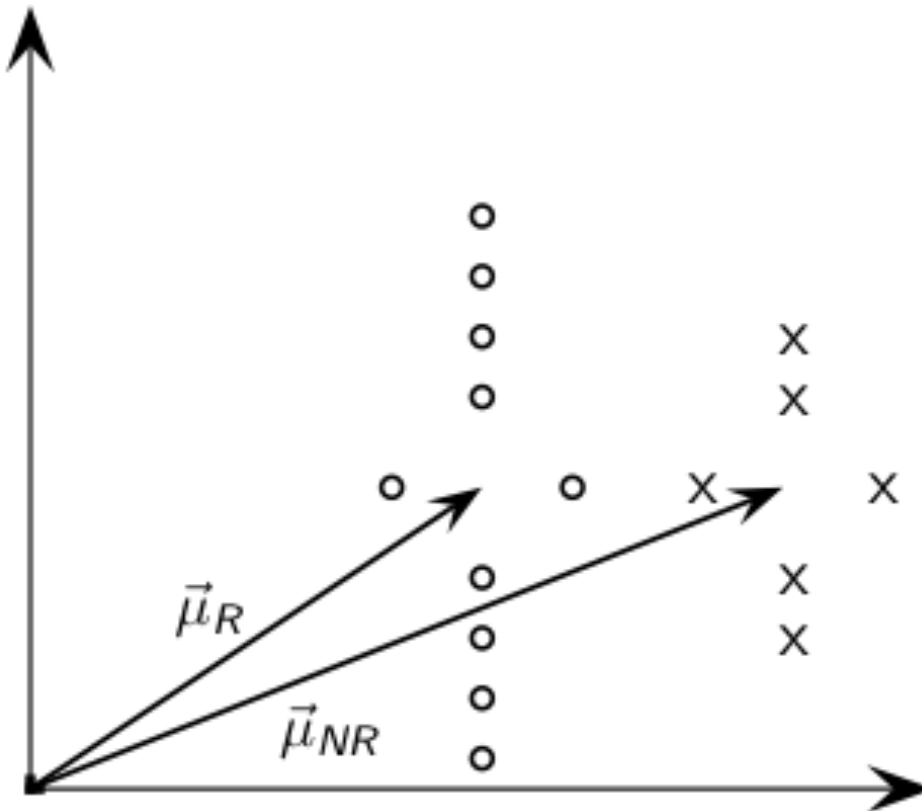
$\vec{\mu}_R$ 不能将相关/不相关文档分开

Rocchio算法图示

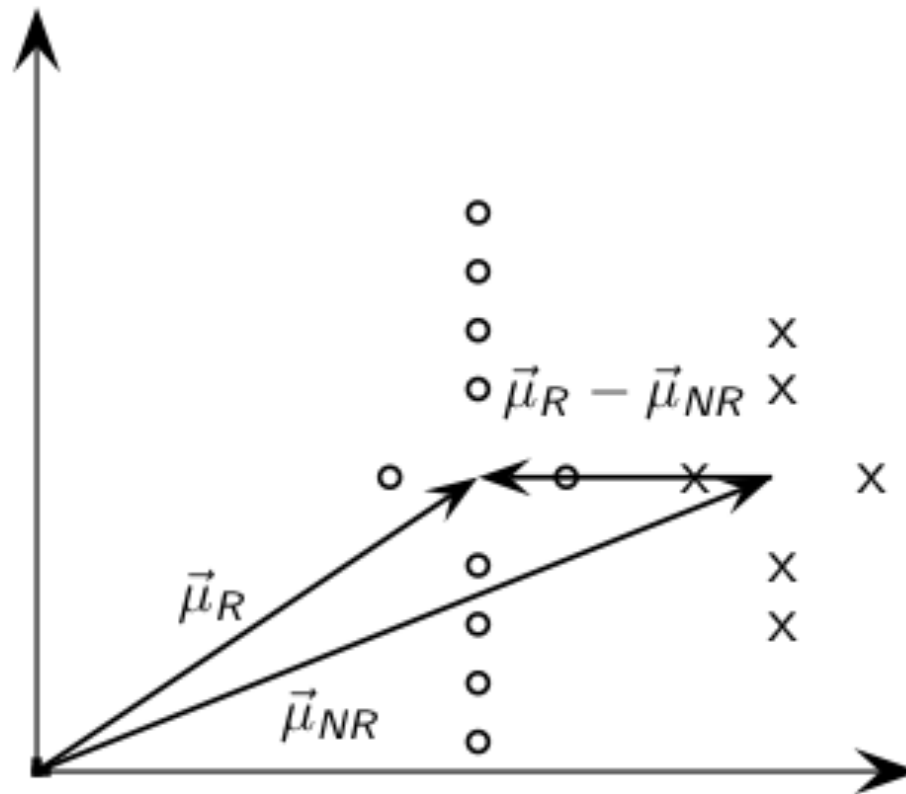


$\vec{\mu}_{NR}$: 不相关文档的质心

Rocchio算法图示

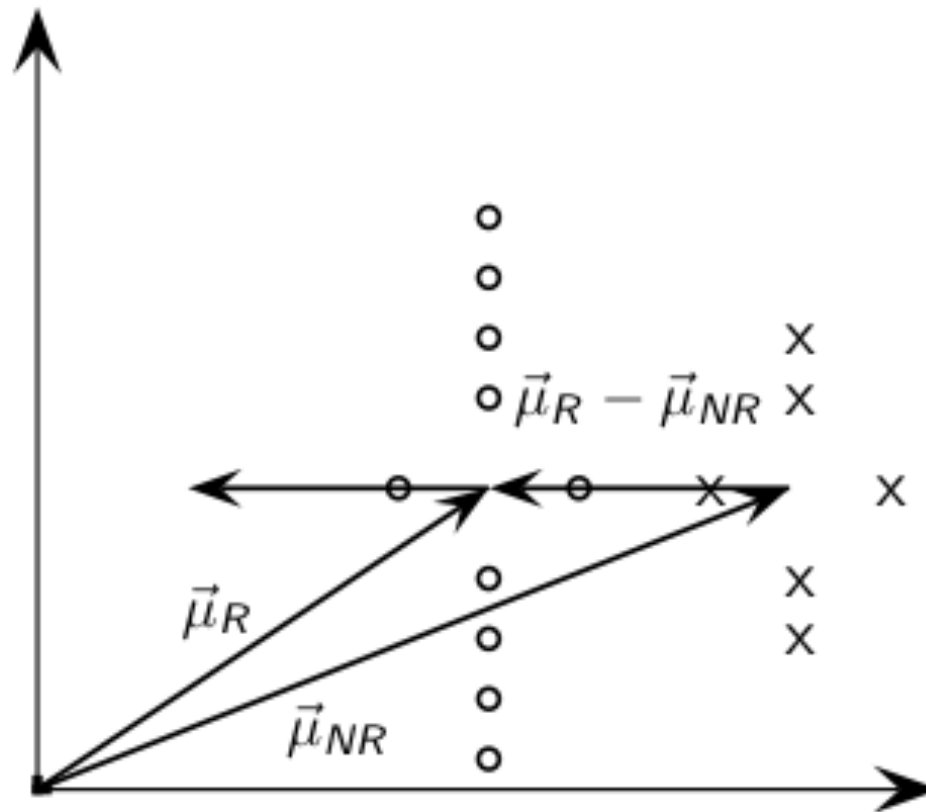


Rocchio' 算法图示



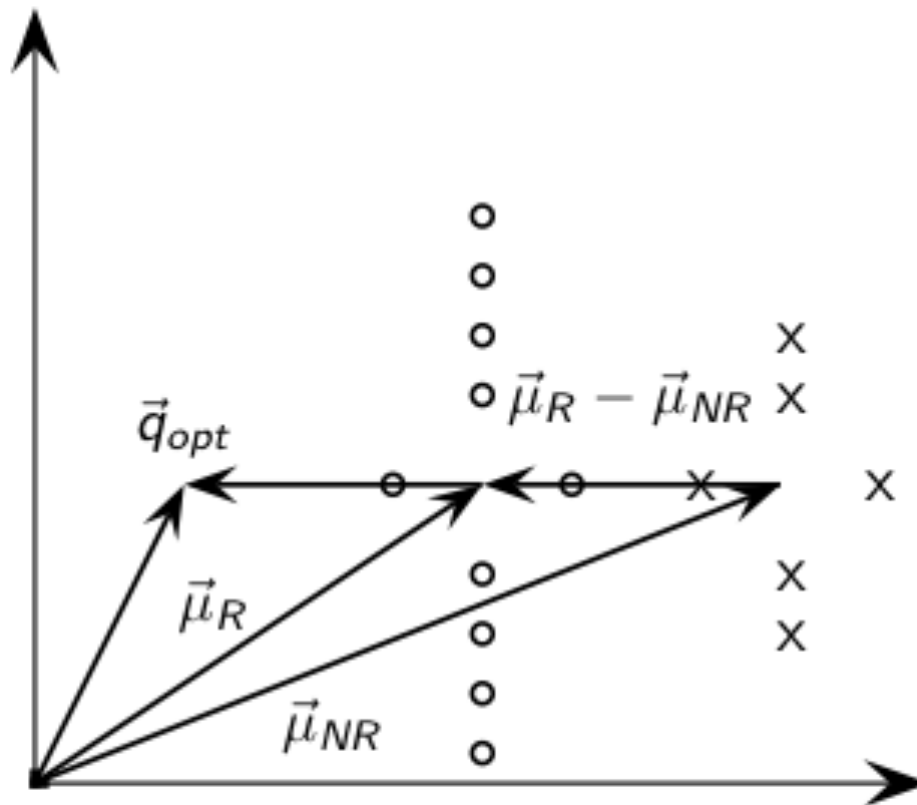
$\vec{\mu}_R - \vec{\mu}_{NR}$: 差异向量

Rocchio算法图示



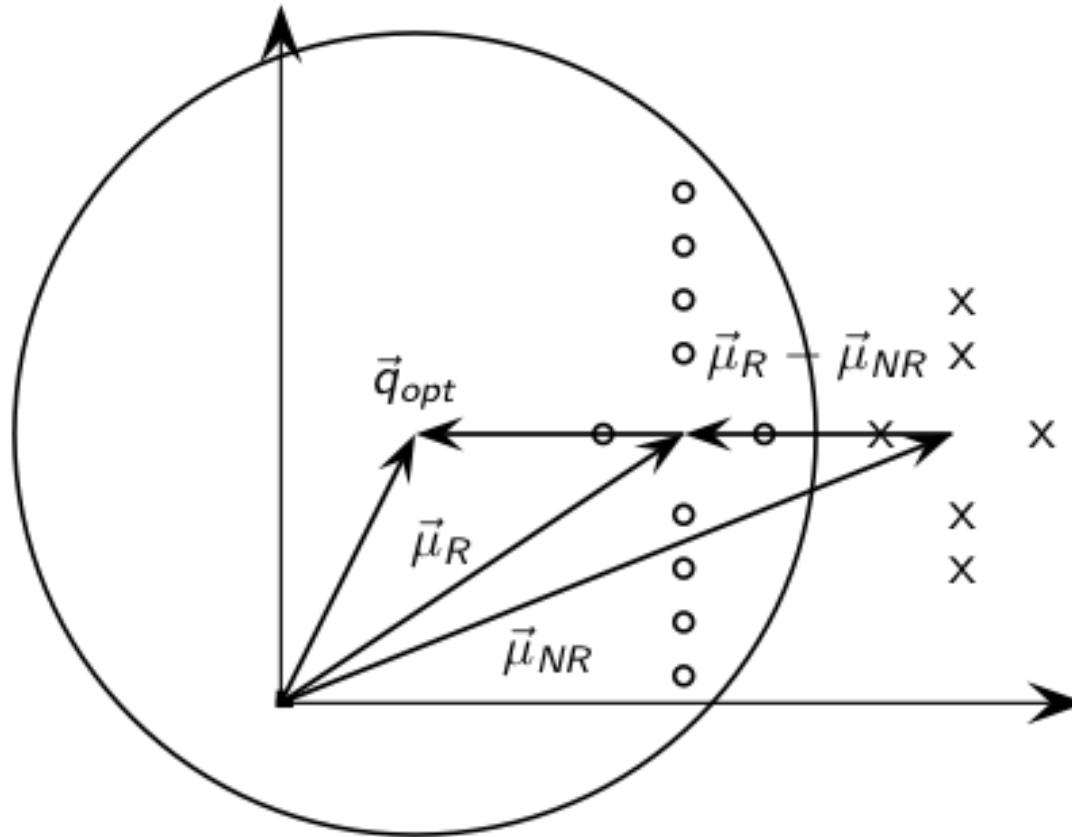
$\vec{\mu}_R$ 加上差异向量

Rocchio算法图示



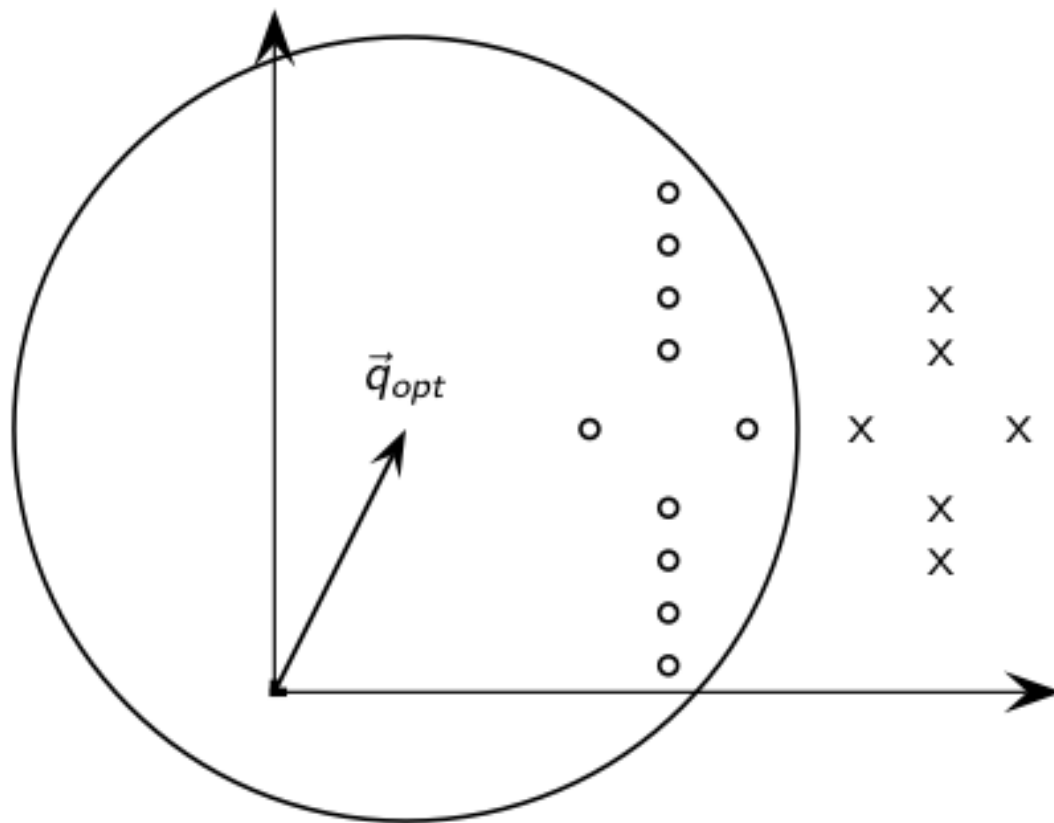
得到 \vec{q}_{opt}

Rocchio算法图示



\vec{q}_{opt} 能够将相关/不相关文档完美地分开

Rocchio算法图示



\vec{q}_{opt} 能够将相关/不相关文档完美地分开
请把上述过程琢磨一下

Rocchio 1971 算法 (SMART系统使用)


实际中使用的公式:

$$\begin{aligned}\vec{q}_m &= \alpha \vec{q}_0 + \beta \mu(D_r) - \gamma \mu(D_{nr}) \\ &= \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j\end{aligned}$$

q_m : 修改后的查询; q_0 : 原始查询;

D_r 、 D_{nr} : 已知的相关和不相关文档集合

α, β, γ 权重

- 新查询向相关文档靠拢而远离非相关文档
- α vs. β/γ 设置中的折中: 如果判定的文档数目很多, 那么 β/γ 可以考虑设置得大一些 
- 一旦计算后出现负权重, 那么将负权重都设为0
- 在向量空间模型中, 权重为负是没有意义的。

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

- 什么时候相关反馈能否提高召回率？
- 假设 A1: 对于某初始查询，用户知道在文档集中使用哪些词项来表达
- 假设 A2: 相关文档中出现的词项类似 (因此，可以基于相关反馈，从一篇相关文档跳到另一篇相关文档)
 - 或者: 所有文档都紧密聚集在某个prototype周围
 - 或者: 有多个不同的prototype, 但是它们之间的用词具有显著的重合率
 - 相关文档和不相关文档之间的相似度很低

假设A1不成立的情况

- 假设 A1: 对于某初始查询，用户知道在文档集中使用哪些词项来表达
- 不成立的情况：用户的词汇表和文档集的词汇表不匹配
- 例子： cosmonaut / astronaut

假设A2不成立的情况

- 假设A2: 相关文档中出现的词项类似
- 假设不成立的查询例子: [contradictory government policies] 互相矛盾的政府政策
- 一些相关的文档集合, 但是文档集合彼此之间并不相似
 - 文档集合1: 烟草种植者的补贴 vs. 禁烟运动
 - 文档集合2: 对发展中国家的帮助 vs. 发展中国家进口商品的高关税
- 有关烟草文档的相关反馈并不会对发展中国家的文档有所帮助

相关反馈的评价

- 选择上一讲中的某个评价指标，比如 $P@10$
- 计算原始查询 q_0 检索结果的 $P@10$ 指标 for original query
- 计算修改后查询 q_1 检索结果的 $P@10$ 指标
- 大部分情况下 q_1 的检索结果精度会显著高于 q_0 !
- 上述评价过程是否公平?

相关反馈的评价

- 公平的评价过程一定要基于存留文档集(residual collection): 用户没有判断的文档集
- 研究表明采用, 采用这种方式进行评价, 相关反馈是比较成功的一种方法
- 经验而言, 一轮相关反馈往往非常有用, 相对一轮相关反馈, 两轮相关反馈效果的提高有限。

有关评价的提醒

- 相关反馈有效性的正确评价，必须要和其他需要花费同样时间的方法
- 相关反馈的一种替代方法：用户修改并重新提交新的查询
- 用户更倾向于修改和重新提交查询而不是判断文档的相关性
- 并没有清晰的证据表明，相关反馈是用户时间使用的最佳方法

课堂练习

- 搜索引擎是否使用相关反馈?
- 为什么?

相关反馈存在的问题

- 相关反馈开销很大
 - 相关反馈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	http://bbs.cixi.cn/dispbbs.asp?Star=4&boardid=46&id=346721&page=1
Page number	1
Rank	7
Anchor text	姑娘们，你们愿意嫁给警察吗？ [慈溪社区]

眼球动作(通过鼠标轨迹模拟)

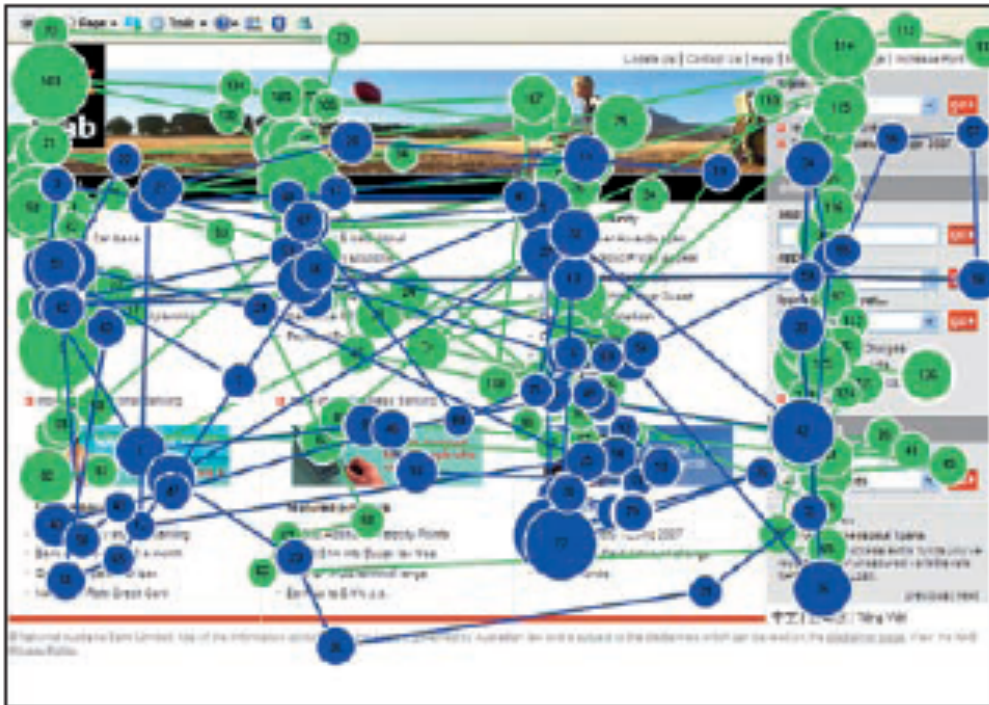


Baidu



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*)

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. l) 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

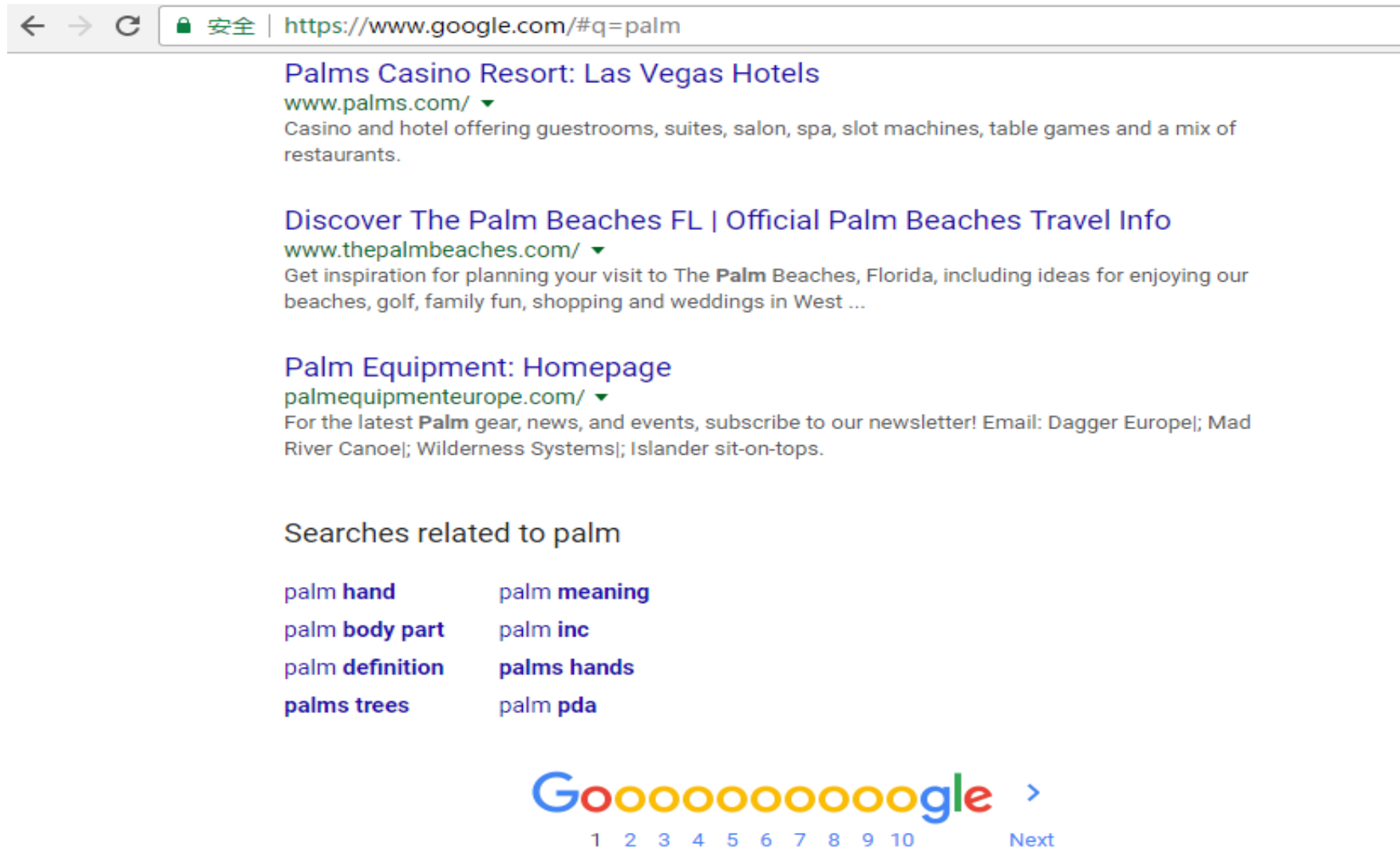
提纲

- 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(Query expansion)

- 查询扩展是另一种提高召回率的方法
- 我们使用“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



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

- 人工构建的同(近)义词词典 (人工编辑人员维护的词典, 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**(有吸引力)
- 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

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

Matching

PubMed Query:

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

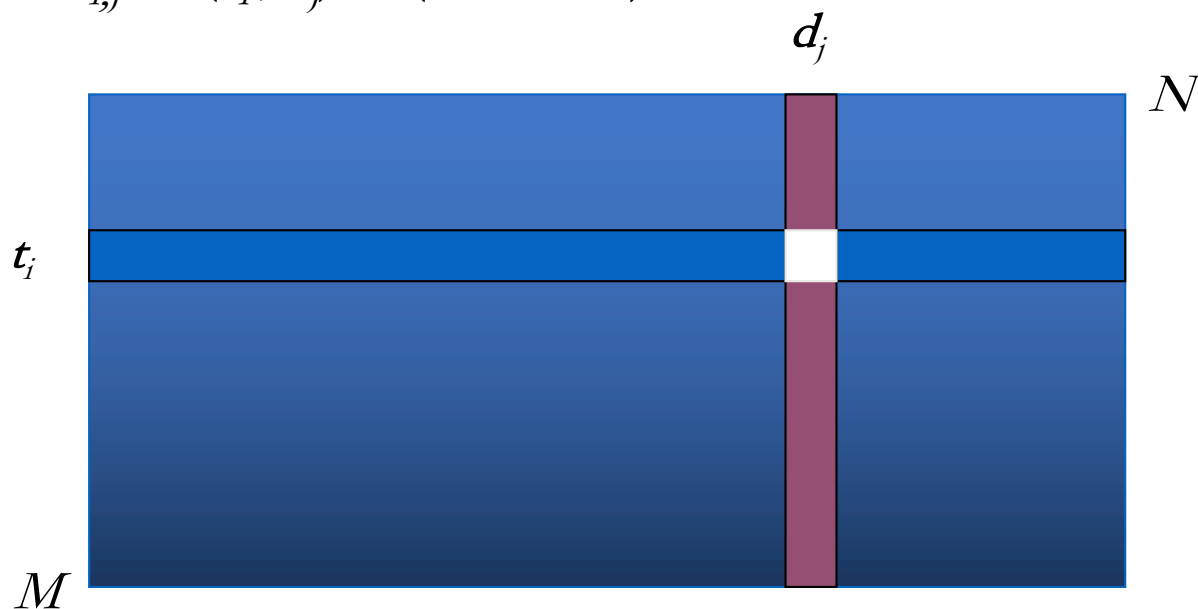
Search URL

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
- 基本的想法是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” \approx “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: 两个词，如果它们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

- 最简单的方法就是通过词典-文档矩阵 A 计算词项-词项的相似度 $C = AA^T$
- $w_{i,j} = (t_i, d_j)$ 的(归一化)权重



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

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

课堂练习

- 请列举下列查询扩展的词汇（每条5个扩展词），请写出理由。
 - 计算机
 - 华东师范大学大学
 - 刘翔

本讲小结

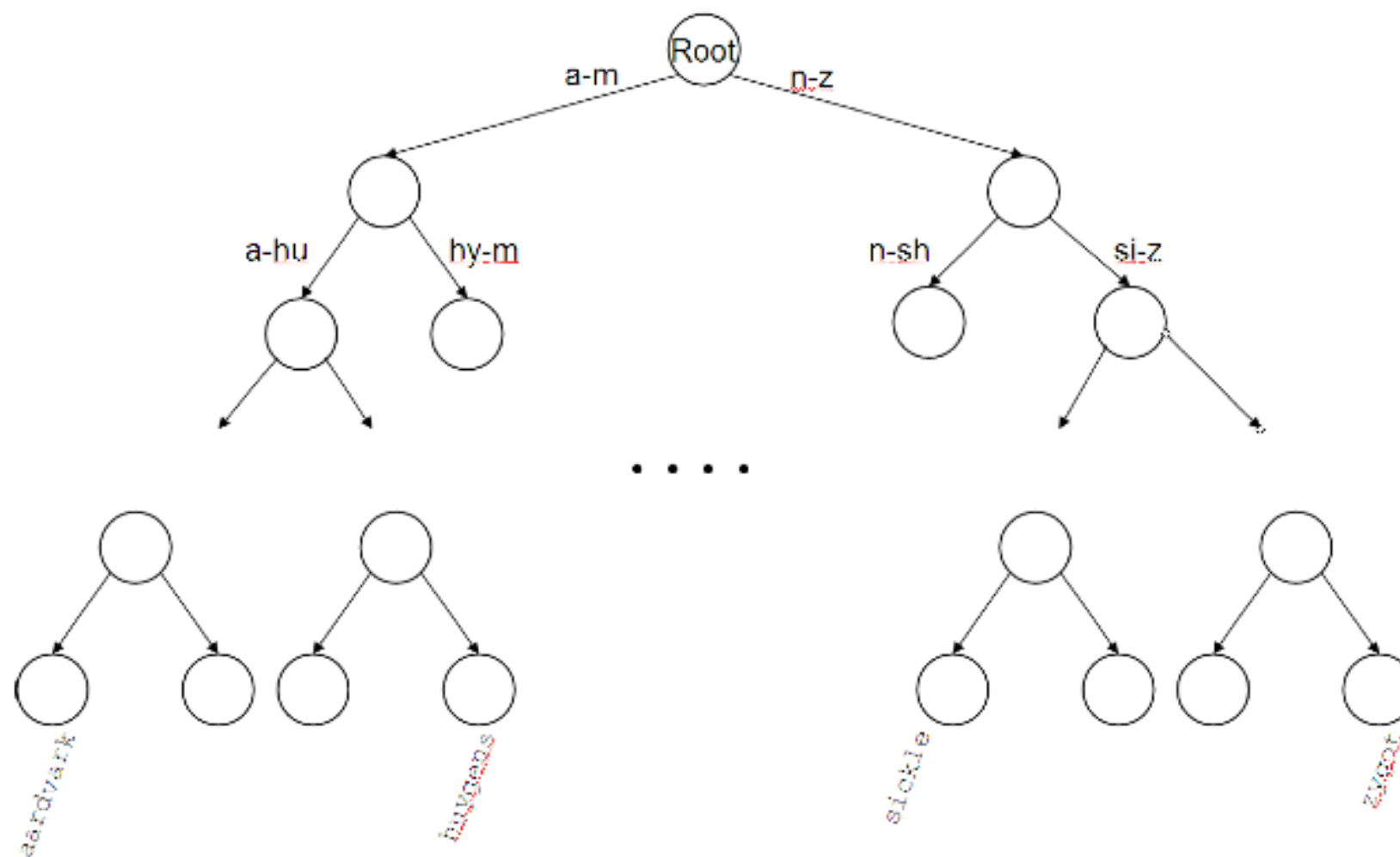
- 交互式相关反馈(Interactive relevance feedback): 在初始检索结果的基础上, 通过用户交互指定哪些文档相关或不相关, 然后改进检索的结果
- 最著名的相关反馈方法: Rocchio 相关反馈
- 查询扩展(Query expansion): 通过在查询中加入同义或者相关的词项来提高检索结果
 - 相关词项的来源: 人工编辑的同义词词典、自动构造的同义词词典、查询日志等等。

补充

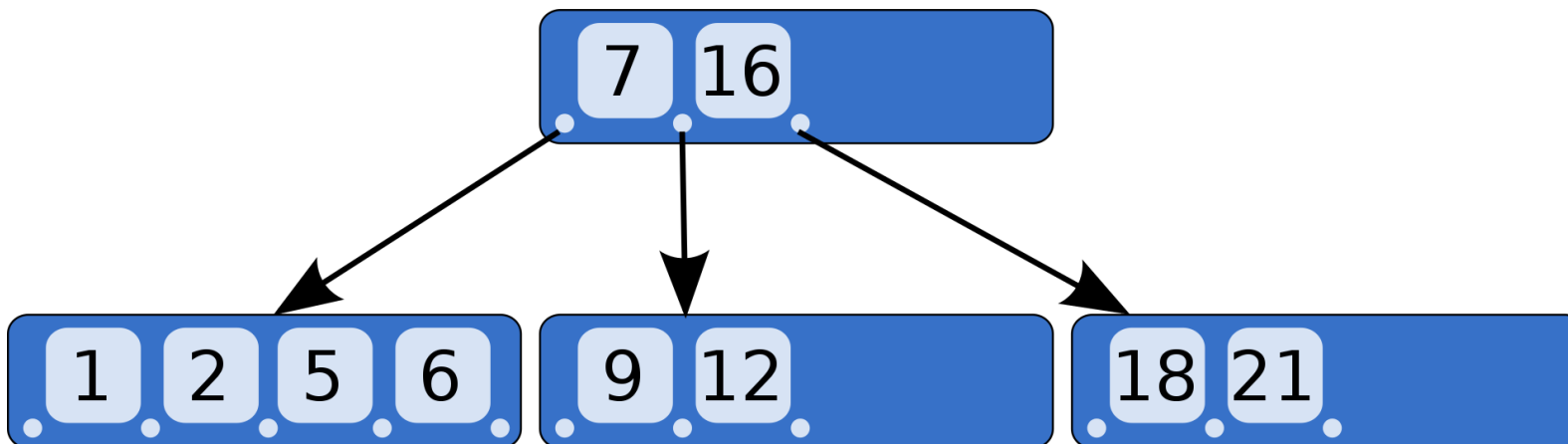
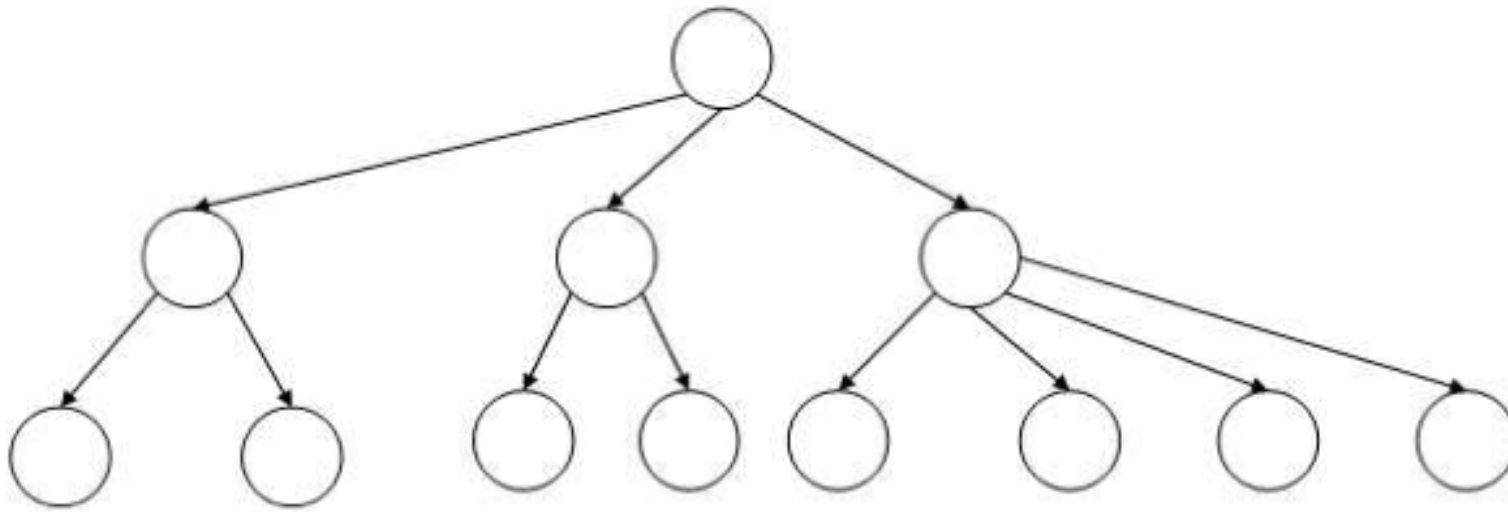
树

- 树可以支持前缀查找
- 最简单的树结构：二叉树
- 搜索速度略低于哈希表方式： $O(\log M)$, 其中 M 是词汇表大小，即所有词项的数目
 - $O(\log M)$ 仅仅对平衡树成立
 - 使二叉树重新保持平衡开销很大
- B-树 能够减轻上述问题
- B-树定义：每个内部节点的子节点数目在 $[a, b]$ 之间，其中 a, b 为合适的正整数, e.g., $[2, 4]$.

二叉树



B-树



B树/B-树/B+树

- 二叉搜索树
- B树=B-树（翻译问题）
 - 一种多路搜索树（并不是二叉的） B-Tree
- B+树：B+树是B-树的变体
 - B+的搜索与B-树也基本相同，区别是B+树只有达到叶子结点才命中（B-树可以在非叶子结点命中）

提纲

① 上一讲回顾

② 词典

③ 通配查询

④ 编辑距离

⑤ 拼写校正

⑥ Soundex

通配查询的处理

- mon^* : 找出所有包含以 *mon* 开头的词项的文档
- 如果采用B-树词典结构，那么实现起来非常容易，只需要返回区间 $mon \leq t < moo$ 上的词项 t
- $*mon$: 找出所有包含以 *mon* 结尾的词项的文档
 - 将所有的词项倒转过来，然后基于它们建一棵附加的树
 - 返回区间 $nom \leq t < non$ 上的词项 t
- 也就是说，通过上述数据结构，可能得到满足通配查询的一系列词项，然后返回任一词项的文档

通配查询的处理

- 怎么处理: $m \times n$ chen

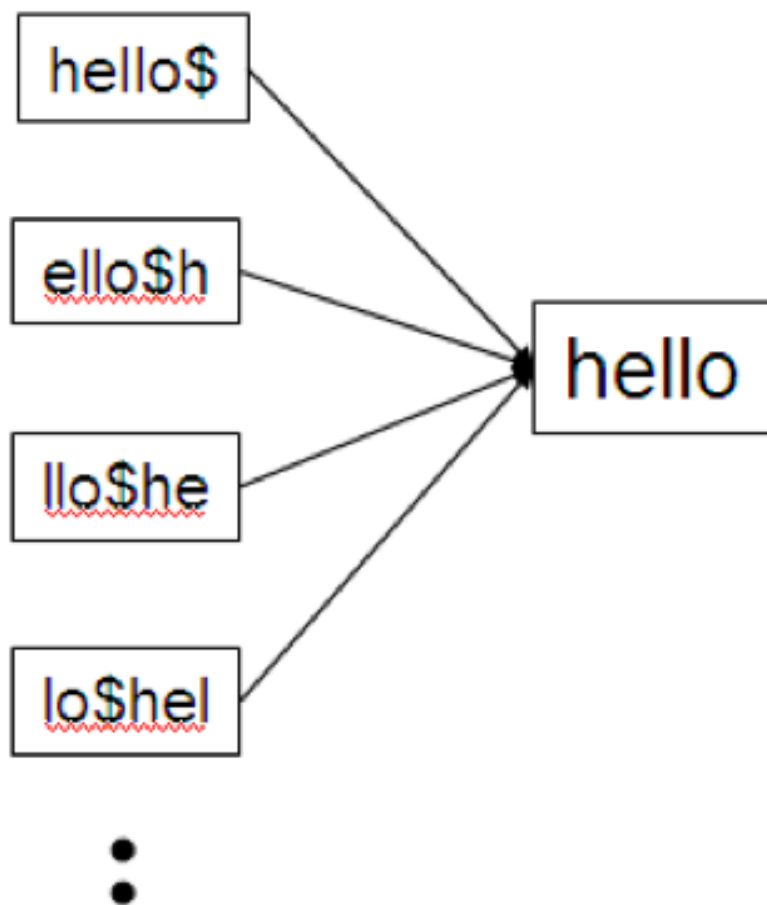
词项中间的 *号处理

- 例子: $m * n$ chen
- 方法1:
 - 在B-树中分别查找满足 $m *$ 和 $*n$ chen的词项集合，然后求交集
 - 这种做法开销很大
- 方法2:
 - 轮排(permuterm) 索引
 - 基本思想：将每个通配查询旋转，使*出现在末尾
 - 将每个旋转后的结果存放在词典中，即B-树中

轮排索引

- 对于词项hello: 将 *hello\$*, *ello\$h*, *llo\$he*, *lo\$hel*, 和 *o\$hell* 加入到 B-树中, 其中 \$ 是一个特殊符号
- 即在词项前面再加一层索引

轮排结果 → 词项的映射示意图



轮排索引

- 对于hello, 已经存储了 *hello\$, ello\$h, llo\$he, lo\$hel, o\$hell, \$hello*
- 查询
 - 对于 x, 查询 x\$
 - 对于 x*, 查询 \$x*
 - 对于 *x, 查询 x\$*
 - 对于 *x*, 查询 x*
 - 对于 x*y, 查询 y\$x*
 - 例子: 假定通配查询为 hel*o, 那么相当于要查询o\$hel*
- 轮排索引称为轮排树更恰当
- 但是轮排索引已经使用非常普遍

现场例子？

Hello, Hello\$

He*, He*\$, \$He*

*lo, *lo\$, lo\$*

*on, *on\$, on\$*

*lo, *lo\$, lo\$*

X*Y, X*Y\$, Y\$X*

使用轮排索引的查找过程

- 将查询进行旋转，将通配符旋转到右部
- 同以往一样查找B-树
- 问题：相对于通常的B-树，轮排树的空间要大4倍以上 (经验值)