现代信息检索

Modern Information Retrieval

第10讲 相关反馈及查询扩展

Relevance Feedback & Query Expansion

*改编自”An introduction to Information retrieval”网上公开的课件，地址 http://nlp.stanford.edu/IR-book/

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

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【实际上也能提高正确率】

- 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，当然有可能提高正确率。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（排名靠后的）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

关于相关反馈和查询扩展

- 相关反馈的本质是将检索返回的文档的相关性判定(不同的判定来源：人工或非人工)作为返回信息，希望提升检索效果（召回率和正确率）。
  - 相关反馈常常用于查询扩展，所以提到相关反馈往往默认为有查询扩展

- 而查询扩展的最初含义是对查询进行扩充，比如：car → car automobile，近年来越来越向查询重构(query reformulation or refinement)偏移，即现在的查询扩展是指对原有查询进行修改。
  - 基于相关反馈(局部方法的代表)进行查询扩展/重构
  - 基于本讲的全局方法进行查询扩展/重构
  - 局部和全局方法相结合的方法(如LCA，本讲没有介绍)

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
- 用户或系统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

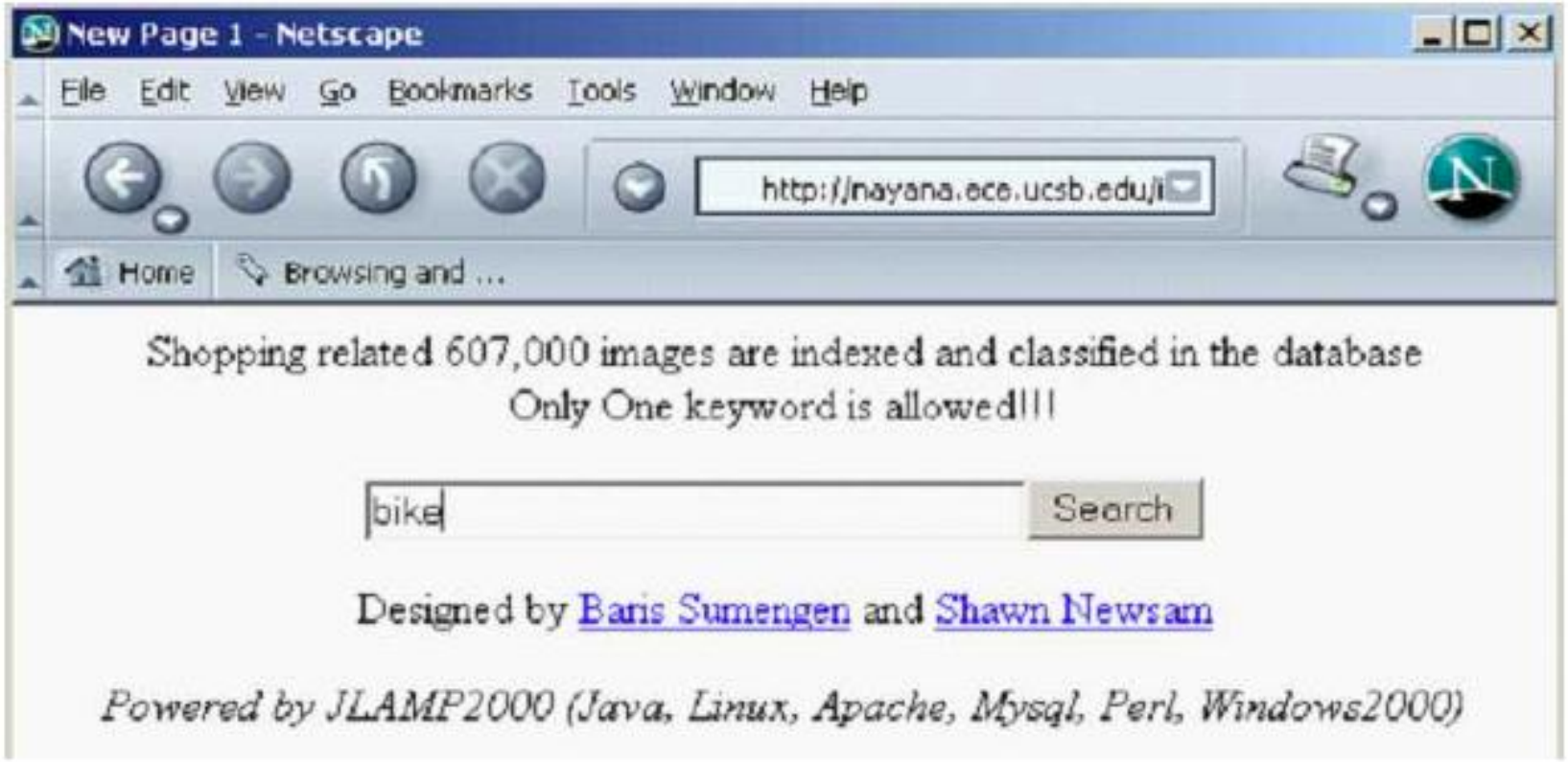
相关反馈分类：根据反馈信息来源

- 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。

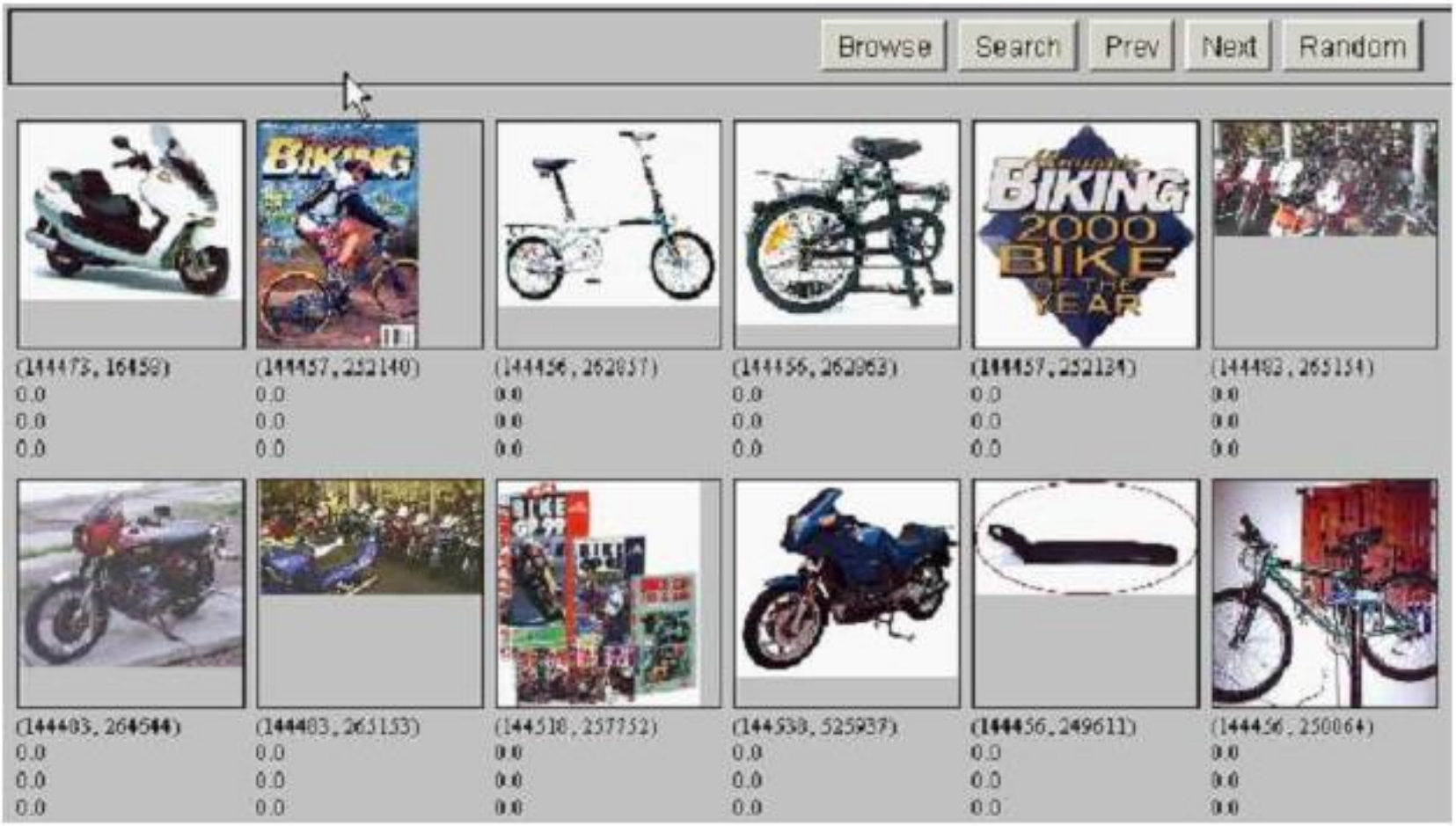
相关反馈

- 相关反馈可以循环若干次
- 将介绍三个不同的(用户)相关反馈的例子

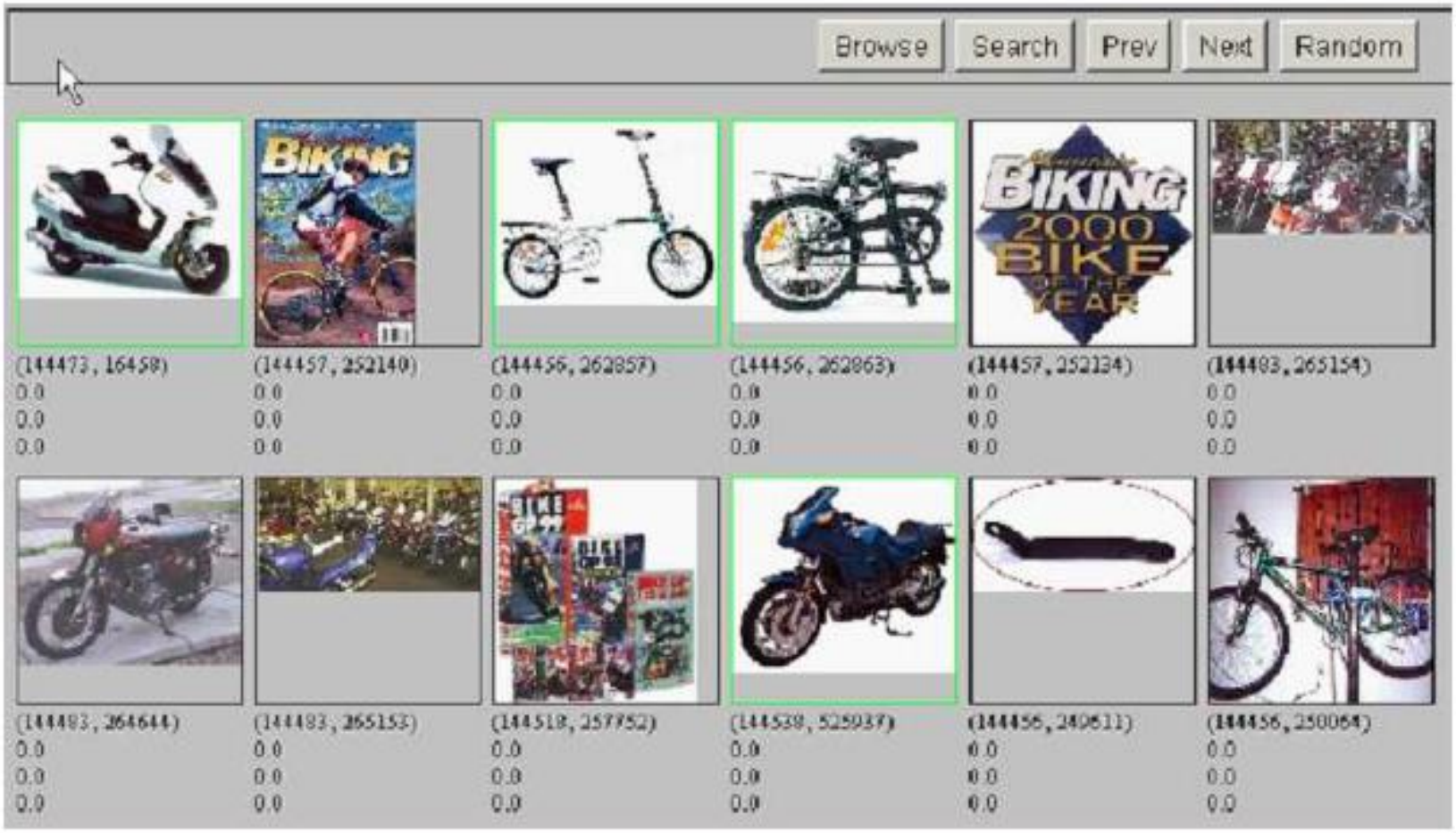
例1：图像搜索的例子


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

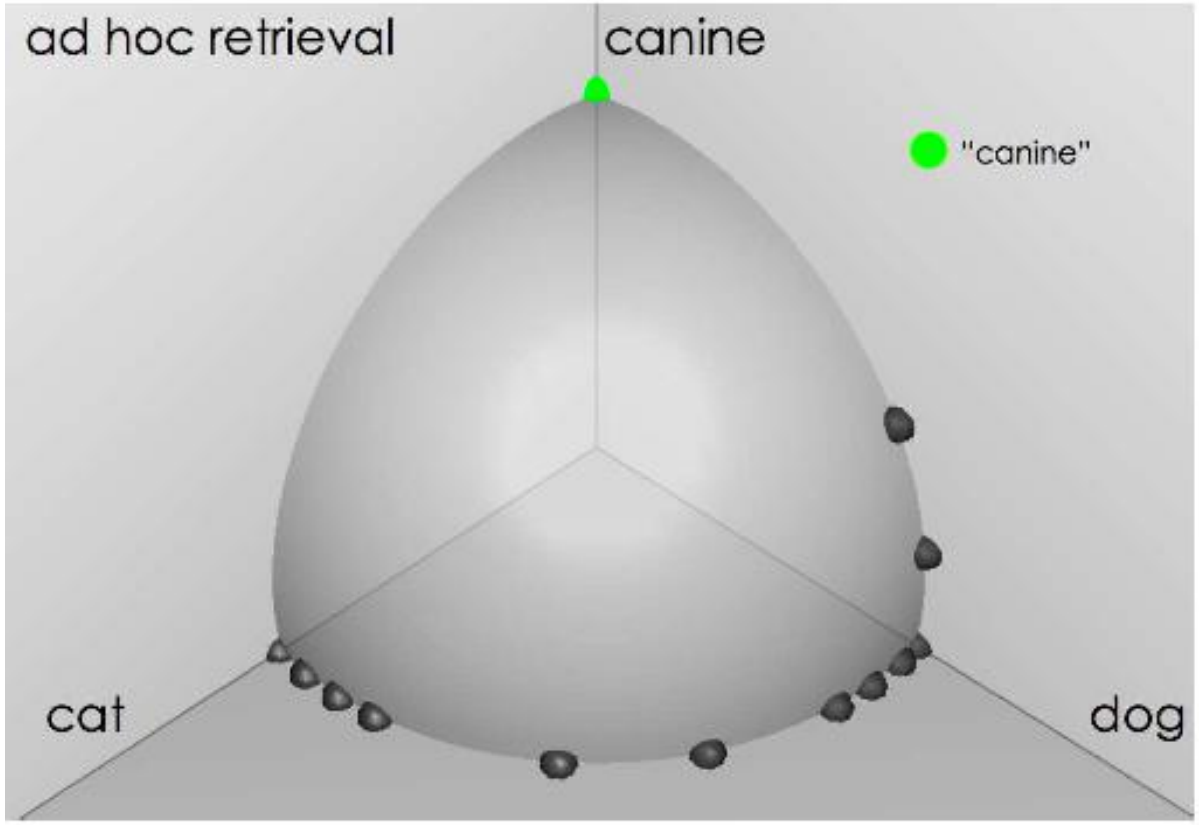

Browse
Search
Prev
Next
Random
BIKING
BIKING 2000 BIKE OF THE YEAR

用户反馈: 选择相关结果


Browse
Search
Prev
Next
Random
BIKING
BIKING 2000 BIKE OF THE YEAR
(144473, 16458)
(144457, 252140)
(144456, 262857)
(144456, 262863)
(144457, 252134)
(144483, 265154)
(144483, 264644)
(144483, 265153)
(144518, 257752)
(144539, 525037)
(144456, 240611)
(144456, 250064)

相关反馈后再次检索的结果

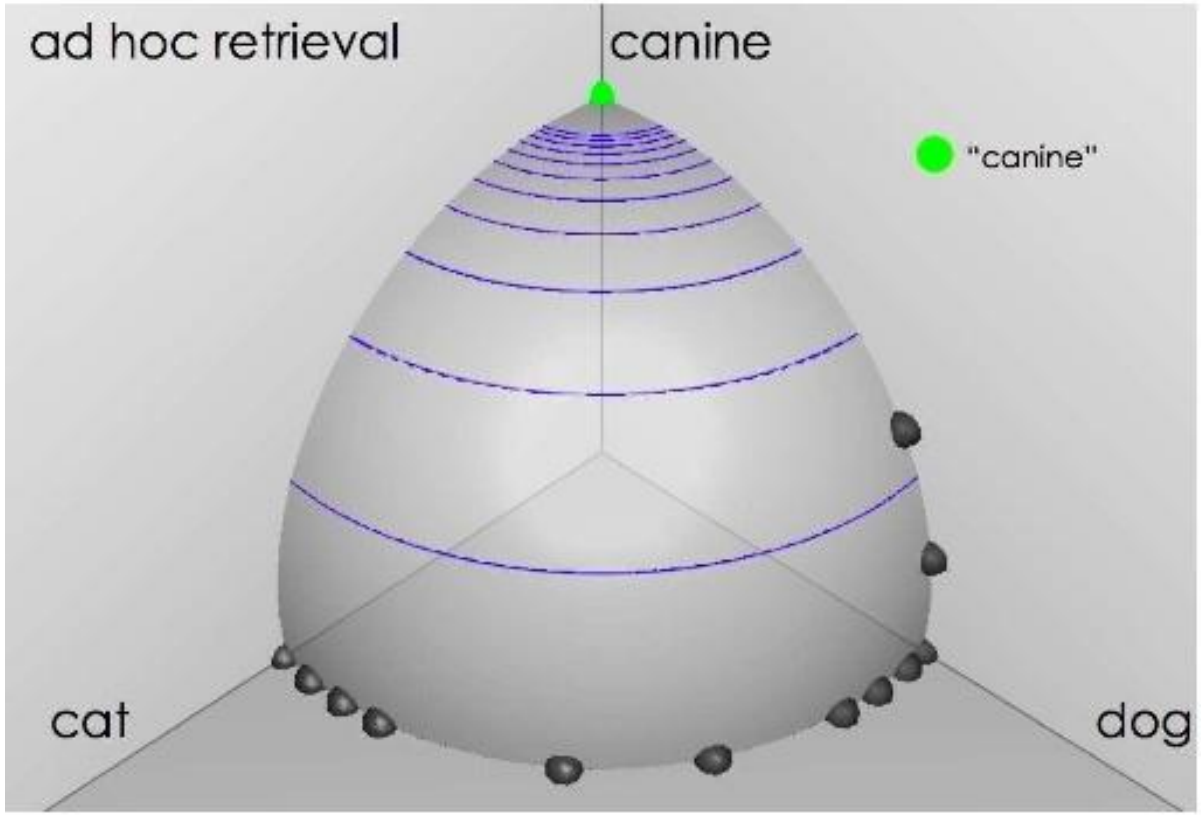
向量空间的例子: 查询 “canine” (1)



Source:

Fernando Díaz

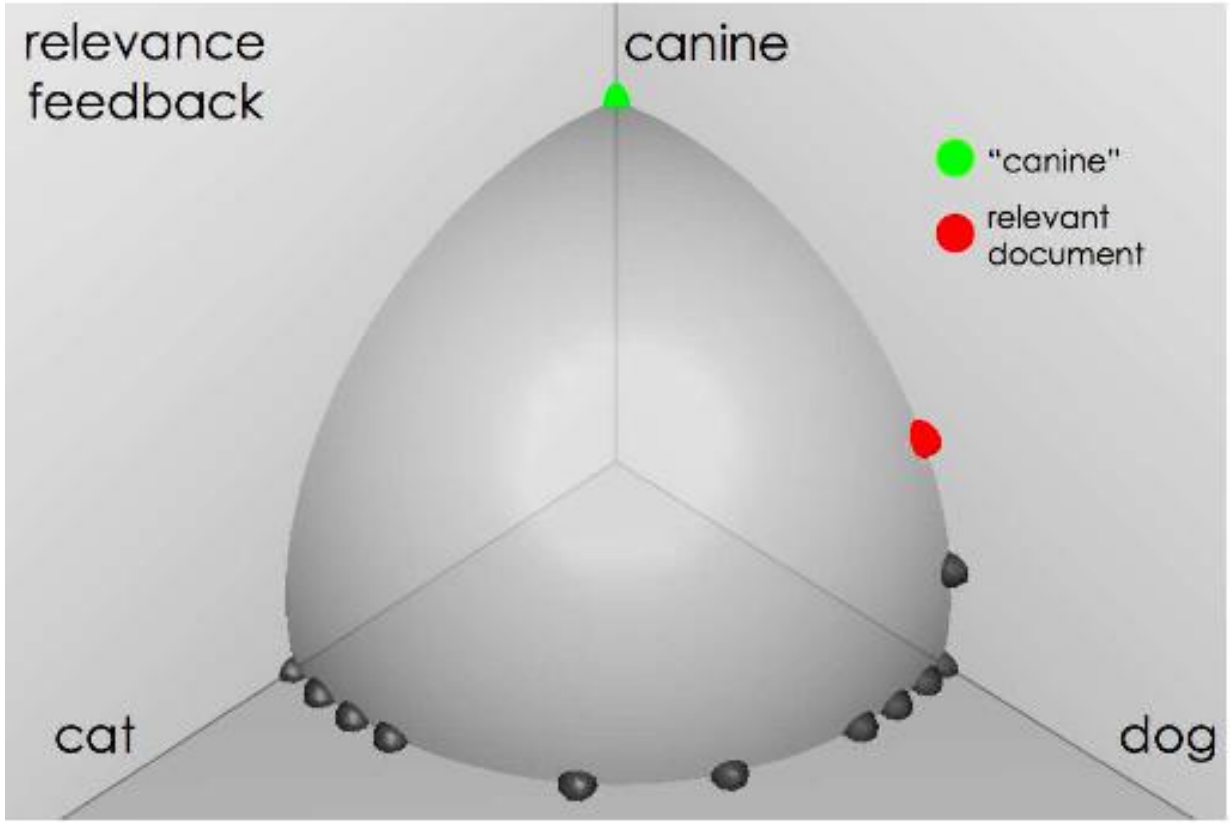
文档和查询 “canine” 的相似度



Source:

Fernando Díaz

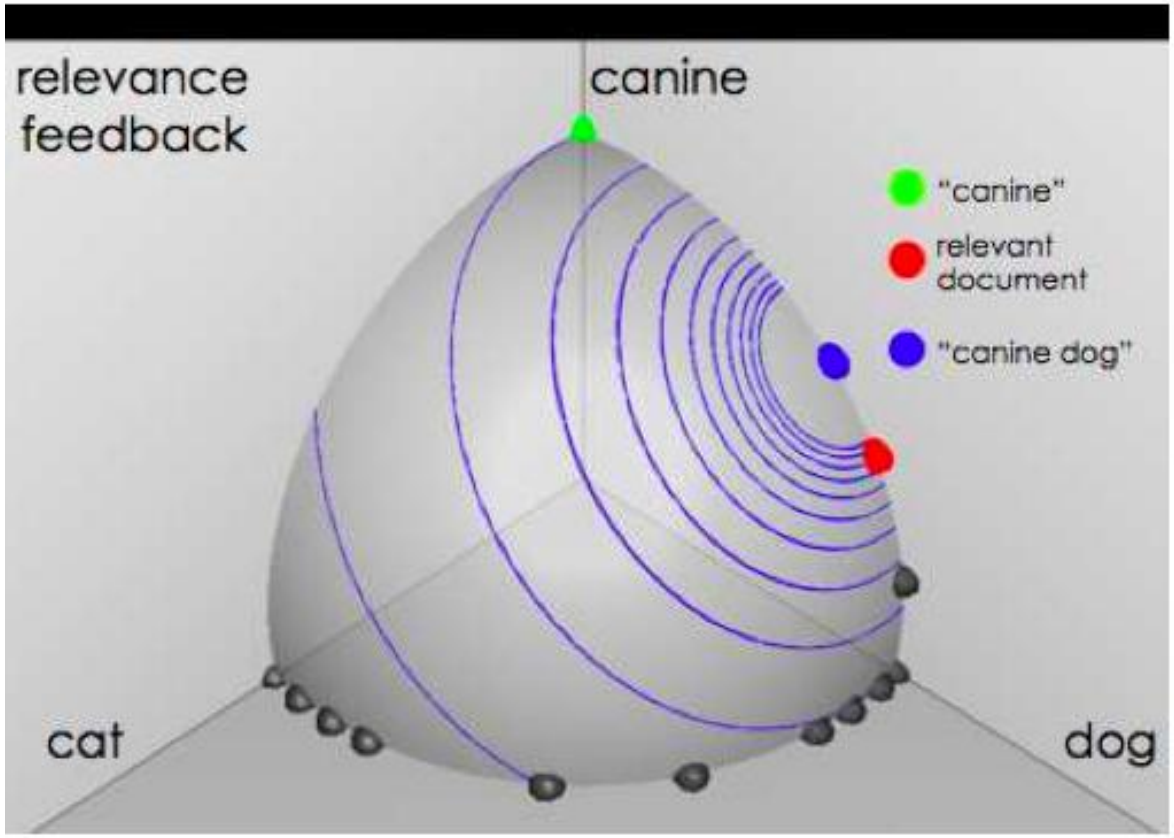
用户反馈: 选择相关文档



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 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

用户将前两篇文档标记为相关 "+". 排名第8的是一篇未标记相关文档

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

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

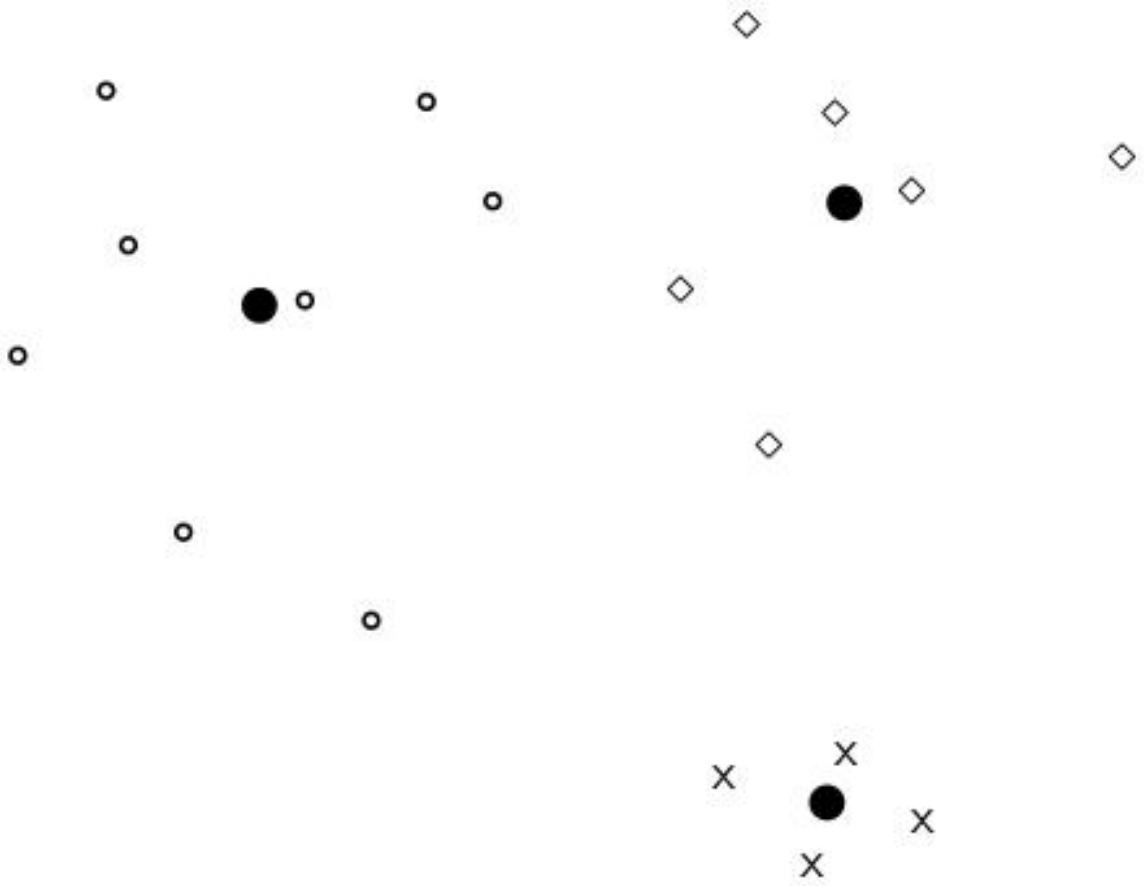
相关反馈中的核心概念：矩心（Centroid）

- 矩心是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矩心

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

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

矩心的例子



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

- 上述公式的意图是 $\vec{q}_{opt}$ 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:

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

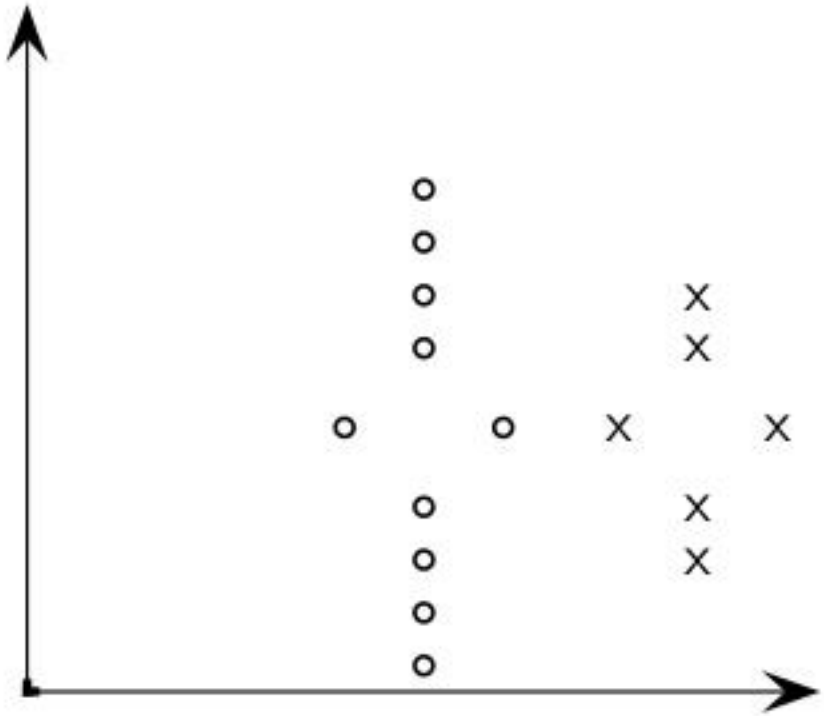
Rocchio算法

- 最优查询向量为：

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

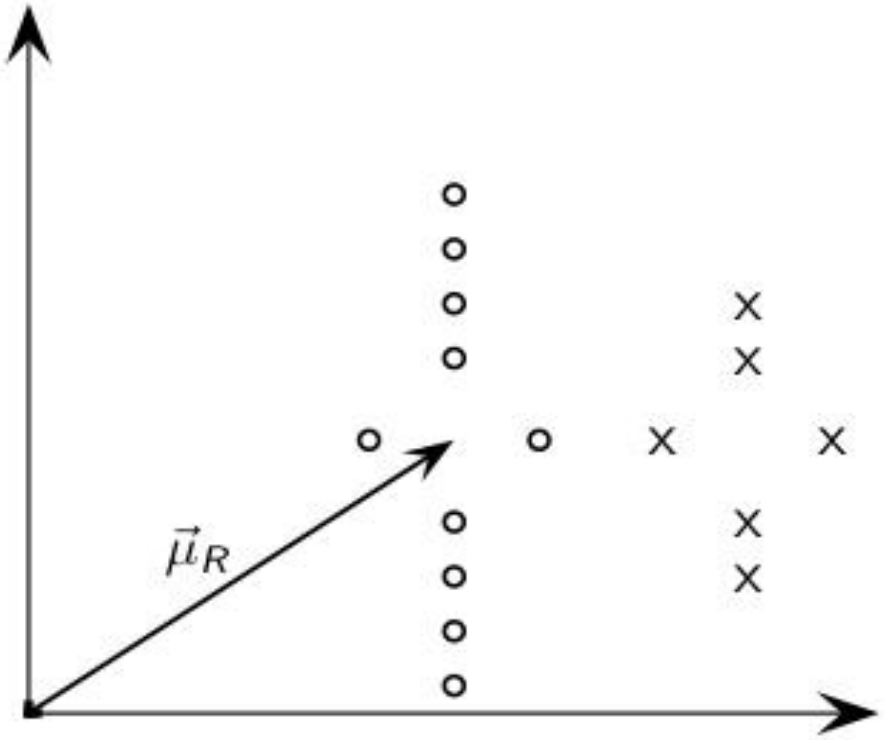
- 即将相关文档的矩心移动一个量，该量为相关文档矩心和不相关文档的差异量

课堂练习: 计算Rocchio向量



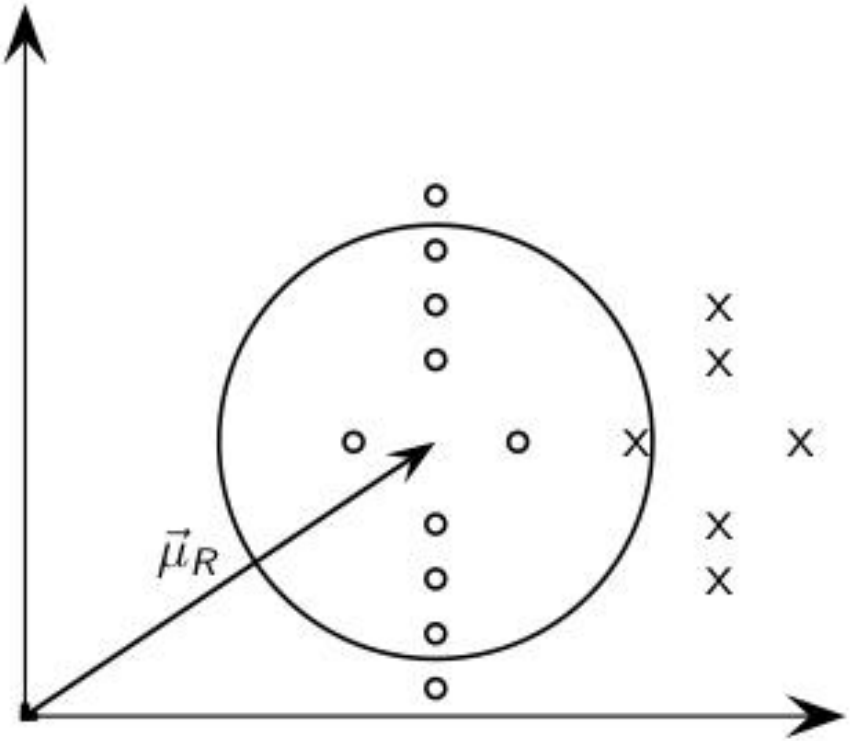
圆形点: 相关文档, 叉叉点: 不相关文档

Rocchio算法图示



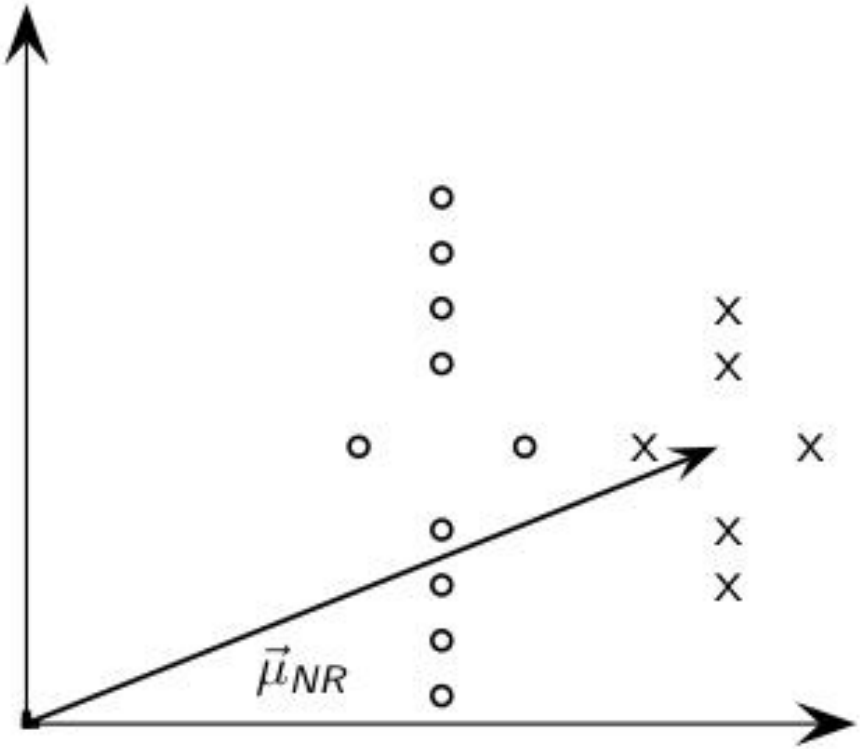
$\vec{\mu}_R$：相关文档的矩心

Rocchio算法图示



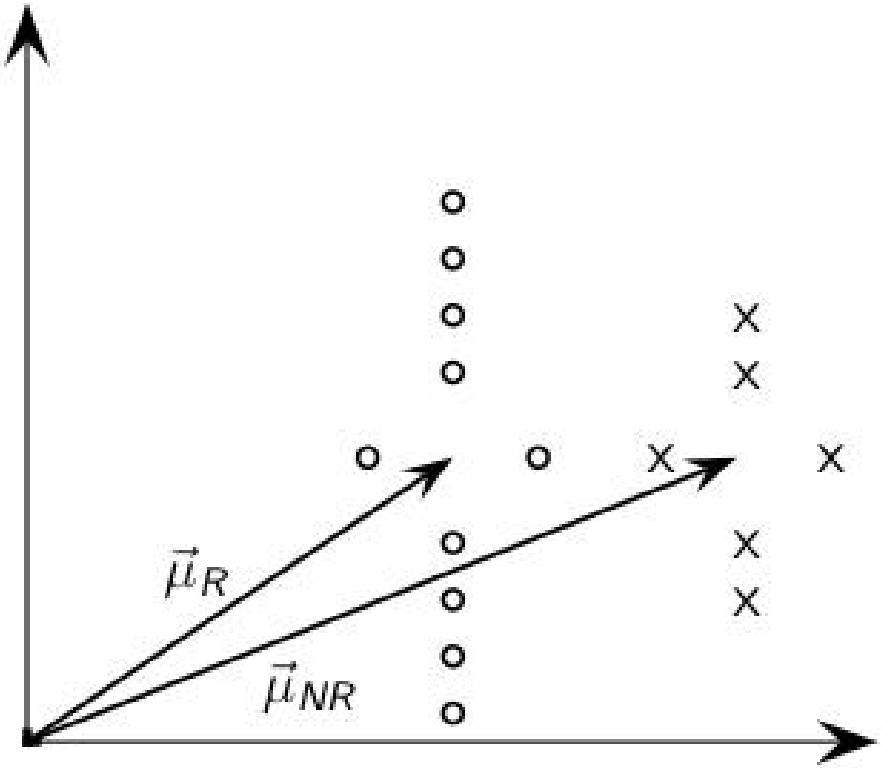
$\vec{\mu}_R$ 不能将相关/不相关文档分开

Rocchio算法图示

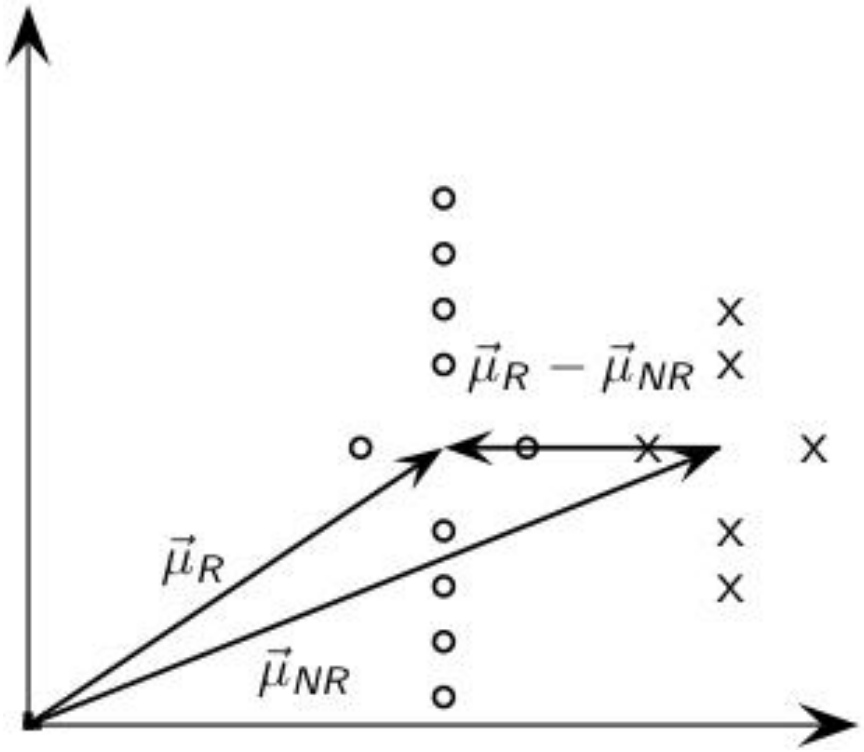


$\vec{\mu}_{NR}$: 不相关文档的矩心

Rocchio算法图示

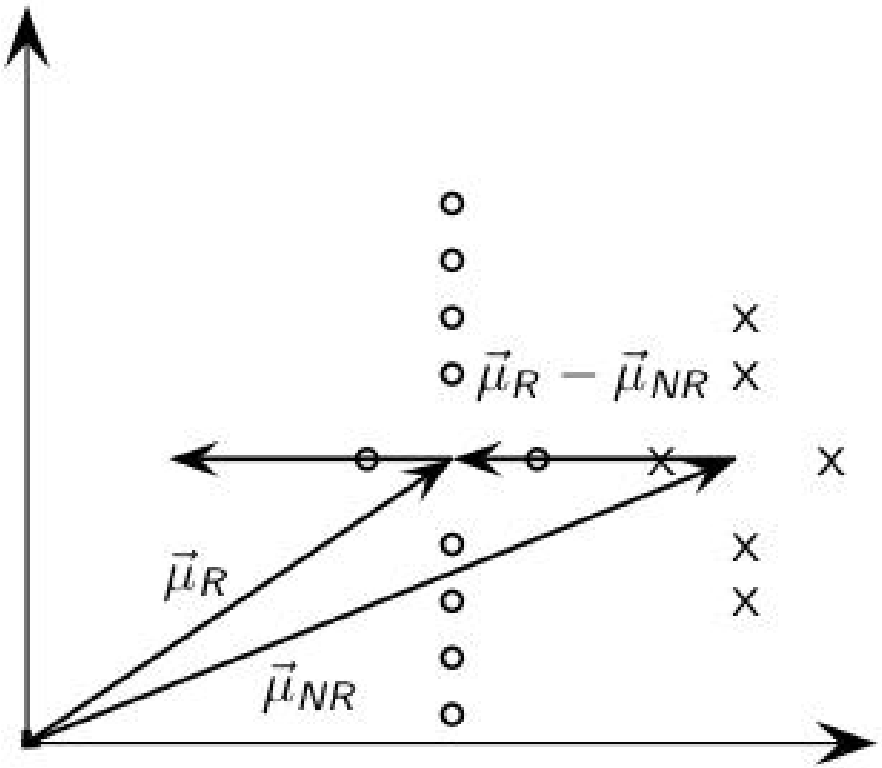


Rocchio’ 算法图示



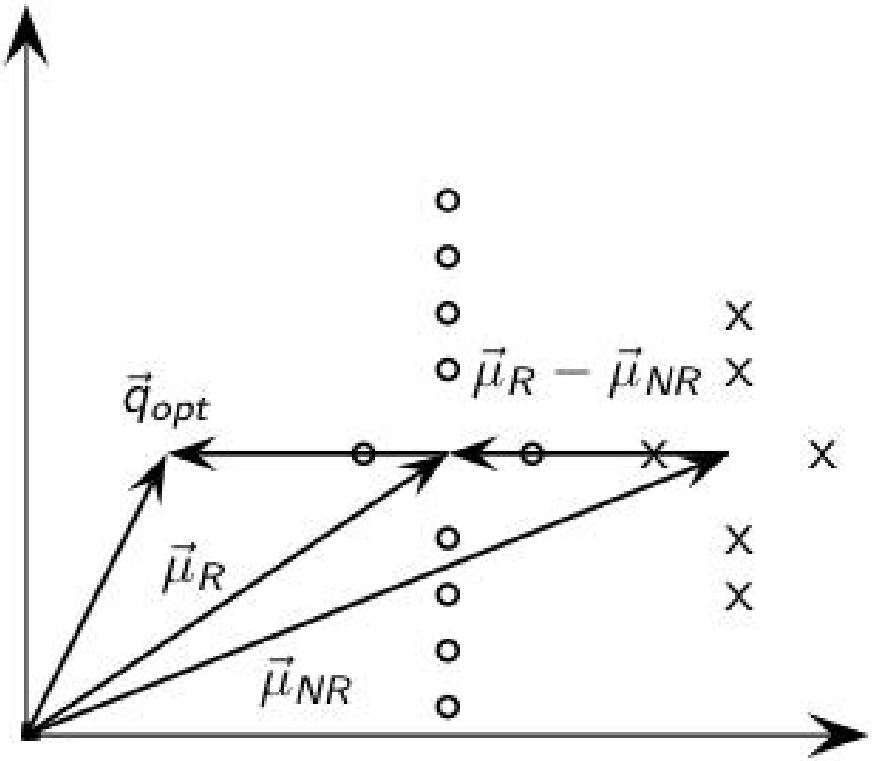
$\vec{\mu}_R$ - $\vec{\mu}_{NR}$: 差异向量

Rocchio算法图示



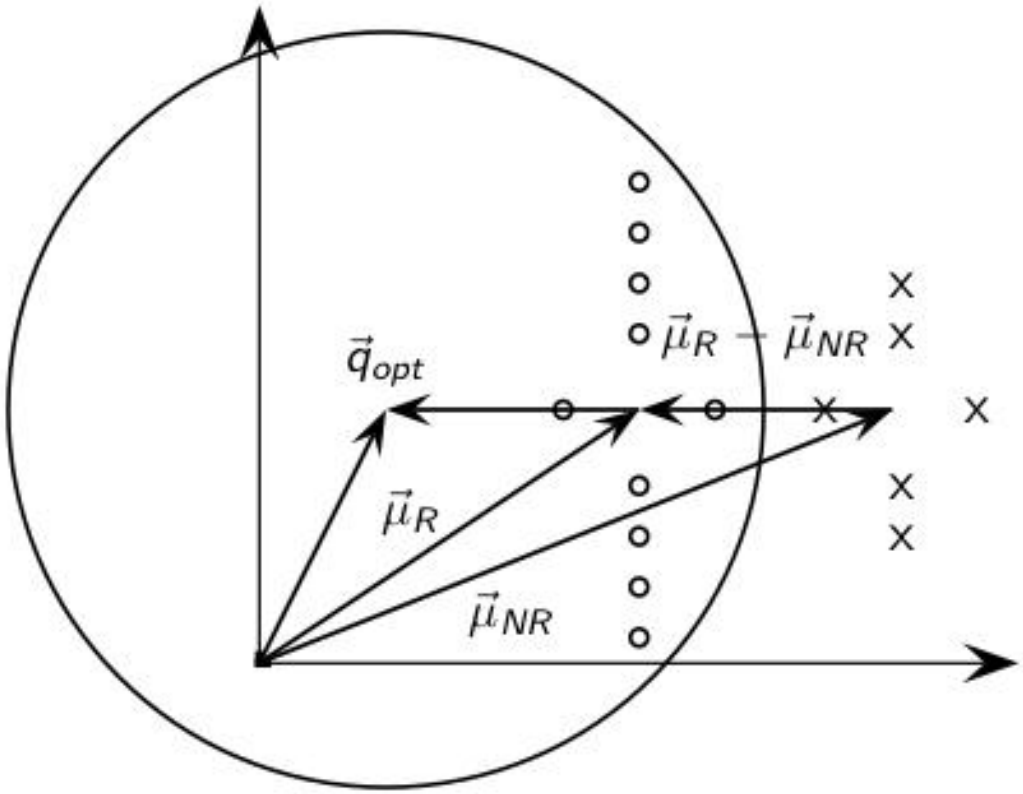
$\vec{\mu}_R$ 加上差异向量

Rocchio算法图示



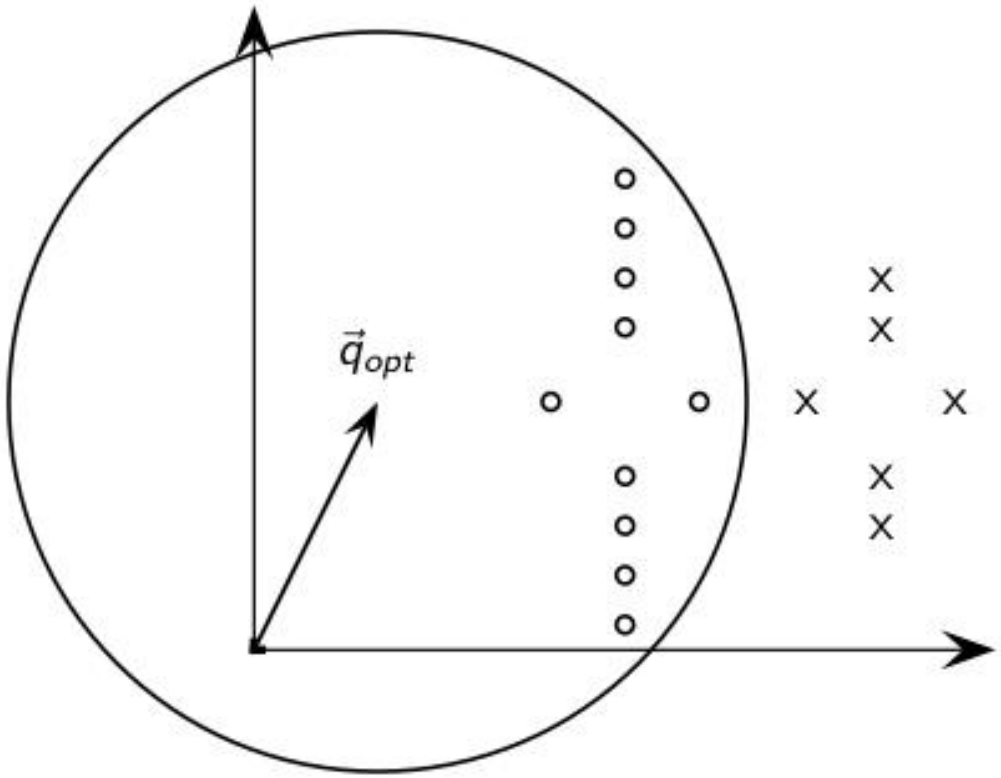
得到 $\vec{q}_{opt}$

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

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: 修改后的查询向量; q_0: 原始查询向量;
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γ=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
  - 注意：伪相关反馈的评价不需要去掉经过判断的文档
- 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?

用户相关反馈存在的问题

- 用户相关反馈开销很大
  - 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

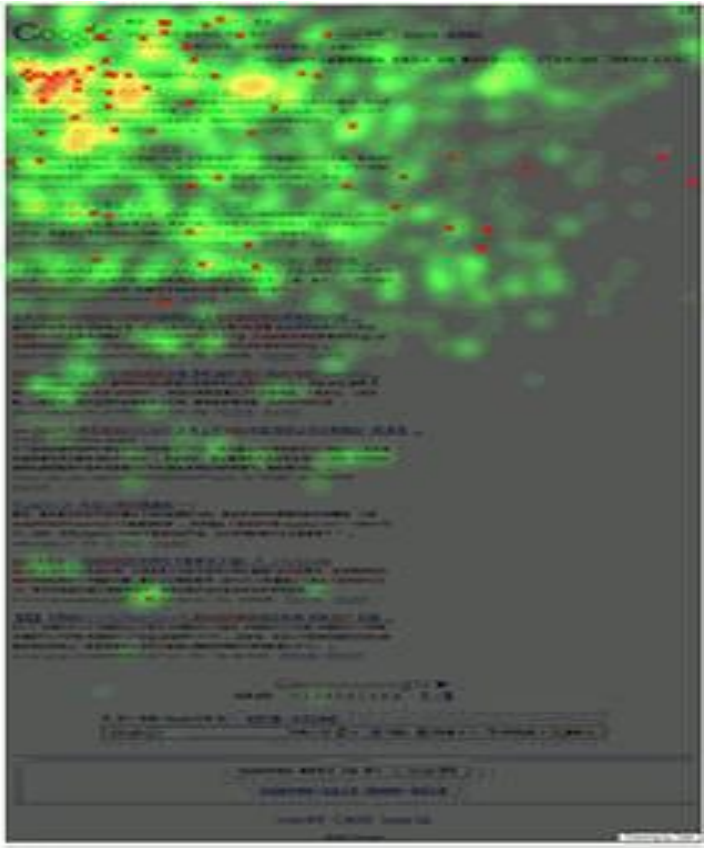
- 鼠标键盘动作：
  - 点击链接、加入收藏夹、拷贝粘贴、停留、翻页等等
- 用户眼球动作
  - Eye tracking可以跟踪用户的眼球动作
  - 拉近、拉远、瞟、凝视、往某个方向转

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<FONT color=#cc0033>嫁给警察</FONT>吗？[慈溪社区]

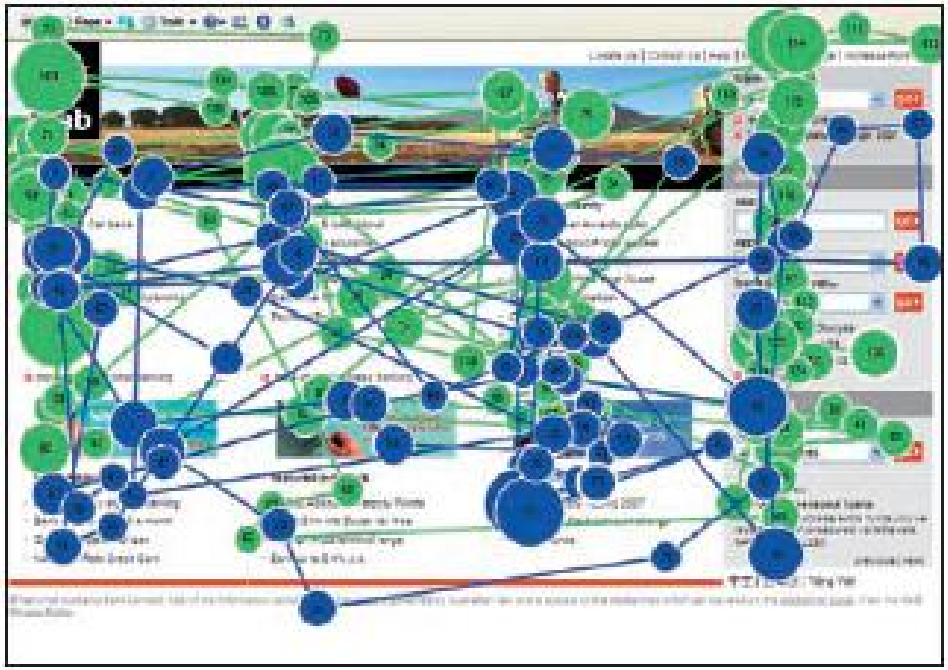
眼球动作(通过鼠标轨迹模拟)

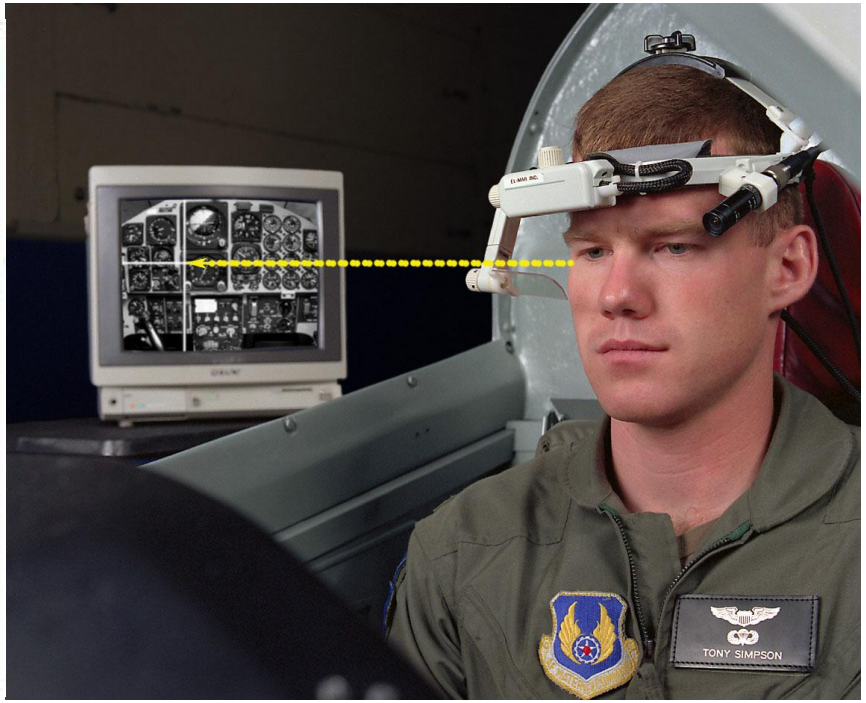
Baidu



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*)

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
lnc.ltc	3210
lnc.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

推荐阅读：Yang Xu, Gareth Jones, Bin Wang, Query Dependent Pseudo Relevance Feedback Based on Wikipedia，SIGIR2009

相关反馈小结

- 文档选择：从检索结果中选择相关或不相关文档。用户显式/隐式，或者系统假设。
- 词项选择：从相关不相关文档中选择需要处理的词项
- 查询扩展/重构：修改原始查询

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

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

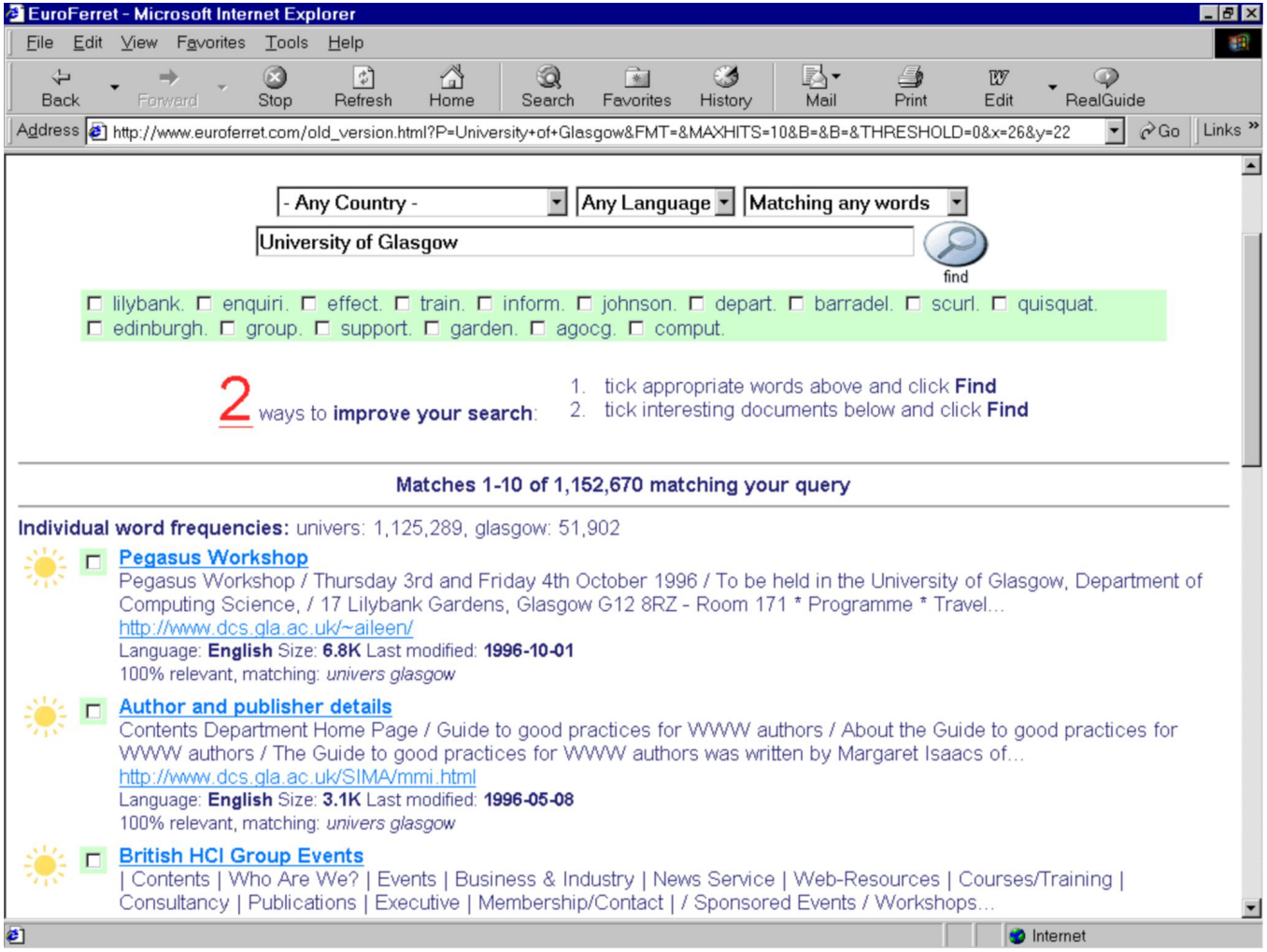
查询扩展的例子


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

查询扩展的方法

- 基于相关反馈的查询扩展(已经介绍)
- 人工词典法：通过人工构建的同(近)义词词典 (人工编辑人员维护的词典，如 PubMed)来扩展原始查询
- 自动词典法：自动导出的同(近)义词词典 (比如，基于词语的共现统计信息)
- 其他外部资源法：比如基于查询日志挖掘出查询等价类(Web上很普遍，比如上面的 “palm” 例子)

交互式查询扩展 (Interactive QE)



交互式QE

- 用户通常很懒
  - 不愿意为提交反馈意见作出额外劳动
  - 用户可能并不理解反馈/扩展的作用
- 用户提交的扩展词项并不一定有用
- Original query: Trump Iran Nuclear Deal
- Expanded terms: white house Schumer

white house Schumer是与查询相关的扩展词，但是与查询的联系并不明显，不一定能提高检索效果

基于词项相似度的查询扩展

- 基于候选词和原始查询词项共现（co-occurrences)的查询扩展
- 相似度指标：如果词项t_1在X个文档中出现，t_2 在Y个文档中出现，并且t_1、t_2在C个文档中共现 they co-occur in C documents:
  - Cosine (t1, t2) =C/sqrt(X*Y)
  - Dice (t1, t2) =2C/(X+Y)
  - Tanimoto (t1, t2) = C/(X+Y－C)
- 将反馈文档集中与原始查询词项最相似的k个候选词加入查询中
- 但是小语料集上的实验表明该方法无效 [Ferber 1996]
  - “Reason(s) of the failures are not really known.”
- 近年来在大语料上实现重复了上述结论

使用WordNet进行查询扩展

- WordNet 是一个在线英语词汇参考系统
  - 英语名词、动词、形容词和副词被组织成同义词集
  - 广泛使用于自然语言处理与人工智能 应用中
- 基于词汇相似性进行查询扩展
- 扩展词项通常都是原始查询词项的同义/近义词
- 例子：对于TREC查询“Scottish highland games”，使用WordNet可以得到以下扩展词项:
- Scotch, Gaelic, upland, hilly, mountainous, bet, gage, stake, punt etc.
- 经验表明使用WordNet对英语查询有效

查询扩展的优点

- 通常可以检索到更多的相关文档
  - 可以提高检索效果，例如Mean Average Precision (MAP)
- 例如:
  - TREC2005 Terabyte Track adhoc task
  - 使用 TF-IDF_R, MAP=0.3024
  - 使用 TF-IDF_R+Bo1, MAP=0.3428
- 统计测试表明MAP显著提高 (p-value=0.008169)

查询扩展可能的问题

- 在伪相关反馈的应用场景下，如果反馈集文档质量很差，会严重降低检索效果
  - 扩展词项很可能和查询无关
  - 反馈集中相关文档数量过少也会导致类似问题
- 可能会产生查询漂移(Query Drift)
  - 反馈文档相关，但是文档中的重要词项与查询无关
- 对于某些查询任务，例如主页搜索，由于相关文档总数非常少，查询扩展通常无效
  - 例如查询“国科大主页” 只有一个相关网址

例子：Adhoc检索任务

- TREC2005 Terabyte Track adhoc task
  - 返回尽量多的相关文档
  - 每一个查询均有大量相关文档
- 使用TF-IDF_R,无QE, MAP=0.3024
- 使用Rocccio QE+Bo1模型可以进一步提高检索结果
- β参数对结果有一定影响

β	*MAP*
0.1	*0.3303*
0.2	*0.3382*
0.3	*0.3432*
0.4	*0.3426*
0.5	*0.3413*
0.6	*0.3401*
0.7	*0.3384*
0.8	*0.3365*
0.9	*0.3347*
1.0	*0.3327*
free	*0.3428*

例子：主题提炼(Topic-distillation)检索任务

- TREC2003 topic distillation task
  - 要求返回关键资源
  - 每一个查询仅有少量相关文档
- 使用TF-IDF_R, MAP=0.0970
- 查询扩展会显著降低检索效果

β	*MAP*
0.1	*0.0607*
0.2	*0.0550*
0.3	*0.0524*
0.4	*0.0507*
0.5	*0.0501*
0.6	*0.0485*
0.7	*0.0478*
0.8	*0.0471*
0.9	*0.0469*
1.0	*0.0464*
free	*0.0497*

使用外部资源进行查询扩展(External QE)

- 在局域网检索、企业内部检索等语料较小的应用中，可以使用外部资源进行查询扩展
- 高质量的外部资源有望提供优质的扩展词项，从而提高检索效果
- [Kwok 2003]：将Google返回的靠前文档作为反馈集，可以显著提高检索效果
- 外部资源：优质资源，例如Wikipedia，或大语料集

选择性查询扩展(Selective QE)

- 在伪相关反馈应用场景，如果预测反馈集质量很低，则不再执行QE
  - 预测方法包括使用一些预测因子，例如扩展词IDF，反馈文档相关性评分，查询词质量评价因子等
  - 可以通过机器学习方法综合考虑多种预测因子
- 适用于对排名靠前文档查准率(early precision)有要求的任务

Giambattista Amati, Claudio Carpineto, Giovanni Romano: Query Difficulty, Robustness, and Selective Application of Query Expansion. ECIR 2004: 127-137

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)，同一会话
  - → “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，也叫用户相关反馈(显式相关反馈)
- 显式相关反馈、隐式相关反馈、伪相关反馈
- 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。

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 (介绍了一个简单的同义词自动构造方法)

课后练习

- 有待补充