信息检索与Web搜索

第10讲 相关反馈及查询扩展

Relevance Feedback & Query Expansion

授课人：高曙明

改编自“现代信息检索”网上公开课件（http://ir.ict.ac.cn/~wangbin）

关于查询优化

- 与查询相关的问题：
  - 构建准确的查询不容易
  - 对于同样查询，不同用户希望得到的结果不尽相同
  - 召回率不够好：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
- 通过对查询进行重构优化，提高检索效果，包括实现与具体用户相关的检索效果
- **解决方法：**相关反馈和查询扩展

相关反馈的基本思想和流程

- **基本思想：**根据用户对查询结果的标注，生成更有效更优化的查询，以提高检索效果
- **流程：**
  - 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更新查询表示
  - 搜索引擎对新查询进行处理，返回新结果

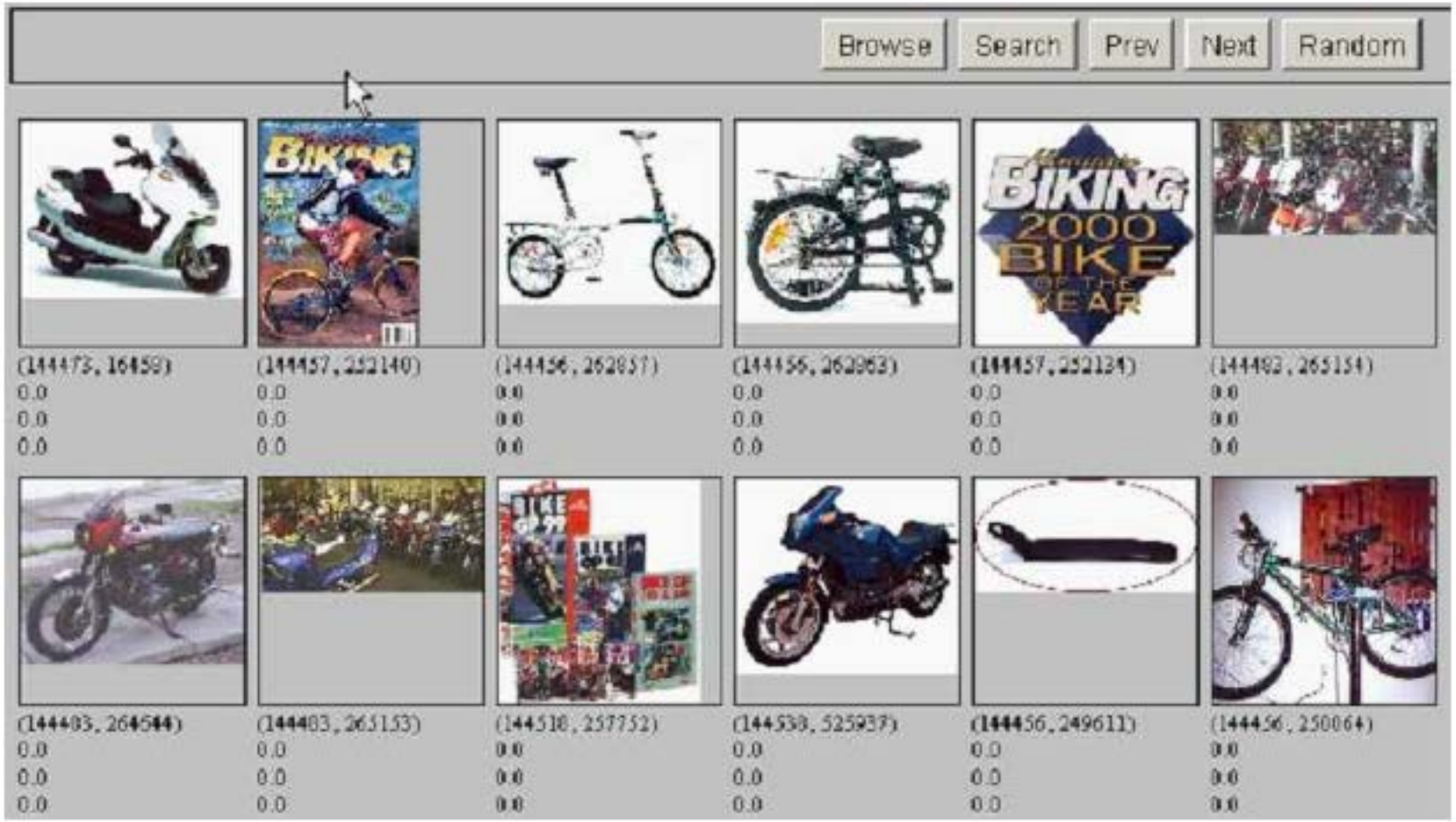
相关反馈分类

- **显式相关反馈**(User Feedback or Explicit Feedback)：用户交互地进行查询结果标注
- **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

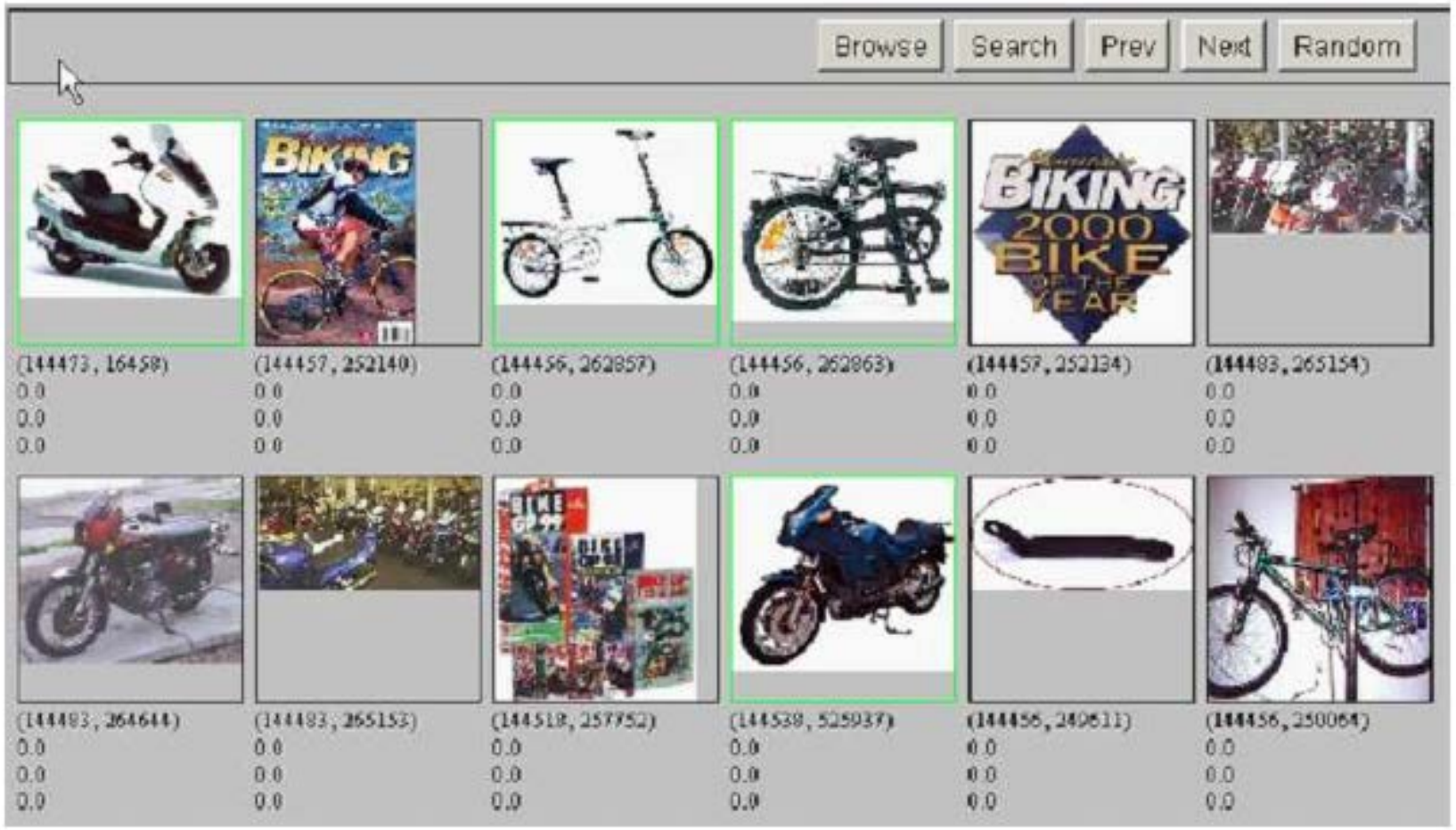
相关反馈举例-1


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


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

相关反馈后再次检索的结果

相关反馈举例-2

初始查询：[new space satellite applications]

初始查询的检索结果：

	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

		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

Rocchio算法核心概念:质心

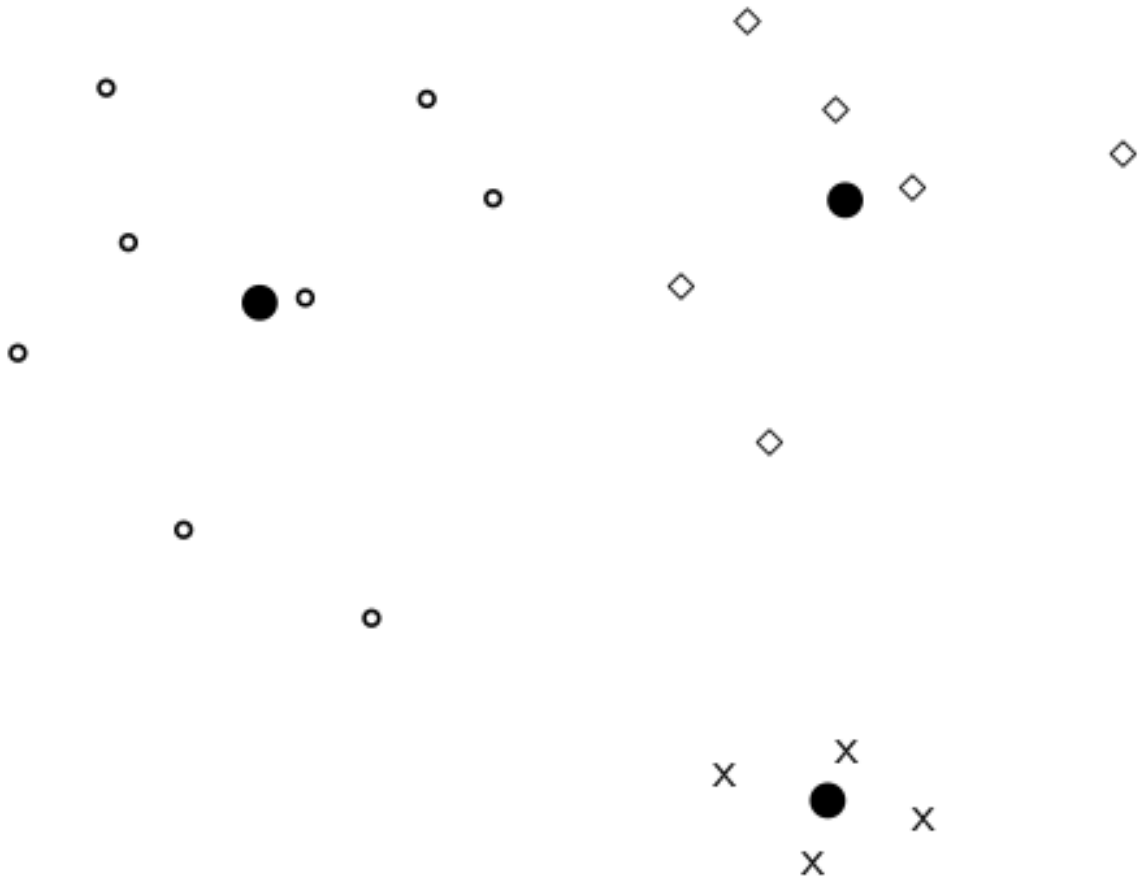
- 质心的定义

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

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

- 质心是一系列点的中心

质心示例



Rocchio算法原理

- 最优查询定义

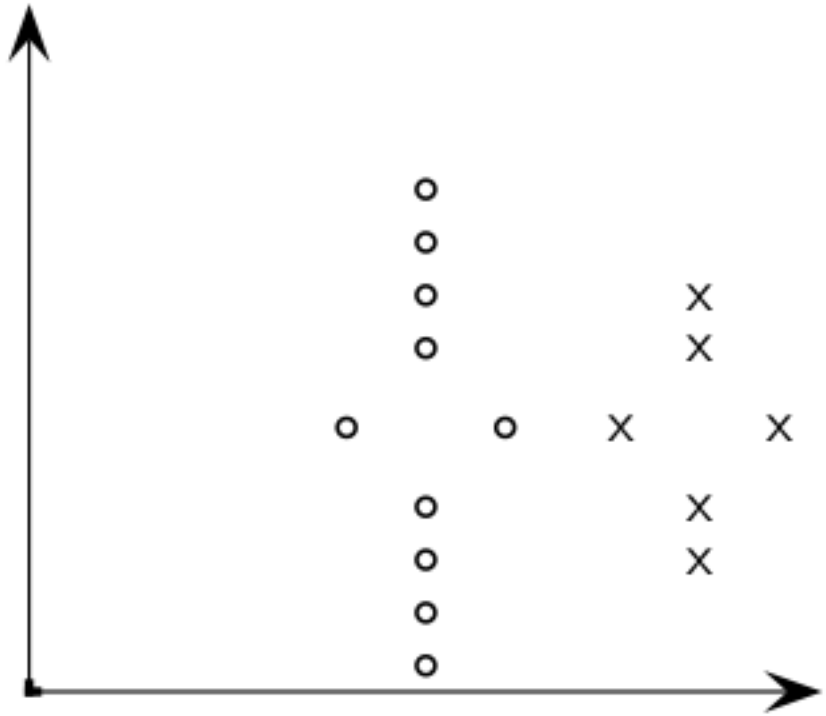
$$\vec{q}_{opt} = \arg\max_{\vec{q}}[\mathrm{sim}(\vec{q}, \mu(D_r)) - \mathrm{sim}(\vec{q}, \mu(D_{nr}))]$$

D_r：相关文档集；D_{nr}：不相关文档集

- 上述公式的意图是 $\vec{q}_{opt}$ 与相关文档相似度最大且同时与不相关文档相似度最小
- 基于余弦相似度，可以将上式改写为：

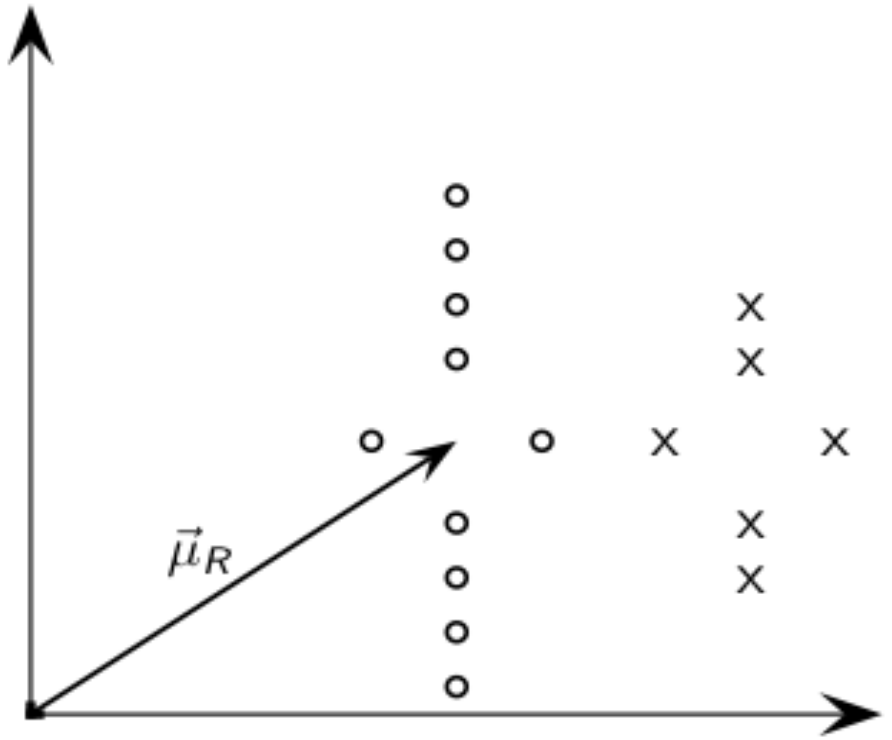
$$\vec{q}_{opt} = \mu(D_r) + [\mu(D_r) - \mu(D_{nr})]$$

举例



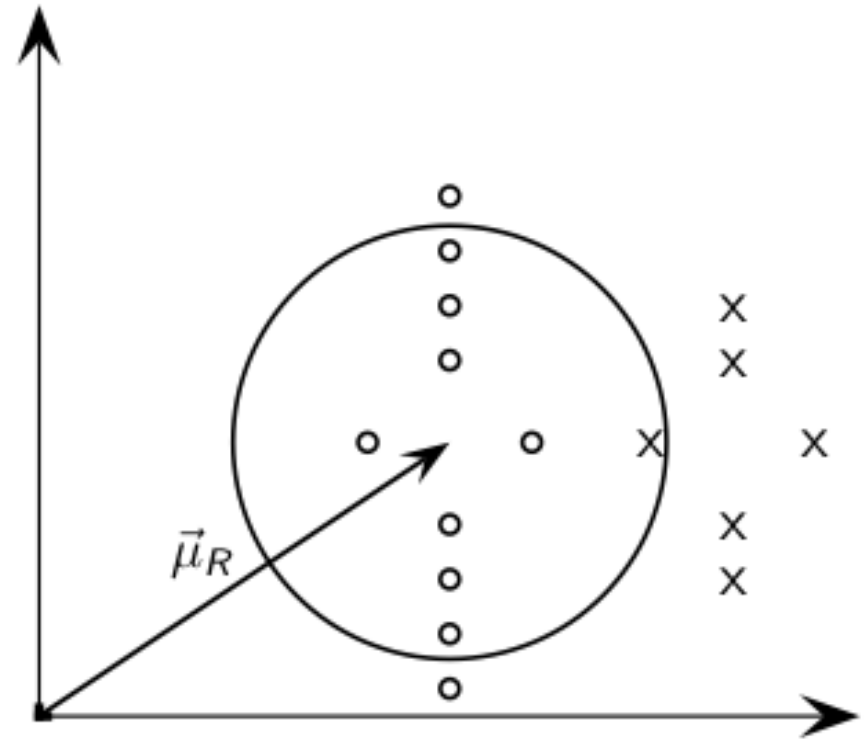
圆形点: 相关文档, 叉叉点: 不相关文档

Rocchio算法原理图示



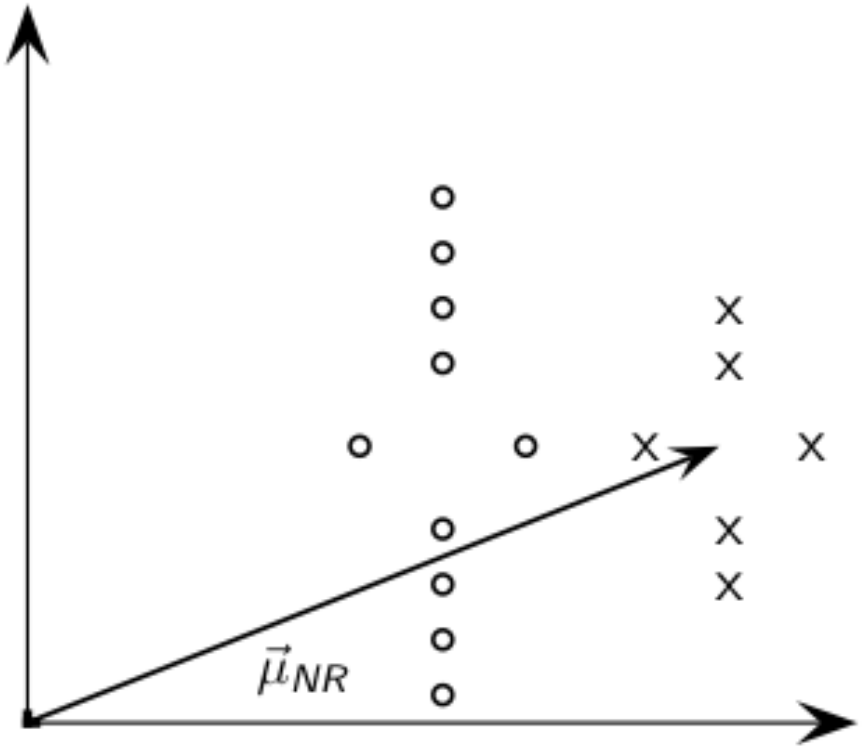
$\vec{\mu}_R$ ：相关文档的质心

Rocchio算法原理图示



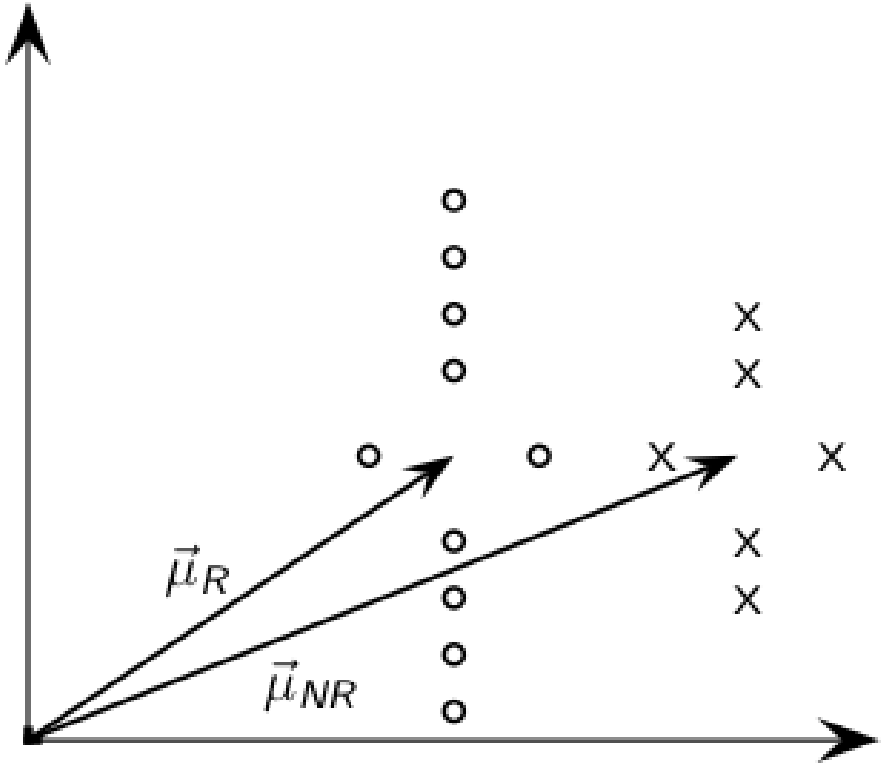
$\vec{\mu}_R$ 不能将相关/不相关文档分开

Rocchio算法原理图示

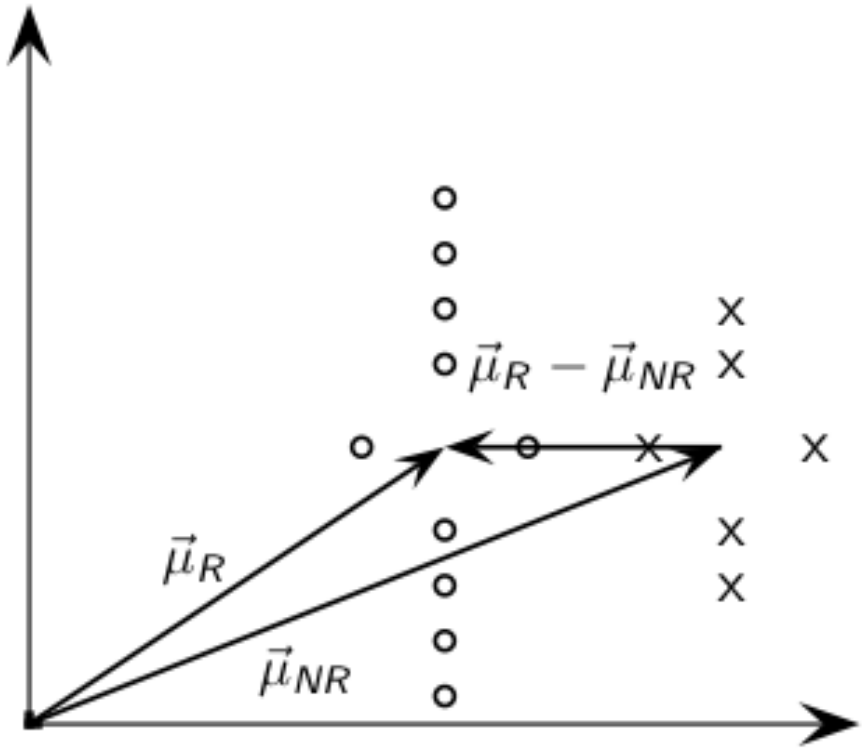


$\vec{\mu}_{NR}$：不相关文档的质心

Rocchio算法原理图示

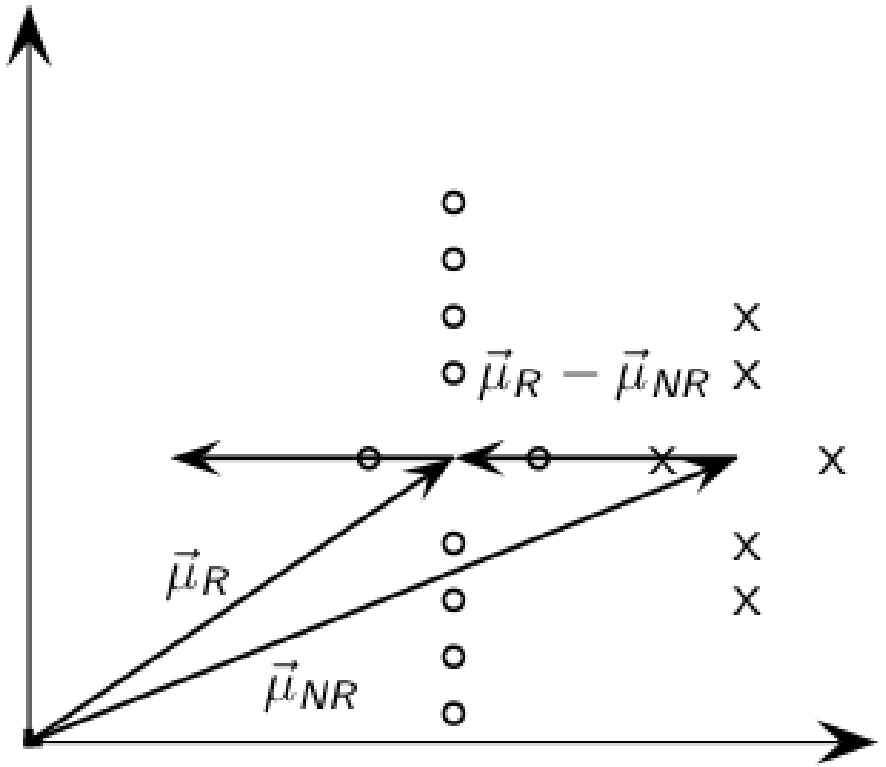


Rocchio算法原理图示



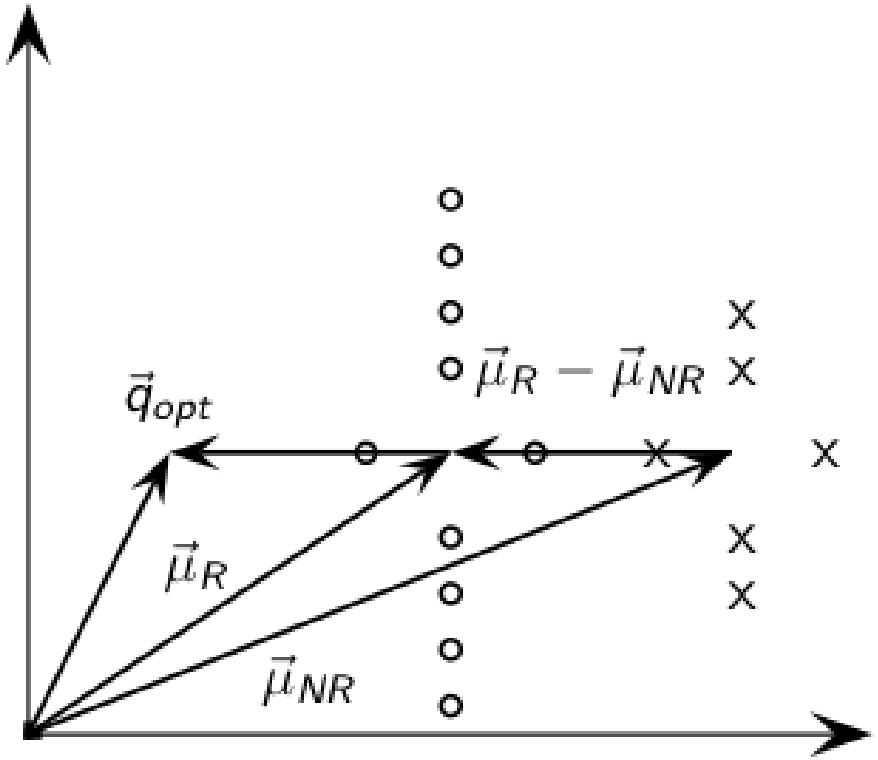
$\vec{\mu}_R$ - $\vec{\mu}_{NR}$：差异向量

Rocchio算法原理图示



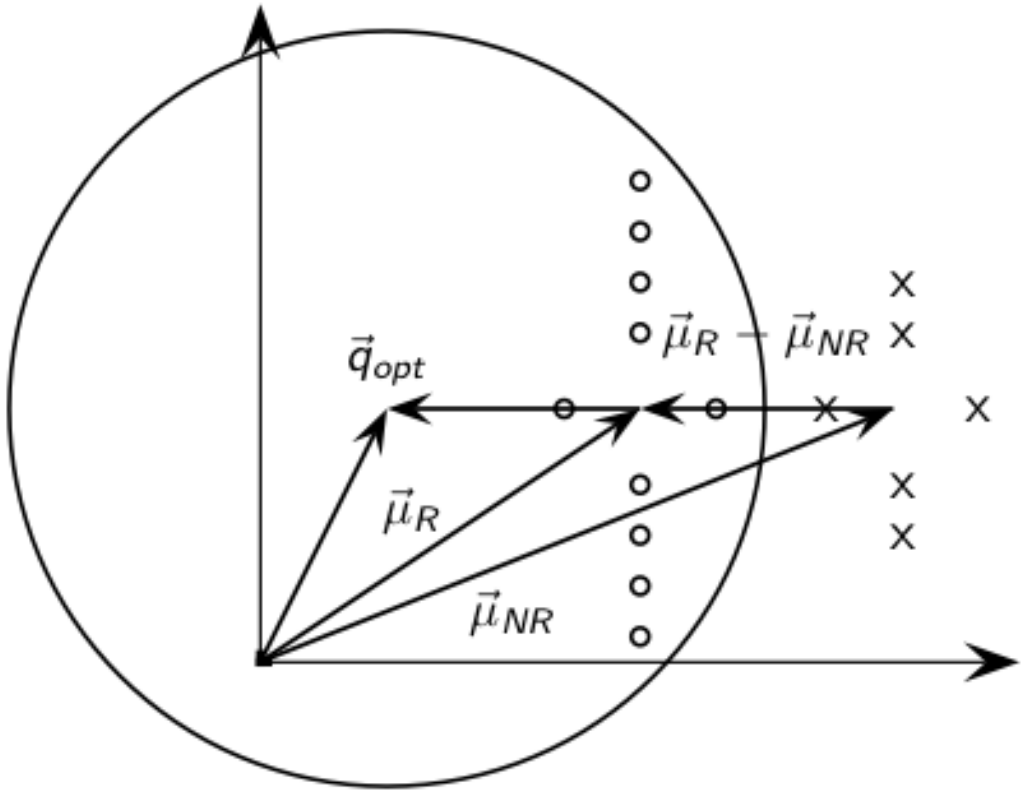
$\vec{\mu}_R$ 加上差异向量

Rocchio算法原理图示



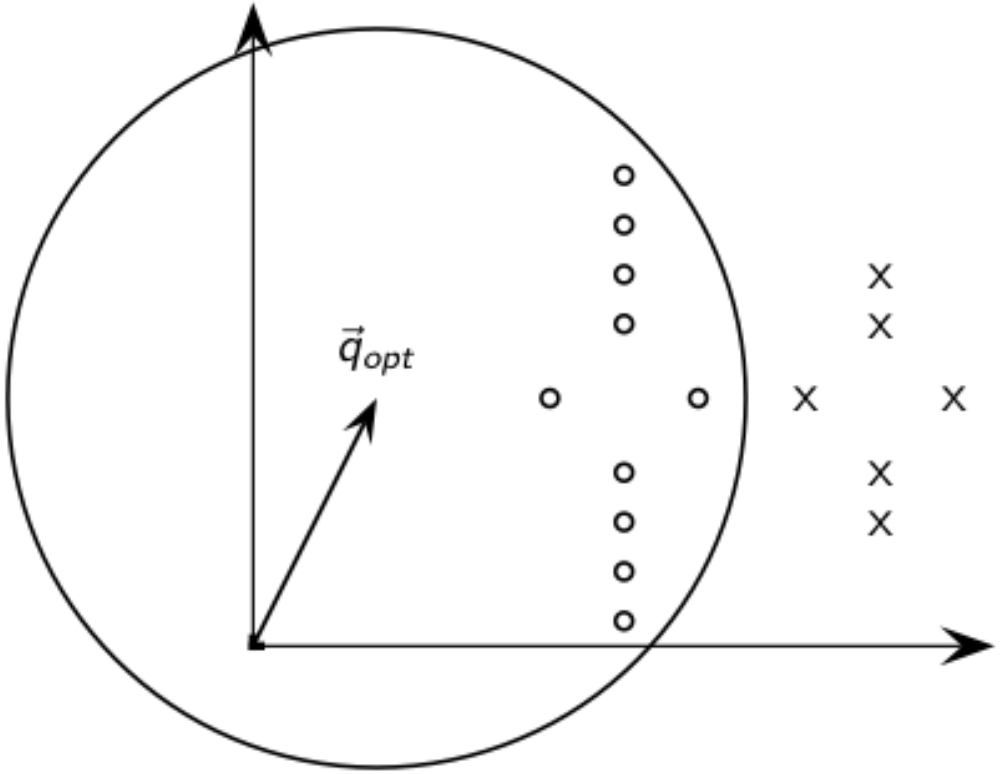
得到 $\vec{q}_{opt}$

Rocchio算法原理图示



$\vec{q}_{opt}$ 能够将相关/不相关文档完美地分开

Rocchio算法原理图示



$\vec{q}_{opt}$ 能够将相关/不相关文档完美地分开

Rocchio 算法

- 实际使用的优化查询确定方法

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

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

- 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

正反馈 vs. 负反馈

- **正（负）反馈：**指用户对相关文档（不相关文档）的标记和反馈
- 正反馈价值往往大于负反馈
- 因此可以通过设置 $\beta = 0.75$, $\gamma = 0.25$来给正反馈更大的权重
- 很多系统甚至只允许正反馈，即γ=0

相关反馈起作用的前提条件

- **并非什么时候相关反馈都能有效地提高召回率**
- **前提条件1：** 用户所构建的初始查询在一定程度上与需求文档相关，即：用户了解文档集词汇表
- **前提条件2：** 相关文档之间非常相似，相关文档和不相关文档之间的相似度很低
- 即所有相关文档都紧密聚集在某个prototype周围

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
- 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

相关反馈的评价

- 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进行对比
- 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，相关反馈是用户需要时间最少的方法

课堂思考

- Web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
- 应用相关反馈之后返回的某些文档的原因不易理解

隐式相关反馈

- **基本思想：**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，这些主要是个性化信息检索(Personalized IR)的内容

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)

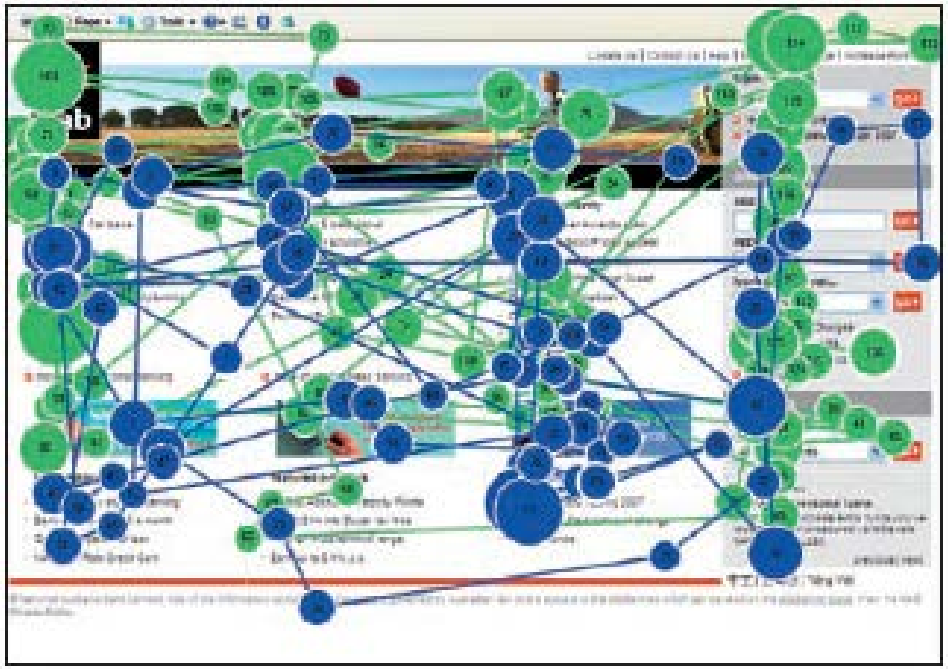
FIELD	VALUE
User ID	1162742023015
Time stamp	06/Nov/2006:00:01:35
Query terms	嫁给警察的理由
URL	http://bbs.cixi.cn/dispbbs.asp?Star=4&boardid=46&id=346721&page=1
Page number	1
Rank	7
Anchor text	姑娘们，你们愿意<FONT color=#cc0033>嫁给警察</FONT>吗？[慈溪社区]

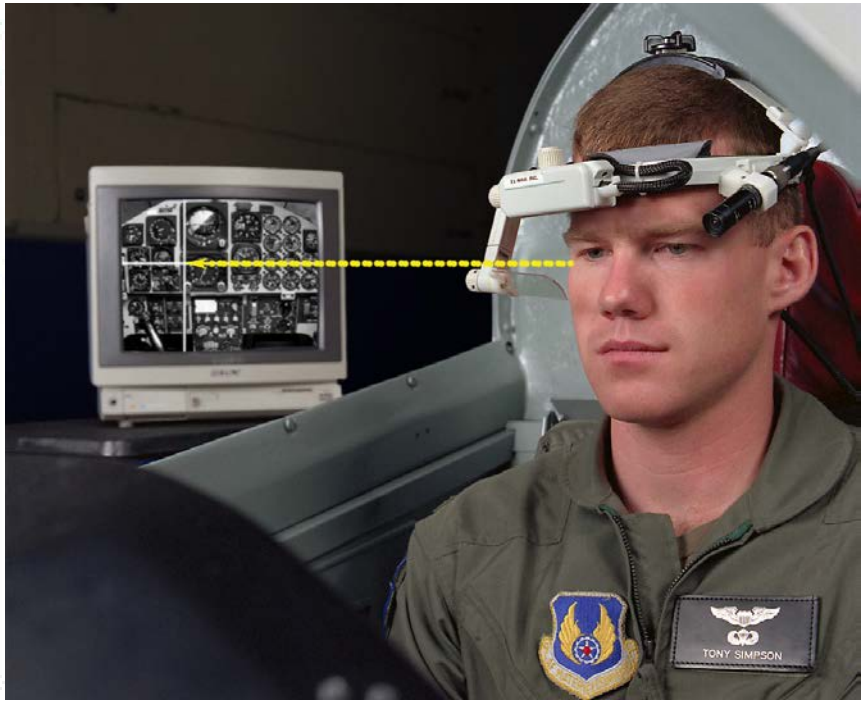
眼球动作(通过鼠标轨迹模拟)

Baidu

Google

关于Eye tracking





隐式相关反馈分析

- 用户行为某种程度上反映出用户的兴趣，因此具有合理性
- **优点：**
  - 不需要用户显式参与，减轻用户负担
- **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)

- **基本思想：** 自动进行用户无关的相关反馈
- **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 (如采用Rocchio算法)
- 平均来讲效果不错
- 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

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

- 实验中的伪相关反馈方法对查询只增加了20个词项 (使用Rocchio算法将增加更多的词项)
- 上述结果表明，伪相关反馈在平均意义上说是有效的方法

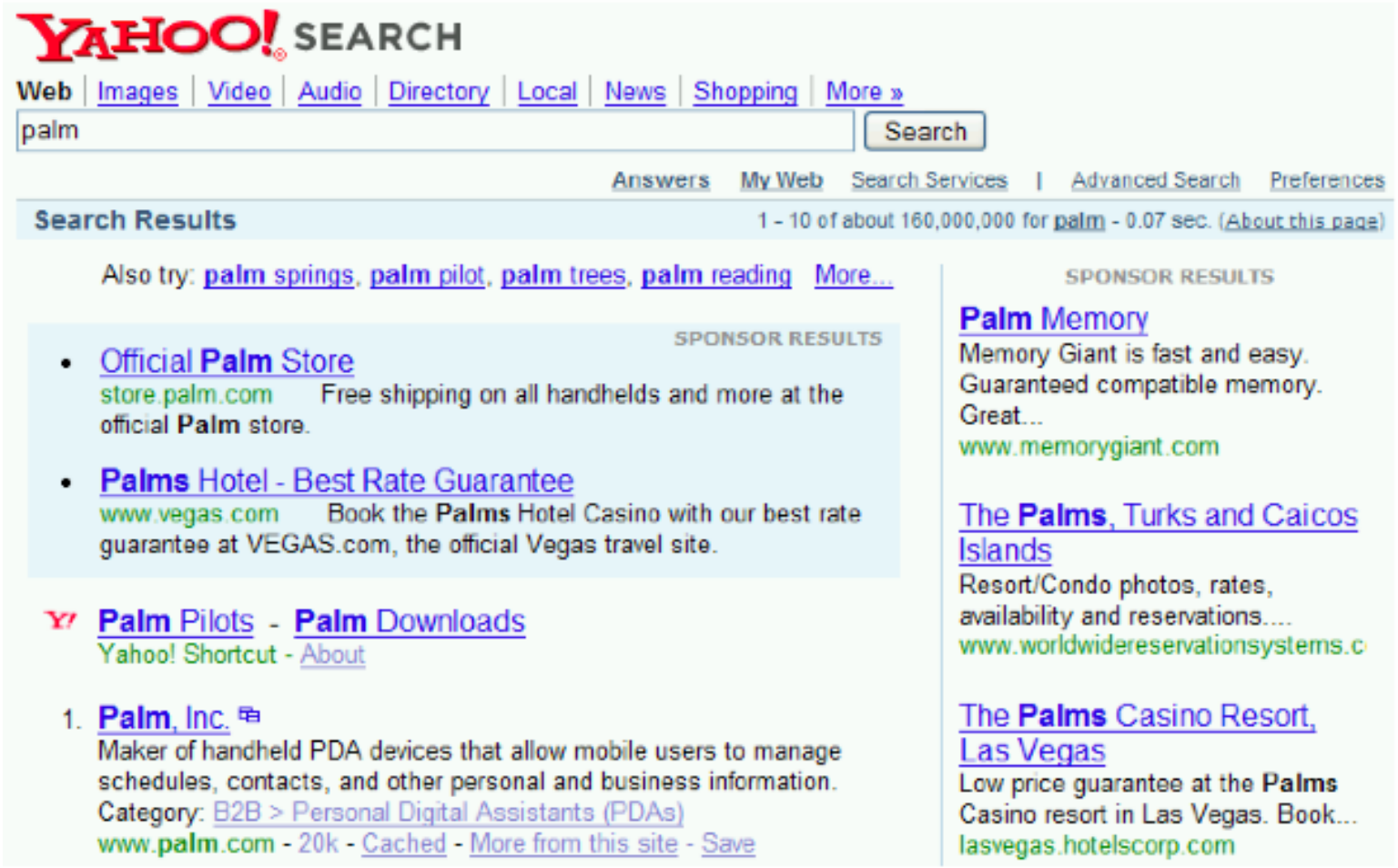
伪相关反馈分析

- **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

查询扩展(Query expansion)

- **基本思想：**通过对查询词或短语添加补充信息，提高检索召回率
- **主要途径：**基于一些全局的资源进行查询扩展，这些资源与查询无关，包括：同义词或近义词词典(thesaurus、wordnet等)
- **查询等价类构建：**
  - 人工构建
  - 自动构建
  - 基于查询日志挖掘

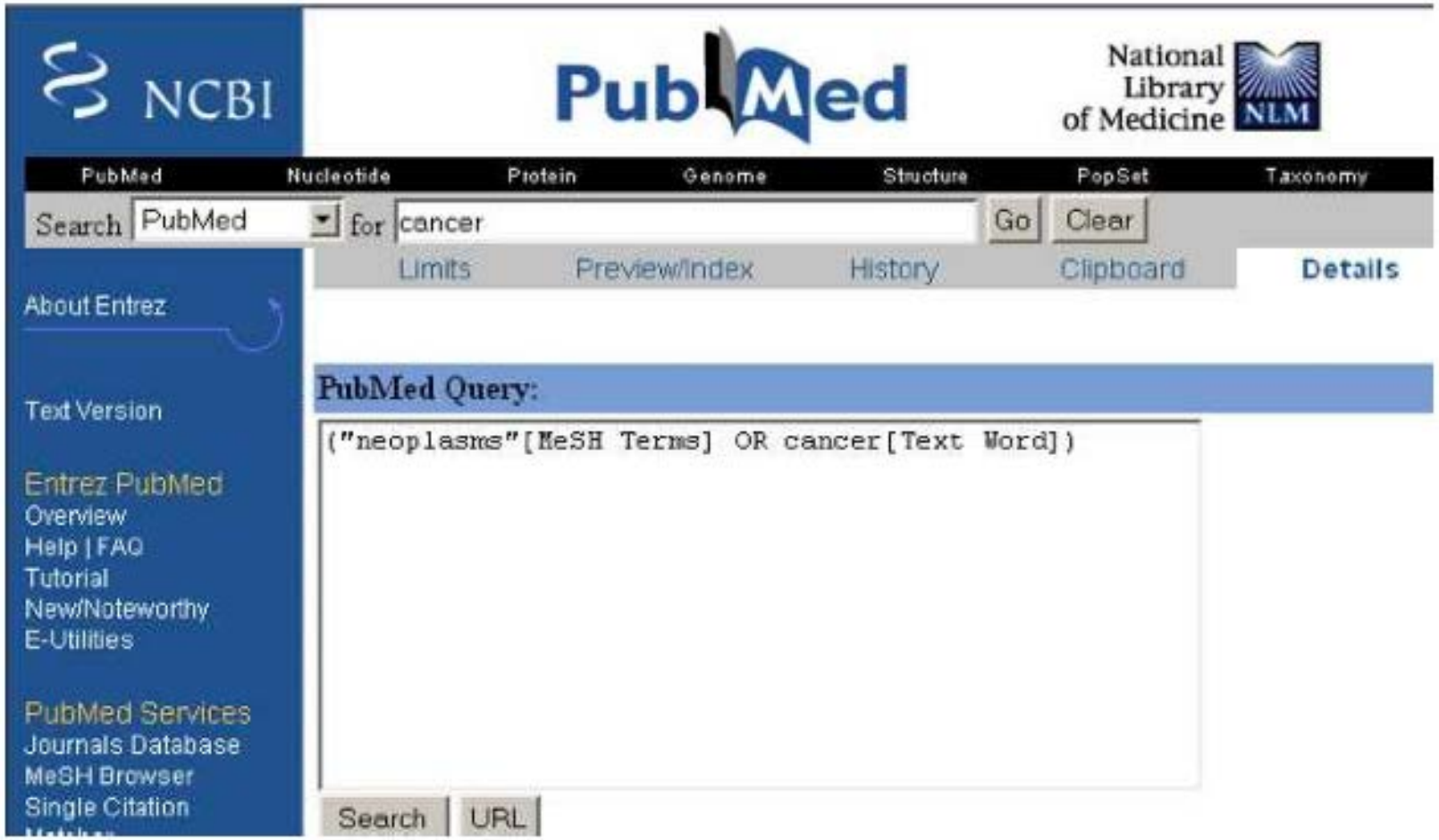
查询扩展的例子


YAHOO! SEARCH
Web | Images | Video | Audio | Directory | Local | News | Shopping | More »
palm
Search
Answers My Web Search Services | Advanced Search Preferences
Search Results
1 - 10 of about 160,000,000 for palm - 0.07 sec. (About this page)
Also try: palm springs, palm pilot, palm trees, palm reading More...
SPONSOR RESULTS
Official Palm Store
store.palm.com Free shipping on all handhelds and more at the official Palm store.
Palms Hotel - Best Rate Guarantee
www.vegas.com Book the Palms Hotel Casino with our best rate guarantee at VEGAS.com, the official Vegas travel site.
Palm Pilots - Palm Downloads
Yahoo! Shortcut - About
1. Palm, Inc.
Maker of handheld PDA devices that allow mobile users to manage schedules, contacts, and other personal and business information.
Category: B2B > Personal Digital Assistants (PDAs)
www.palm.com - 20k - Cached - More from this site - Save
SPONSOR RESULTS
Palm Memory
Memory Giant is fast and easy. Guaranteed compatible memory. Great...
www.memorygiant.com
The Palms, Turks and Caicos Islands
Resort/Condo photos, rates, availability and reservations....
www.worldwidereservationsystems.c
The Palms Casino Resort, Las Vegas
Low price guarantee at the Palms Casino resort in Las Vegas. Book...
lasvegas.hotelscorp.com

基于同(近)义词词典的查询扩展

- **具体方法：** 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
- 广泛应用于特定领域(如科学、工程领域)的搜索引擎中
- **前提条件：** 有一个好的同（近）义词词典

基于人工词典的扩展样例: PubMed



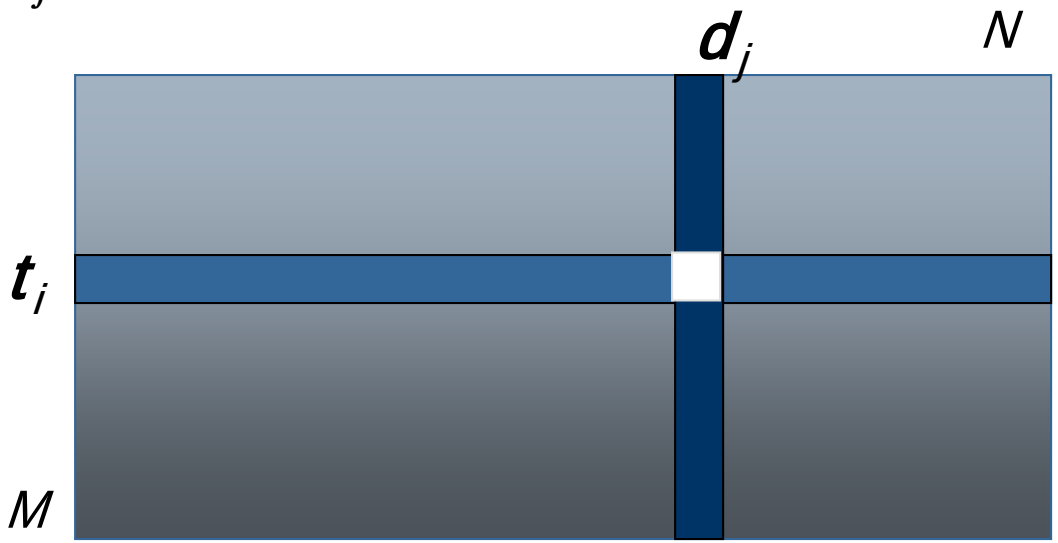
PubMed：著名的医学文献数据库

同(近)义词词典的自动构建

- **基本思想：** 通过计算文档集中词语之间的相似度来自动生成同(近)义词词典
- **相似度度量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
- **相似度度量2:** 两个词，如果它们同某些词具有某种给定的语法关系的话，那么它们类似
  - apples 和pears与harvest, peel, eat, prepare具有一样的动宾关系，因此 apples 和pears肯定彼此类似
- 共现关系一般更加鲁棒，而语法关系可能更加精确

基于共现的同(近)义词典构造

- 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



如果矩阵A是0/1矩阵，那么C的每一项是什么？

- 对每个t_i, 选择C中对应行向量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
- 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

查询扩展分析

- **优点：**
  - 能够提高检索的召回率
  - 其效果容易为用户所理解
- **缺点：**
  - 有时候可能会明显降低正确率
  - 总体上不如相关反馈成功

参考资料

- 《信息检索导论》第9章
- http://ifnlp.org/ir
  - Salton and Buckley 1990 (原始的相关反馈论文)
  - Spink, Jansen, Ozmultu 2000: Relevance feedback at Excite
  - Schütze 1998: Automatic word sense discrimination (接扫了一个简单的同义词自动构造方法)

课后作业

- 见课程网页：

  http://10.76.3.31