

信息检索导论

An Introduction to Information Retrieval

第9讲 相关反馈及查询扩展

Relevance Feedback & Query Expansion

授课人：林政

中国科学院信息工程研究所/国科大网络空间安全学院

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 (查准率)

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

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

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

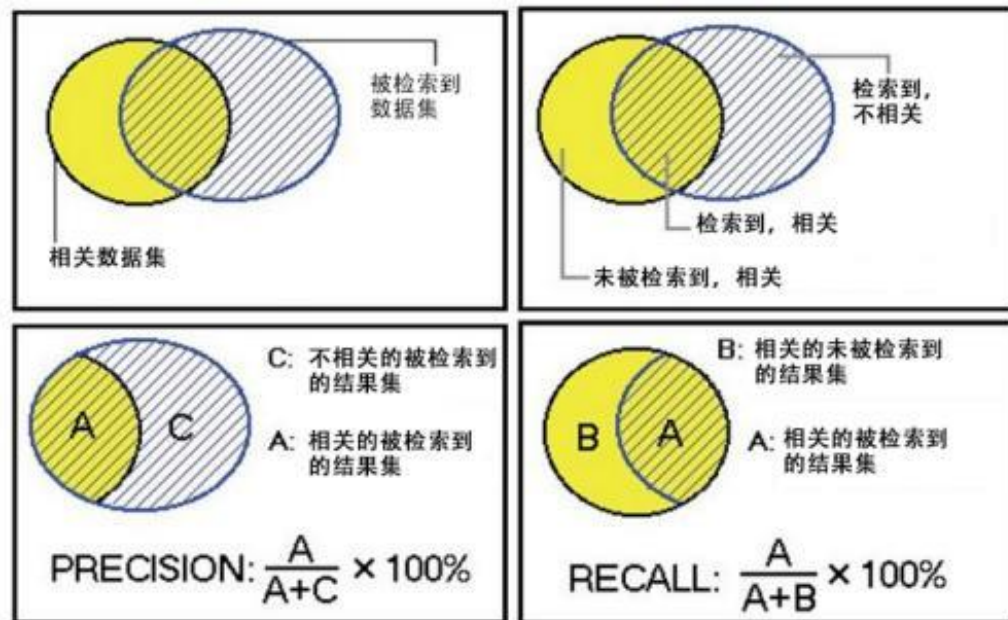
正确率 vs. 召回率

	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)

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

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

大规模语料召回率难以计算，怎么办？
 一般采pooling的方法：对多个检索系统的Top N个结果组成的集合进行标注，标注出的相关文档集合作为整个相关文档集合。



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-Measure是Precision和Recall加权调和平均）

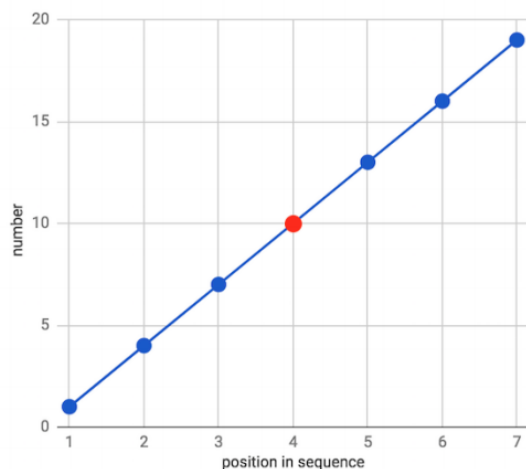
$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1 - \alpha}{\alpha}$$

如果beta=1，就是常见的F1

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

调和平均数，又称倒数平均数，是总体各统计变量倒数的算术平均数的倒数。

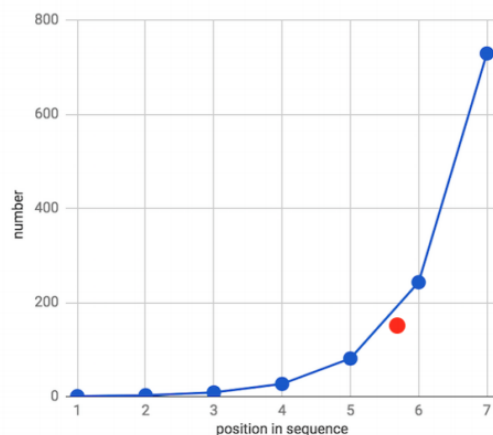
关于平均数的扩展



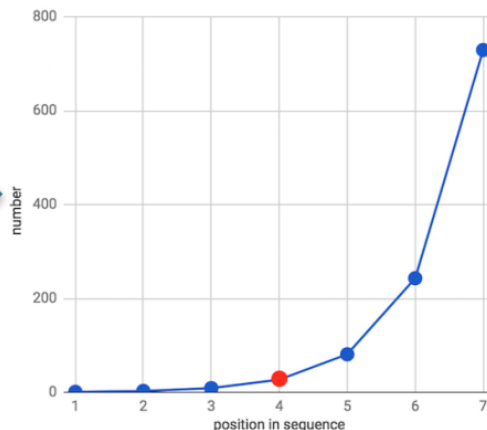
算数平均数



有些数据集内部存在乘法或指数关系



算数平均数



几何平均数

n 个观察值连乘积的 n 次方根

假设我们有一笔5年期存款，本金为\$100,000，每年的利率是变动的。年利率：1%、9%、6%、2%、15%
我们想要找到平均年利率，并据此计算5年后本金和利息的总和？

考虑一次去便利店并返回的行程：去程速度为30km/h，
返程时堵车速度为10km/h，去程和返程走的是同一路线，也
就是说距离一样（5 miles），整个行程的平均速度是多少？

F值为什么采用调和平均数

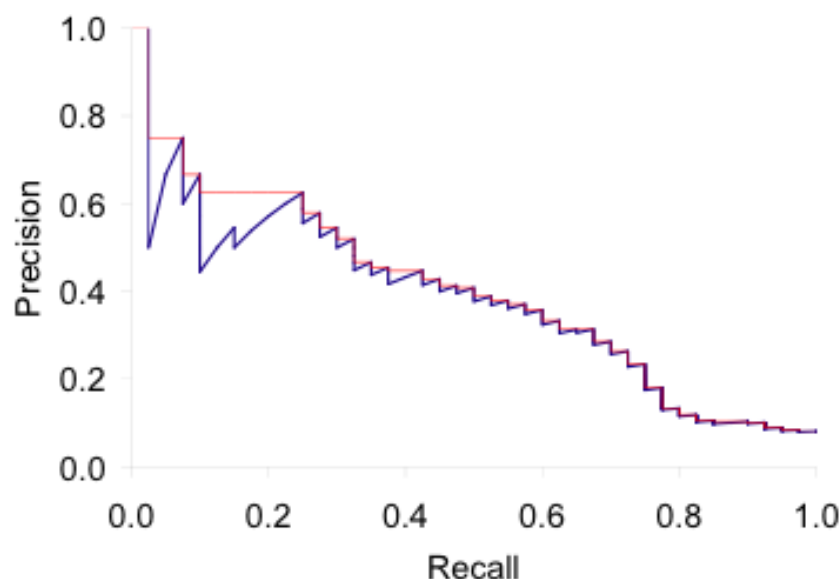
算数、几何、调和这三种平均数各有利弊，但调和平均数受极端值影响较大，更适合评价不平衡数据的分类问题

已知三种模型得到的 P 和 R 值如下，分别计算三种平均数

	P	R	\bar{X}	G	F_1
algorithm 1	0.5	0.4	0.45	0.45	0.44
algorithm 2	0.7	0.1	0.4	0.27	0.18
algorithm 3	0.02	1.0	0.51	0.14	0.04

可以看出算法3的 P 值非常小，我们认为此模型效果不好，但是利用算数平均数和几何平均数来衡量并不能表现出来，只有 F_1 对极端值比较重视，能够感受到这种变化。

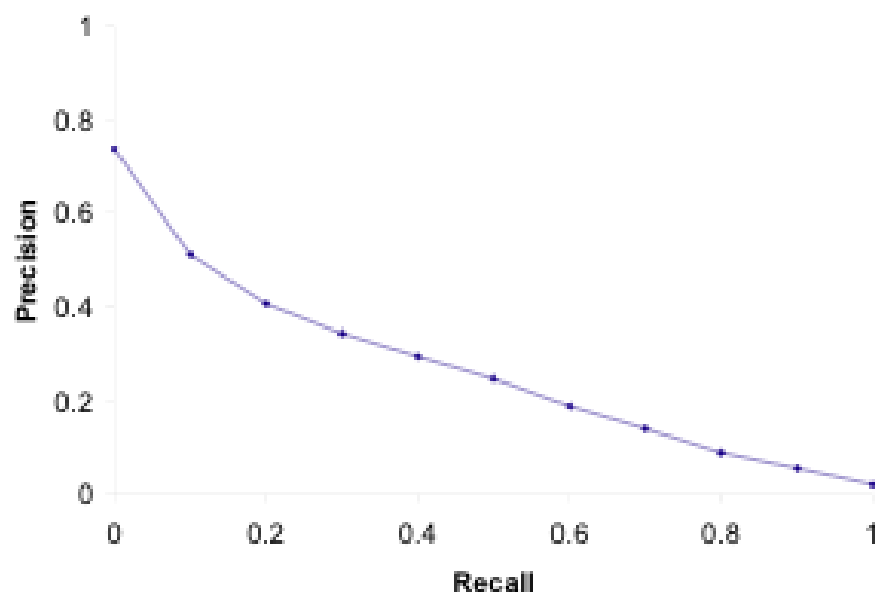
正确率-召回率曲线



- 插值 (红色): 从观察到的值, 得到未观察数据的值, 将来所有点上的最高结果
- 插值后可以得到一个较平滑的曲线, 来去掉细小变化的影响, 更合理的评价搜索引擎。

- 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
- 准确率和召回率是互相影响的, 理想情况下肯定是做到两者都高, 但是一般情况下准确率高、召回率低, 召回率低、准确率高, 当然如果两者都低, 那是什么地方出问题了

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
- 11点平均正确率是一个带插值的平均正确率，没有考虑序

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
 - 未插值的平均准确率: 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考虑了召回率、准确率和序(位置)
- MAP是IR领域使用最广泛的指标之一
- 对于大规模IR系统, 因为召回率通常很难计算, 没有大规模真实标注, 所以不考虑召回率时也经常使precision@N

多个查询求平均：宏平均和微平均

- 假如有两次查询：
- **宏平均**（先分别计算每个查询的P和R，再计算平均值）

$$\text{Macro-average precision} = (P1+P2)/2$$

$$\text{Macro-average recall} = (R1+R2)/2$$

- **微平均**（把所有查询一次性都考虑进来）

$$\text{Micro-average precision} = (TP1+TP2)/(TP1+TP2+FP1+FP2)$$

$$\text{Micro-average recall} = (TP1+TP2)/(TP1+TP2+FN1+FN2)$$

如何每个查询的样本数量相差不大，那么宏平均和微平均差异也不大

如果每个查询的样本相差较大：

更注重样本量多的查询：使用微平均

更注重样本量少的查询：使用宏平均

NDCG（归一化折损累计增益）

每个文档不仅仅是相关和不相关，还有相关程度差别，比如0 1 2 3

- 对某个查询 q , Directed **G**ain

$$G' = \langle 3, 2, 3, 0, 0, 1, 2, 2, 3, 0, \dots \rangle.$$

- Cumulated Gain(**CG**) vector

$$CG[i] = \begin{cases} G[1], & \text{if } i = 1 \\ CG[i - 1] + G[i], & \text{otherwise.} \end{cases}$$

$$CG' = \langle 3, 5, 8, 8, 8, 9, 11, 13, 16, 16, \dots \rangle.$$

- **Discounted CG** vector(${}^b\log i$ 表示以 b 为底对 i 取对数)

- 相关度级别越高的结果越多越好

$$DCG[i] = \begin{cases} CG[i], & \text{if } i < b \\ DCG[i - 1] + G[i] / {}^b\log i, & \text{if } i \geq b. \end{cases}$$

- 相关度级别越高的结果越靠前越好

$$DCG' = \langle 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61, \dots \rangle.$$

NDCG

- BV(Best Vector): 假定 m 个 3, l 个 2, k 个 1, 其他都是 0

$$BV[i] = \begin{cases} 3, & \text{if } i \leq m, \\ 2, & \text{if } m < i \leq m + l, \\ 1, & \text{if } m + l < i \leq m + l + k, \\ 0, & \text{otherwise.} \end{cases}$$

$$I' = \langle 3, 3, 3, 2, 2, 2, 1, 1, 1, 1, 0, 0, 0, \dots \rangle.$$

$$CG'_1 = \langle 3, 6, 9, 11, 13, 15, 16, 17, 18, 19, 19, 19, 19, \dots \rangle$$

$$DCG'_1 = \langle 3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 11.21, 11.53, 11.83, 11.83, 11.83, \dots \rangle.$$

理想的DCG每一维都比真实的DCG值大，真实DCG除以理想DCG就是NDCG

NDCG

- **N**ormalized (**D**)**CG**

$$\text{norm-vect}(\mathbf{V}, \mathbf{I}) = \langle v_1/i_1, v_2/i_2, \dots, v_k/i_k \rangle.$$

$$\begin{aligned} \text{nCG}' &= \text{norm-vect}(\text{CG}', \text{CG}'_I) \\ &= \langle 1, 0.83, 0.89, 0.73, 0.62, 0.6, 0.69, 0.76, 0.89, 0.84, \dots \rangle. \end{aligned}$$

- $\text{N(D)CG}@k$: 表示第 k 个位置上的 N(D)CG 值

标准的评价会议: TREC

- TREC = Text Retrieval Conference (TREC)
- TREC是文本检索领域人气最旺、最权威的评测会议，由美国国防部高等研究计划署 (Defense Advanced Research Projects Agency, 简称 DARPA) 与美国国家标准和技术局 (National Institute of Standards and Technology, 简称NIST) 联合主办。
- 该会议细分为几大主要方向：问题回答 (QA)、特定领域检索 (Legal、Genomics、Enterprise、Blog)、传统Web检索等。
- 会议负责组织收集并向与会者提供标准的语料库 (Corpus)、检索条件和问题集 (Query Set)、以及评测办法 (Evaluation)，与会者则被要求在规定的时间内构造检索系统并提交检索结果

动态摘要

- 结果片段包括文档的标题以及一段自动抽取的摘要
 - 静态摘要：永远保持不变，并不随查询变化而变化
 - 动态摘要：基于查询的摘要，根据查询推导出信息需求来进行个性化生成，并试图解释在给定查询下返回当前文档的原因
-
- 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
 - 出现查询短语和查询词项的snippet优先
 - 最终将所有snippet都显示出来作为摘要

一个动态摘要的例子

查询: “**new guinea economic development**” Snippets (加黑标识)
that were extracted from a document: . . . **In recent years, Papua New Guinea has faced severe economic difficulties and**
economic growth has slowed, partly as a result of weak governance
and civil war, and partly as a result of external factors such as the
Bougainville civil war which led to the closure in 1989 of the
Panguna mine (at that time the most important foreign exchange
earner and contributor to Government finances), the Asian
financial crisis, a decline in the prices of gold and copper, and a fall
in the production of oil. **PNG’s economic development record**
over the past few years is evidence that governance issues
underly many of the country’s problems. Good governance, which
may be defined as the transparent and accountable management of
human, natural, economic and financial resources for the purposes
of equitable and sustainable development, flows from proper public
sector management, efficient fiscal and accounting mechanisms,
and a willingness to make service delivery a priority in practice. . . .

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, 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。

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**【实际也有可能提高正确率】
- 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

- 相关反馈的本质是对检索返回的文档进行相关性判定(人工或者系统判定)。
 - 相关反馈常常用于查询扩展，所以提到相关反馈默认为有查询扩展

- 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)

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

- ① 用户提交一个(简短的)查询
- ② 搜索引擎返回一系列文档
- ③ 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
- ④ 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
- ⑤ 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。一般是公司做, 能拿到用户行为数据。
- **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。一般是研究机构做。

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 图像检索















图像检索里一般很方便让用户参与标注，进行显式相关反馈

初始查询的结果













Initial search results interface showing a grid of images and associated data.

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

					
(144473, 16459)	(144457, 252140)	(144456, 262037)	(144456, 262063)	(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
					
(144403, 264544)	(144403, 265153)	(144510, 257752)	(144530, 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













用户反馈: 选择相关结果

Interface showing a grid of 12 image results, each with a green border indicating selection. The interface includes navigation buttons: Browse, Search, Prev, Next, Random.

					
(144473, 16458)	(144457, 252140)	(144456, 262857)	(144456, 262863)	(144457, 252134)	(144493, 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)	(144539, 525937)	(144456, 240611)	(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

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 0.23853	(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

用户（显式）相关反馈小结

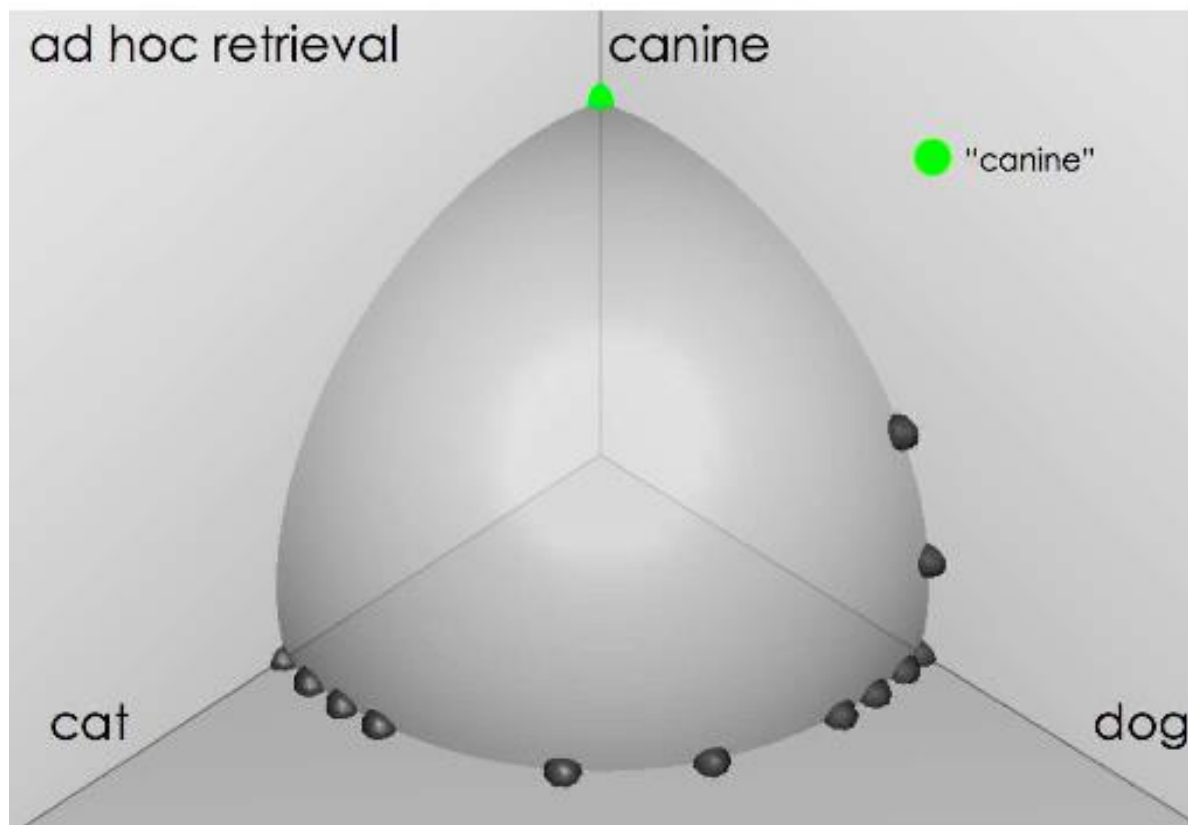
- 优点：

- 以显式的方法让用户进行标识，标识的正确率高

- 缺点：

- 需要用户的直接参与，而大多数用户不愿意进行标注
 - 需要精心的界面设计

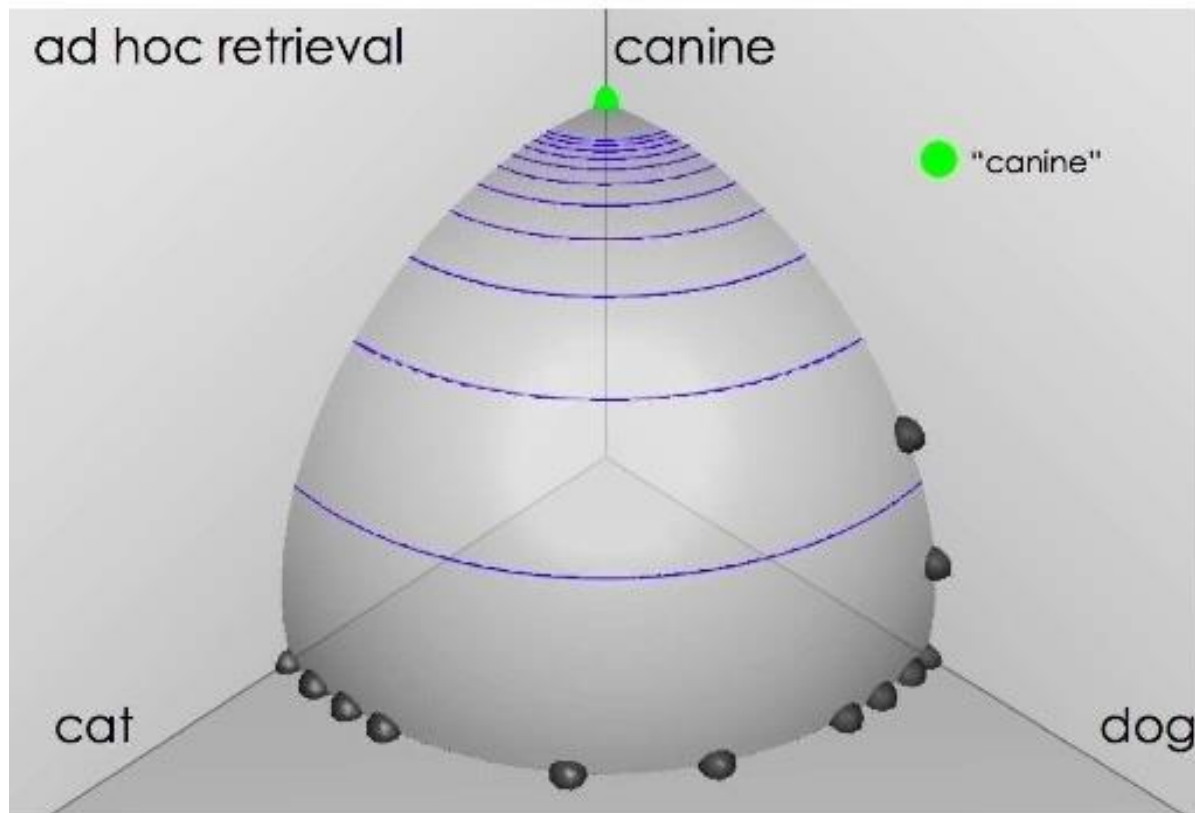
例2 向量空间的例子: 查询 “canine”



Source:
Fernando Díaz

canine: 犬科

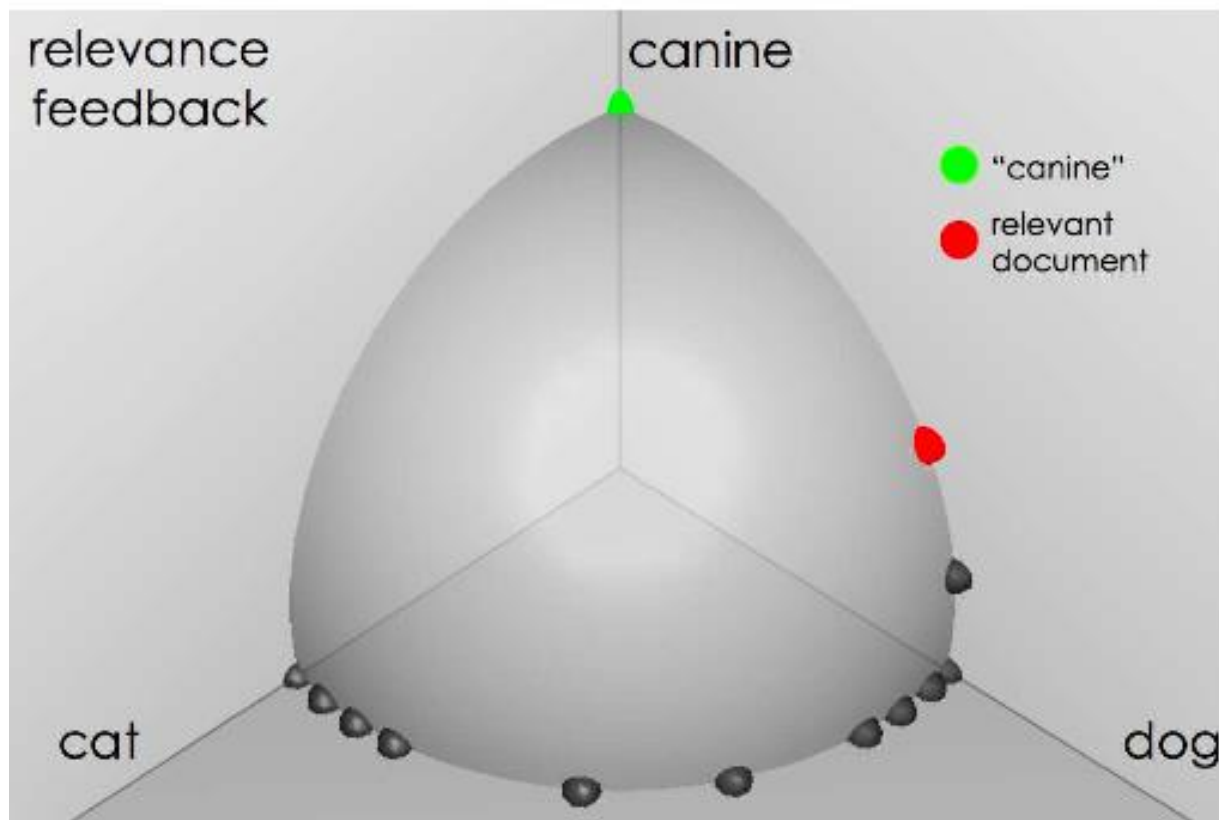
文档和查询“canine”的相似度



Source:

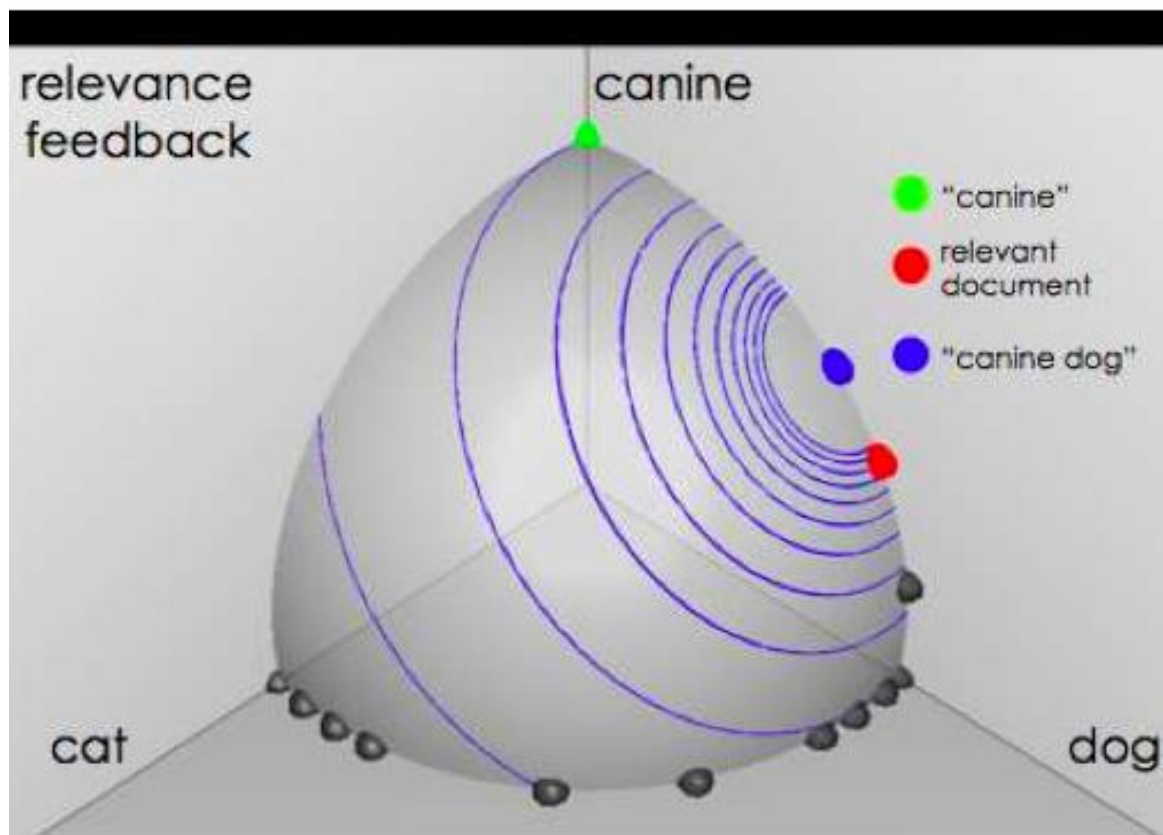
Fernando Díaz

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

原始查询: [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 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

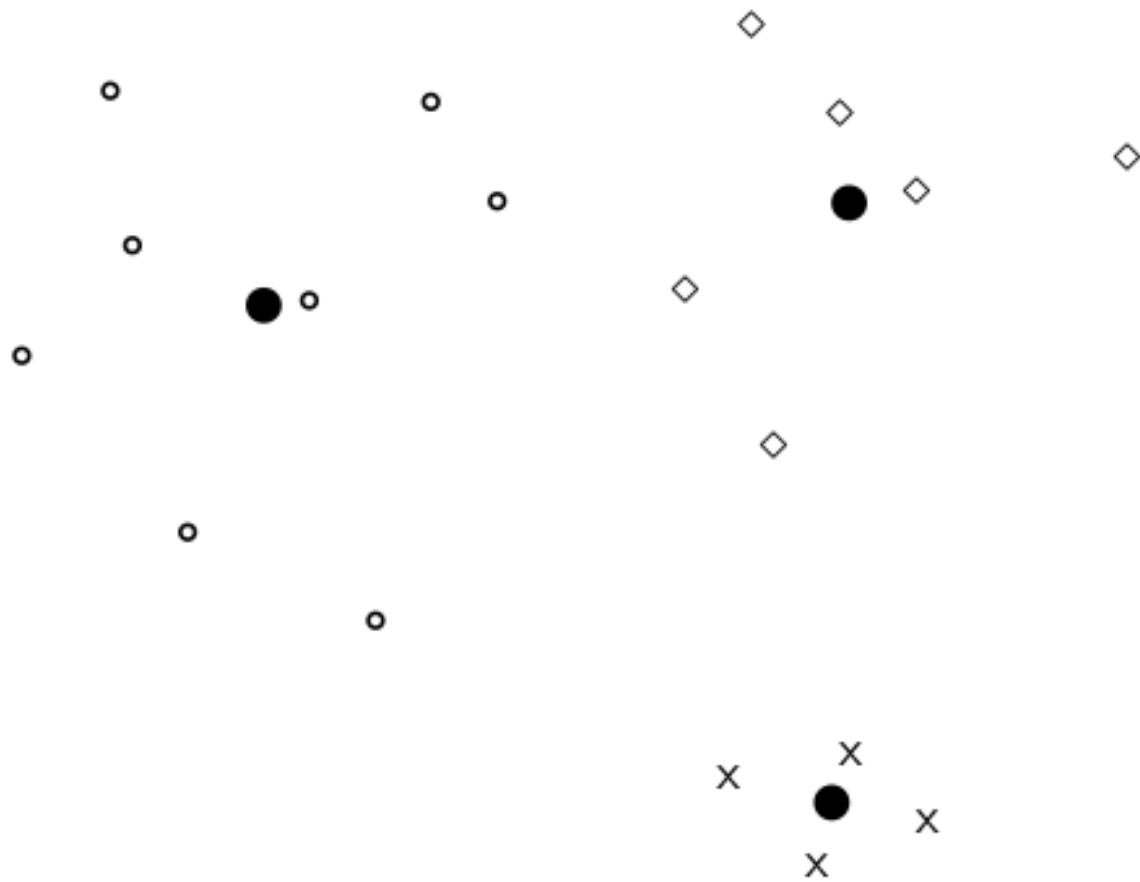
相关反馈中的核心概念：质心

- 质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:

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

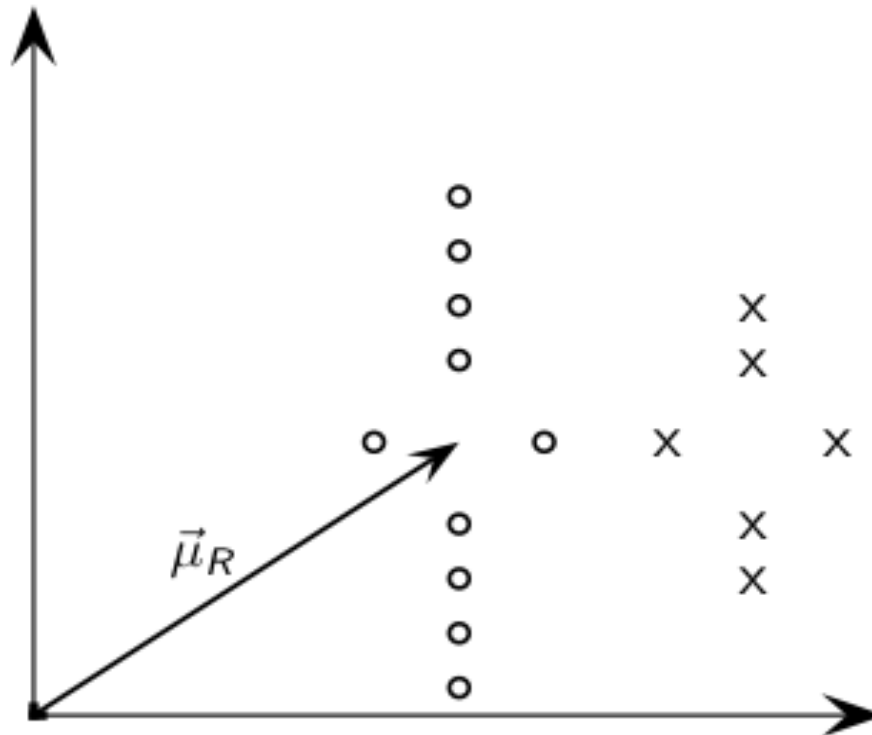
Rocchio算法

- 最优查询向量为：

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

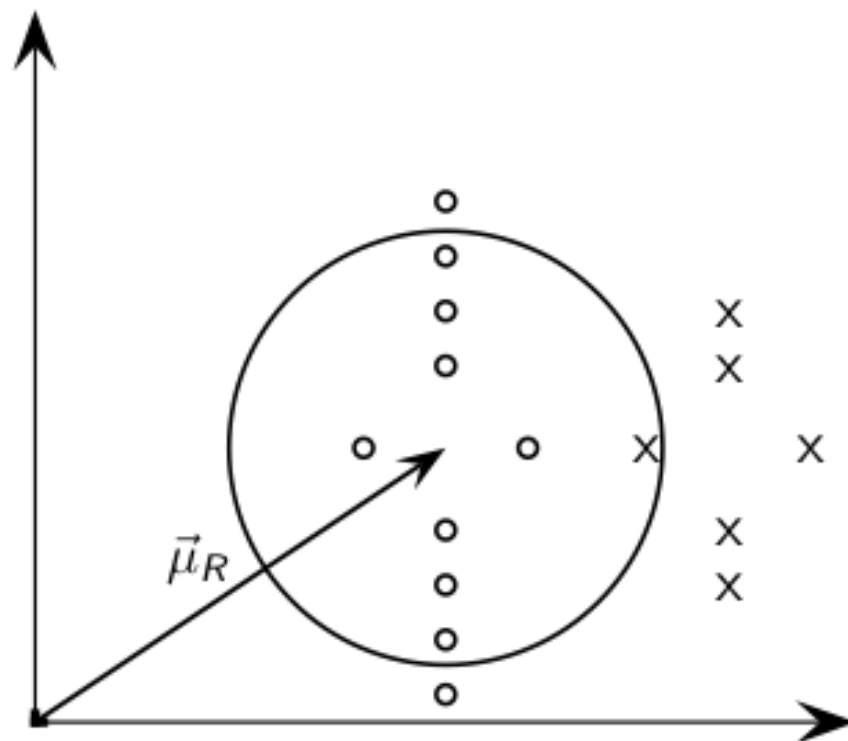
- 即将相关文档的质心移动一个量，该量为相关文档质心和不相关文档的差异量

Rocchio算法图示



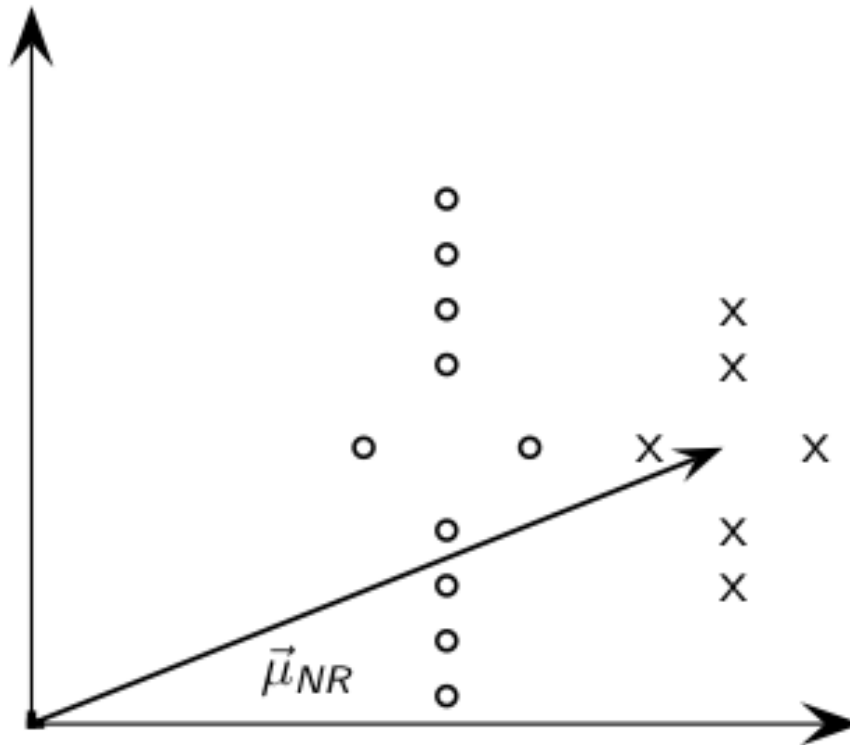
$\vec{\mu}_R$: 相关文档的质心

Rocchio算法图示



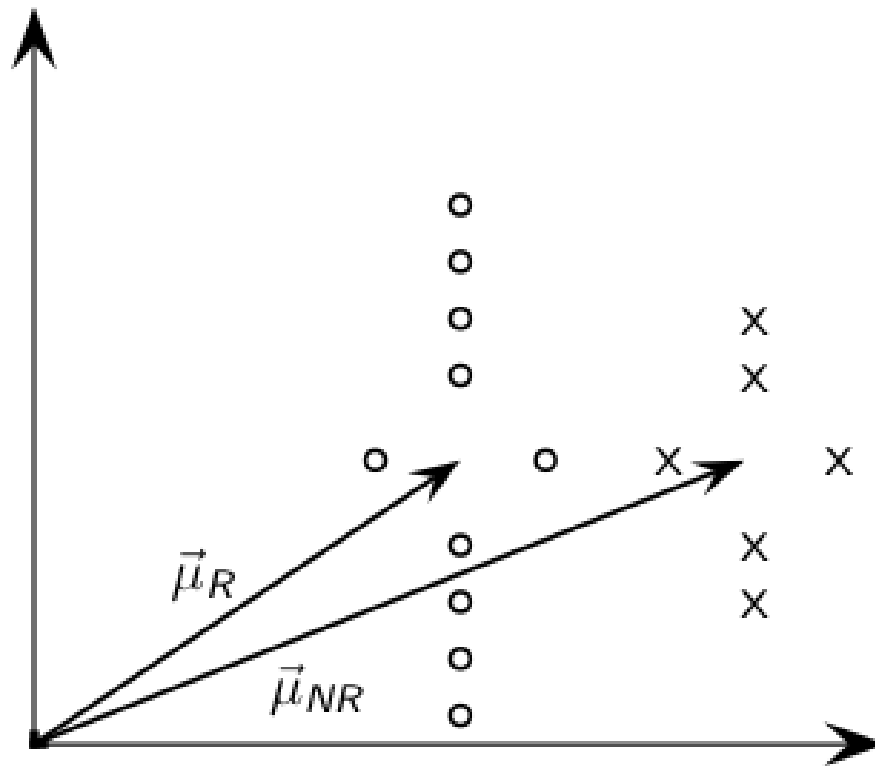
$\vec{\mu}_R$ 不能将相关/不相关文档分开

Rocchio算法图示

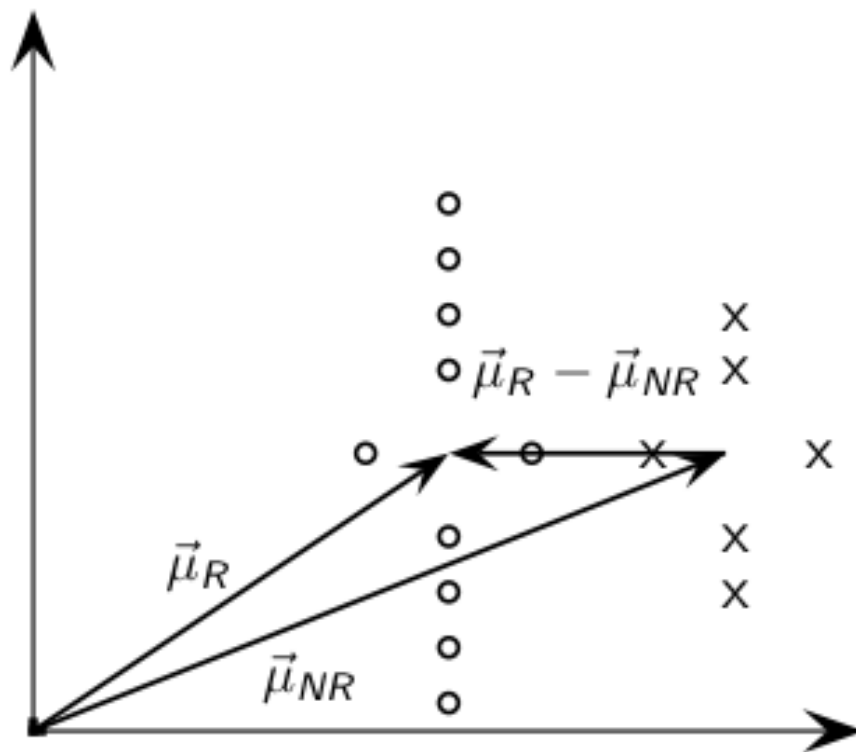


$\vec{\mu}_{NR}$: 不相关文档的质心

Rocchio算法图示

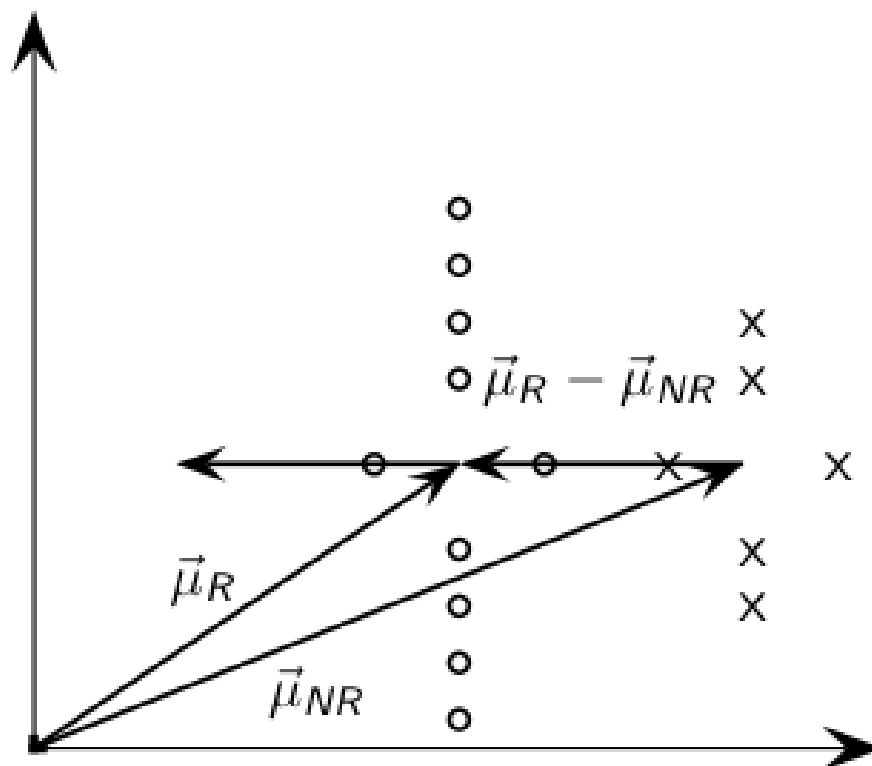


Rocchio算法图示



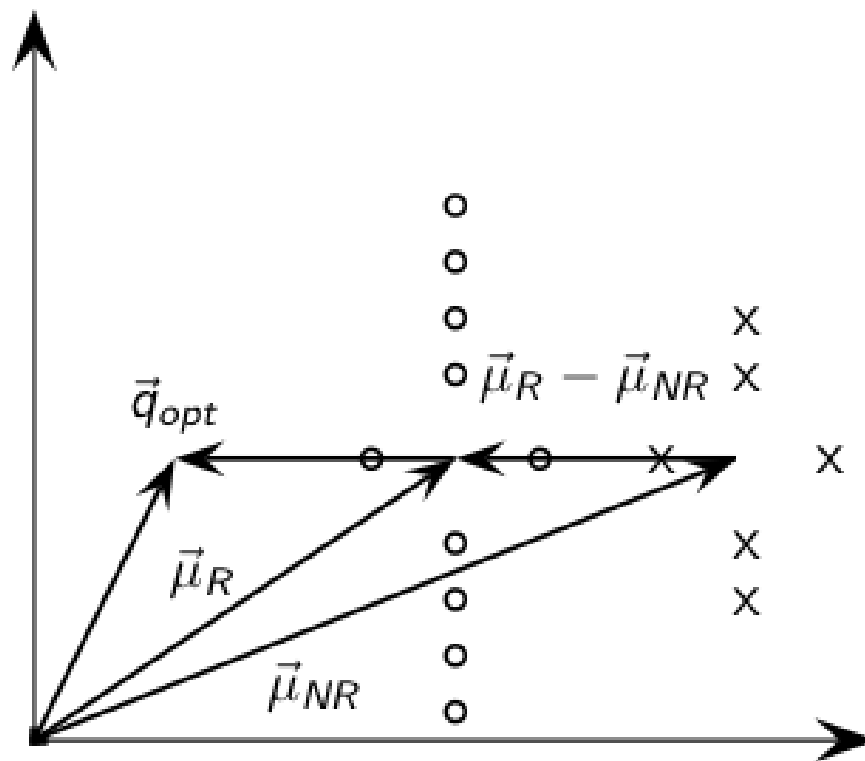
$\vec{\mu}_R - \vec{\mu}_{NR}$: 差异向量

Rocchio算法图示



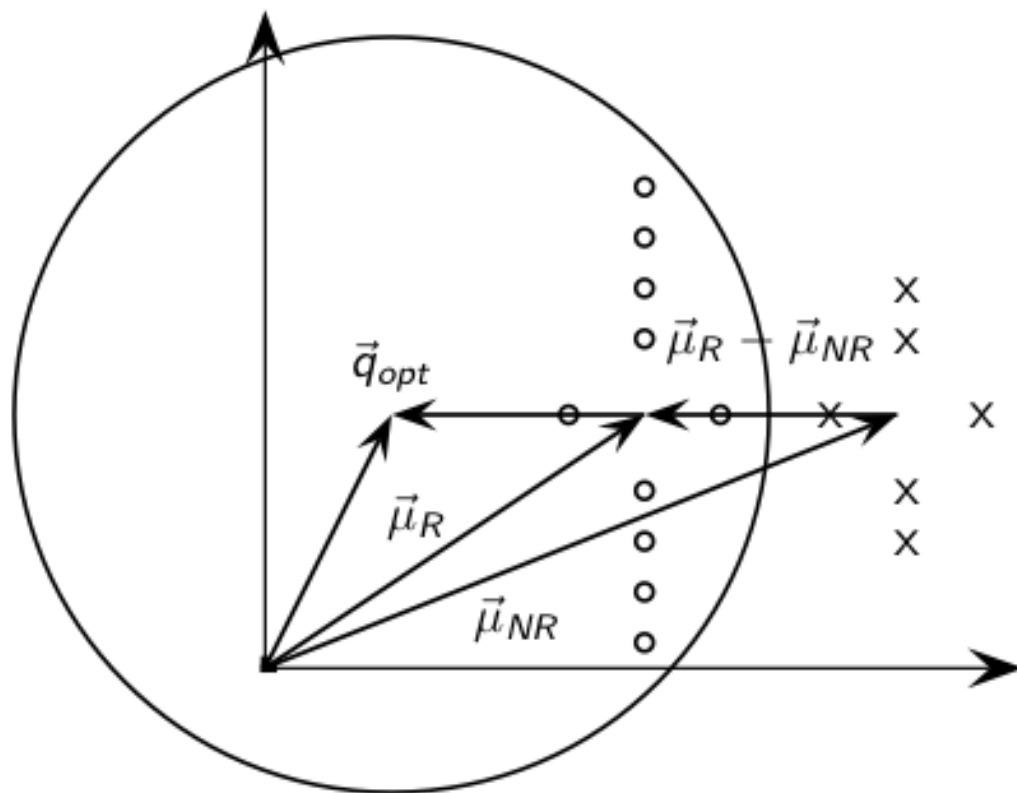
$\vec{\mu}_R$ 加上差异向量

Rocchio算法图示



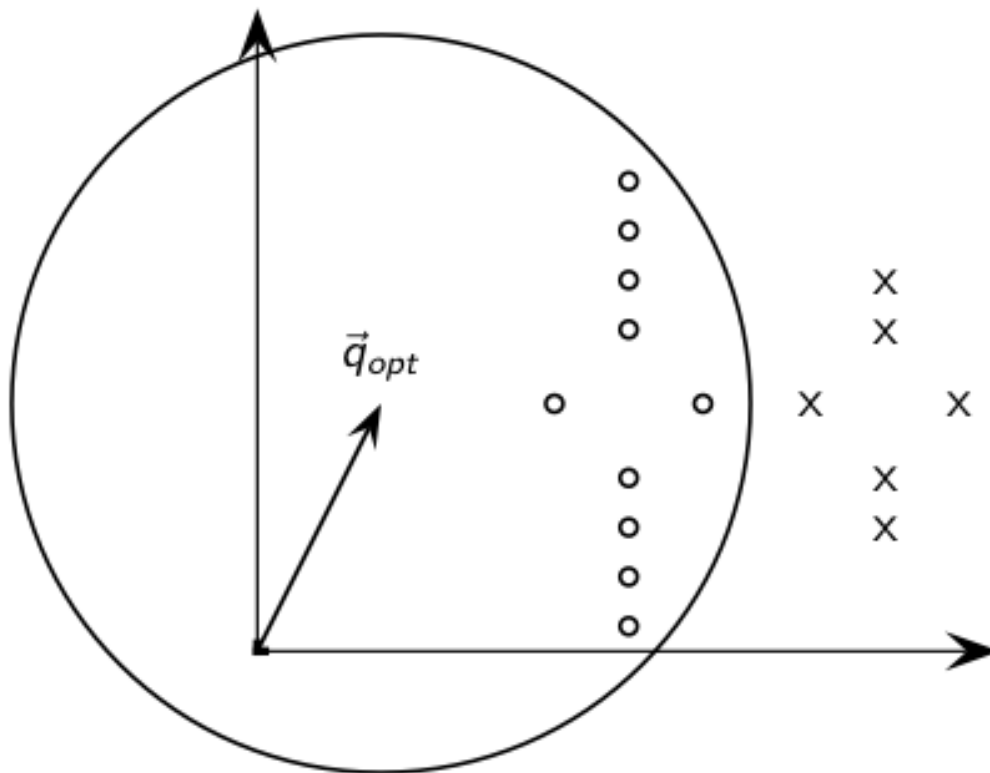
得到 \vec{q}_{opt}

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

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 $\gamma = 0$

练习题：

1. 相关反馈不能解决的问题（）（多选）
- A. 拼写错误
 - B. 跨语言IR
 - C. 用户的词汇表和文档集的词汇表不匹配
 - D. 提高召回率

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: 对中国发展中国家的帮助 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中国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比较
- 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?

用户相关反馈界面例子

The screenshot shows a web browser window displaying a list of search results. On the right side, there is a sidebar with two main sections: 'Your query was rewritten to:' and 'Refine your search using these keywords found in the results:'. The 'Your query was rewritten to:' section shows the original query 'information filtering' and a rewritten version 'information filtering' with a link to 'Repeat without rewrite'. The 'Refine your search using these keywords found in the results:' section lists various keywords such as 'bayesian', 'distribution of errors', 'filtered', 'filtering algorithm', 'information needs', 'information retrieval system', 'information science', 'information sources', 'irrelevant information', 'knowledge-based', 'latent', 'query expansion', 'selective dissemination of information', 'social information', 'web information', and 'word-of-mouth'. Below this list is a section 'Or refine using:' with a dropdown menu set to 'All of the words' and a 'Refine' button.

Search Results List:

1. [Information filtering and information brokering: revisiting the role of the Librarian](#) [1K]
Jun 2001
Information filtering and information brokering: revisiting the role of the Librarian Derek Law, MA,D.Univ.,FRSE Director of Information Strategy University of Strathclyde in Glasgow Next slide Back to first slide View graphic version
[more hits from](#) [http://www.bireme.br/crics5/proceedings/Law/01Havana/t...] [similar results](#)
2. [Information Filtering Bibliography - Steve Gant](#) [59K]
Sep 1997
Information Filtering Bibliography by Steve Gant revised...Bellcore Workshop on High-Performance **Information Filtering** (Morriston, N.J.). Baclace, Paul E (1992): Competitive Agents for **Information Filtering** . Commun. ACM 35(12, December...
[more hits from](#) [http://ils.unc.edu/gants/filterbib.html] [similar results](#)
3. [A Case-Based Information Filtering System for the World Wide Web](#) [27K]
Mauro Marinilli / Alessandro Micarelli / Filippo Sciarrone, Jun 1999
...Case-Based Approach to Adaptive **Information Filtering** for the WWW Mauro Marinilli...for the realization of an **Information Filtering** system for the Web. The...case-based approach to adaptive **Information Filtering** Keywords: **Information**...
[more hits from](#) [http://www.wis.win.tue.nl/asum99/marinilli/marinilli.htm...] [similar results](#)
4. [Information Filtering and Retrieval](#) [2K]
Jun 2000
Slide 20 of 21 **Information Filtering** and Retrieval Combine Retrieval of geographic, visual, audio and text information. Reduce the amount of data returned through filtering for redundancy, maximal information, compression, summarization. --
[more hits from](#) [http://www.informedia.cs.cmu.edu/eod/EODforWeb/EOD_Int...] [similar results](#)
5. [The Information Filtering Module](#) [6K]
Feb 1995
...Extracting Document Representations Up: Newt: An Implementation Previous: Genetic Algorithm The **Information Filtering** Module The **Information Filtering** module (called YAIF, for Yet Another Information Filter) is responsible for actually retrieving...
[more hits from](#) [http://agents.www.media.mit.edu/groups/agents/publicat...] [similar results](#)
6. [CFP Wks. Pract. Appl. Information Filtering](#) [8K]
May 1996
...Thread] CFP Wks. Pract. Appl. **Information Filtering** Subject : CFP Wks. Pract. Appl. **Information Filtering** From sorenson@csvgx1.ucc.ie Date...Workshop on Practical Applications of **Information Filtering** to be held in conjunction with First...
[more hits from](#) [http://documents.cfar.umd.edu/newsgroups/miscarchive/m...] [similar results](#)
7. [SIFTER Information Filtering Digital Libraries IU Indiana University Bloomington Indianapolis](#) [1K]
Oct 2004
Comprehensive information on the SIFTER project Go to About SIFTER Publications Projects Sponsors Related Links URL: httpsifter.indiana.edu/index.shtml Updated: June 11, 2002 Contact: webmaster@sifter.indiana.edu
[more hits from](#) [http://sifter.indiana.edu/] [similar results](#)
8. [PhD Studentship in Multimedia Information Filtering](#) [5K]
Oct 2004

Microsoft PowerPoint - [现代信息检索第4章]

开始 2 Messenger 2 Total Comm... Foxmail Microsoft Pow... 2 Internet E... 13:33

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（看过的内容再看一遍？）
- 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设备可以跟踪用户的眼球动作
 - 拉近、拉远、瞟、凝视、往某个方向转

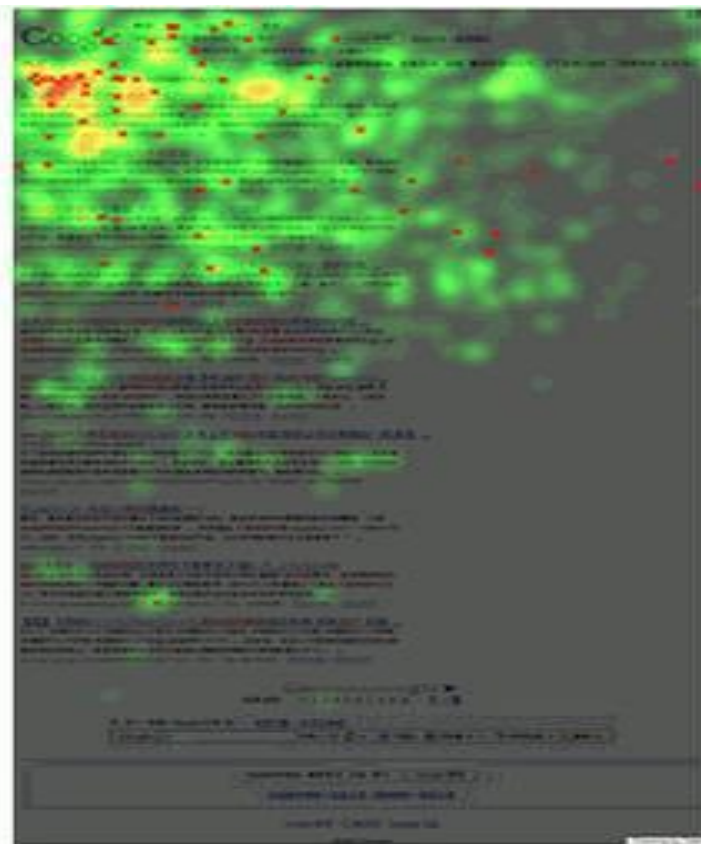
点击行为(Click through behavior)

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)

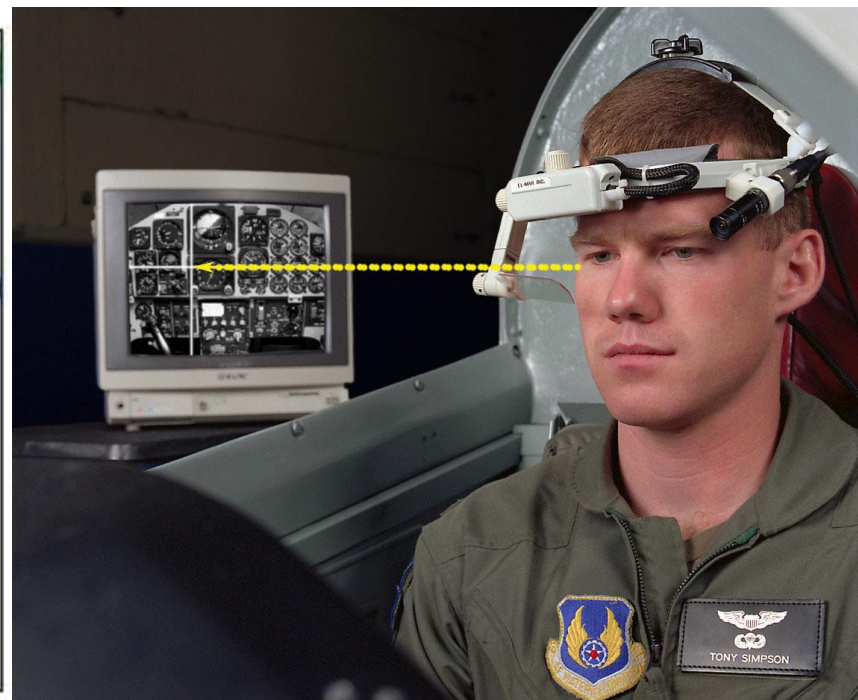
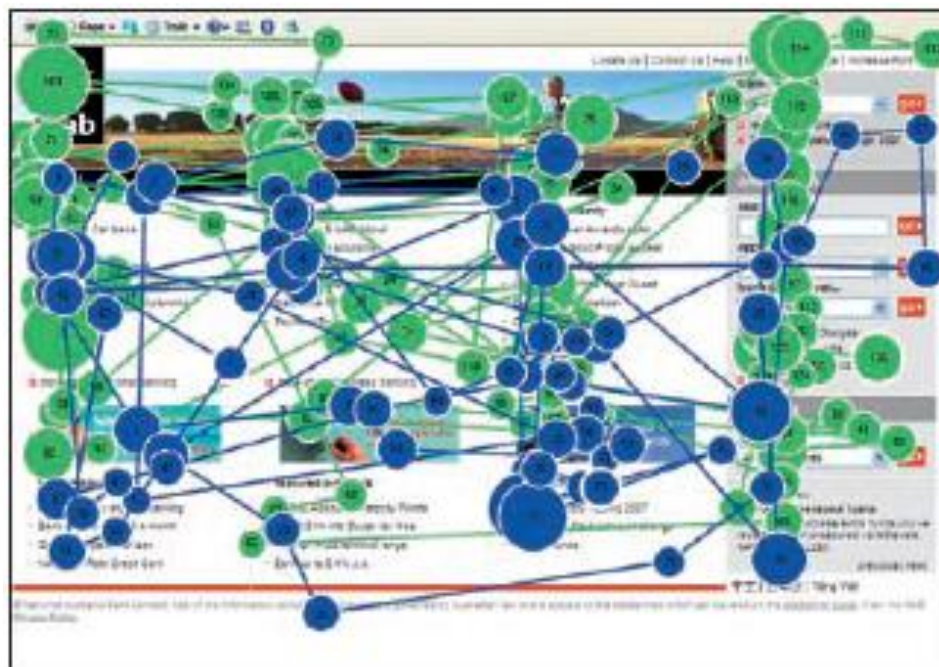


Baidu



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)

- 也称为盲相关反馈(Blind 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 (R vs. RF)
- 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

练习题：

2. 下面关于伪相关反馈的说法，错误的有（）

- A. 不需要额外的训练数据
- B. 平均效果不错
- C. 多次循环迭代后有可能导致主题漂移，如果是有歧义的查询，漂移情况会更严重
- D. 伪相关文档的质量和数量不太影响检索性能

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”，大部分情况下不做区分。
- **查询扩展**：最初含义是对原始查询增加**term**
- **查询重构**：对初始查询进行修改，以便能更接近用户的查询需求
 - 增删**term**
如：“计算机” → “计算机 电脑”
 - 修改**term**的权重或者概率分布参数
如：“科学院1 研究生1” → “科学院2 研究生1”

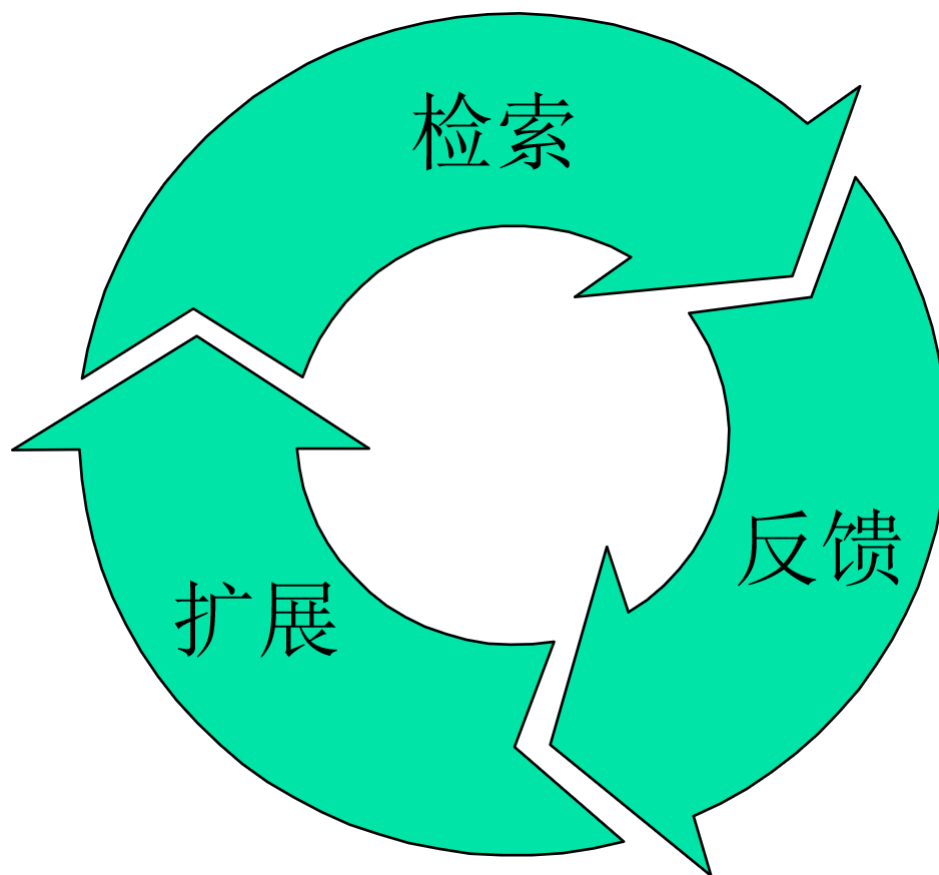
为什么要进行查询扩展

- 用户有明确的查询目的，但是用户不知道如何构造好的查询，因此，初始查询往往不能很好地表达其查询需求
 - 用户没有经验
 - 查询的表达方式很多
 - 对检索系统不熟悉
 - 对目标文档分布不熟悉
- 用户不知道要具体查询什么，需要查看结果以后才逐渐细化

相关反馈与查询扩展

- 相关反馈是给返回的结果文档进行相关性标注，查询扩展是对原始的查询进行修改。
- 相关反馈的结果常常用于查询扩展，查询扩展也可以不依赖于相关反馈
- 注：在很多教材中，相关反馈通常被认为是查询扩展的技术之一。基于相关反馈的查询扩展和相关反馈是一个意思。这里为了理解的方便，做了一点点区分。

检索—相关反馈—查询扩展循环



查询扩展方法分类

- **基于(用户)相关反馈的查询扩展**：用户参与反馈过程(显式/隐式)，系统利用用户反馈结果对原始查询进行调整。
- **基于自动局部分析的查询扩展**：利用上次检索结果的前N个结果文档进行自动分析
- **基于自动全局分析的查询扩展**：利用全部的文档集合进行分析。
- **基于(外部)资源的方法**：基于已有的词典(wordnet、Hownet、同义词词林)、网络资源(wiki、social tag)、基于其他搜索引擎(如google)的结果、搜索日志(search log)等等

基于（用户）相关反馈的查询扩展

- 向量空间模型中基于(用户)相关反馈的查询扩展
（前面介绍的Rocchio算法）
- 概率模型中基于(用户)相关反馈的查询扩展
- 语言模型中基于(用户)相关反馈的查询扩展
- 说明：
 - 实际上各种形式的相关反馈方法都可以用于这里的查询扩展

向量空间模型中的查询扩展

标准Rocchio公式(中心向量法)

$$\vec{q}_m = \alpha \vec{q} + \beta \frac{\sum_{i \in R} \vec{d}_i}{n_R} - \gamma \frac{\sum_{i \notin R} \vec{d}_i}{n_{\bar{R}}}$$

修改后的查询向量是原有查询向量、相关文档的平均文档向量及不相关文档的平均文档向量的加权求和。

$\alpha\beta\gamma$ 为加权重。几种常用取值方法： $\alpha=\beta=\gamma=1$, 或
 $\alpha=\beta=1, \gamma=0$

基于用户相关反馈的查询扩展的优缺点

- 优点：
 - 为用户提供了方便反馈的接口，但是不参与查询扩展过程
 - 将整个搜索过程分成几个小步骤，每个步骤都很容易把握
 - 提供了一个受控过程来强调或者弱化某些term
- 缺点：
 - 主要的缺点是需要人工的参与，加重用户的负担。
- IR中集中研究不需要用户相关反馈而自动进行QE的方法。

局部分析 VS 全局分析

- 局部分析(local analysis): 利用对初始查询检索得到的结果, 特别是排名靠前的那些文档(称为局部文档)进行分析后来改善查询, 需要基于上次检索的结果(基于伪相关反馈)
- 全局分析(global analysis): 利用文档集合中的全部文档进行分析, 不需要基于初始检索的结果。

基于自动局部分析的查询扩展

- 基于局部聚类(local clustering)的方法
- 基于局部上下文分析(local context analysis, LCA)的方法

基于自动局部分析的查询扩展

- 基于局部聚类(local clustering)的方法 ←
- 基于局部上下文分析(local context analysis, LCA)的方法

基本思路

- 利用局部文档对term进行聚类，即将相关的term聚在一起，聚类的结果称为一个个簇(cluster)，于是利用簇中的相关term对查询q进行扩展。
- 关键：定义term之间的相似度，不同的相似度定义得到不同的簇。
- 一种思路：关联簇Association clusters (还有其他计算term之间相似度方法)

关联簇(1)

- 局部文档的向量表示矩阵（词-文档矩阵），其中每个 $a_{ij} = f(t_i, d_j)$ ，即原始 TF

$$A_{m \times n} = \begin{matrix} & d_1 & d_2 & \dots & d_n \\ \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{matrix} & \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{matrix}$$

- 矩阵 AA^T 称为关联矩阵 (association matrix)，其中的第 u 行、第 v 列元素 $c_{u,v}$ 表示的是 t_u 和 t_v 的相似度 (即 A 中第 u 行、第 v 行对应的向量的相似度)

$$c_{u,v} = \sum_j f(t_u, d_j) * f(t_v, d_j) \longrightarrow \text{内积!}$$

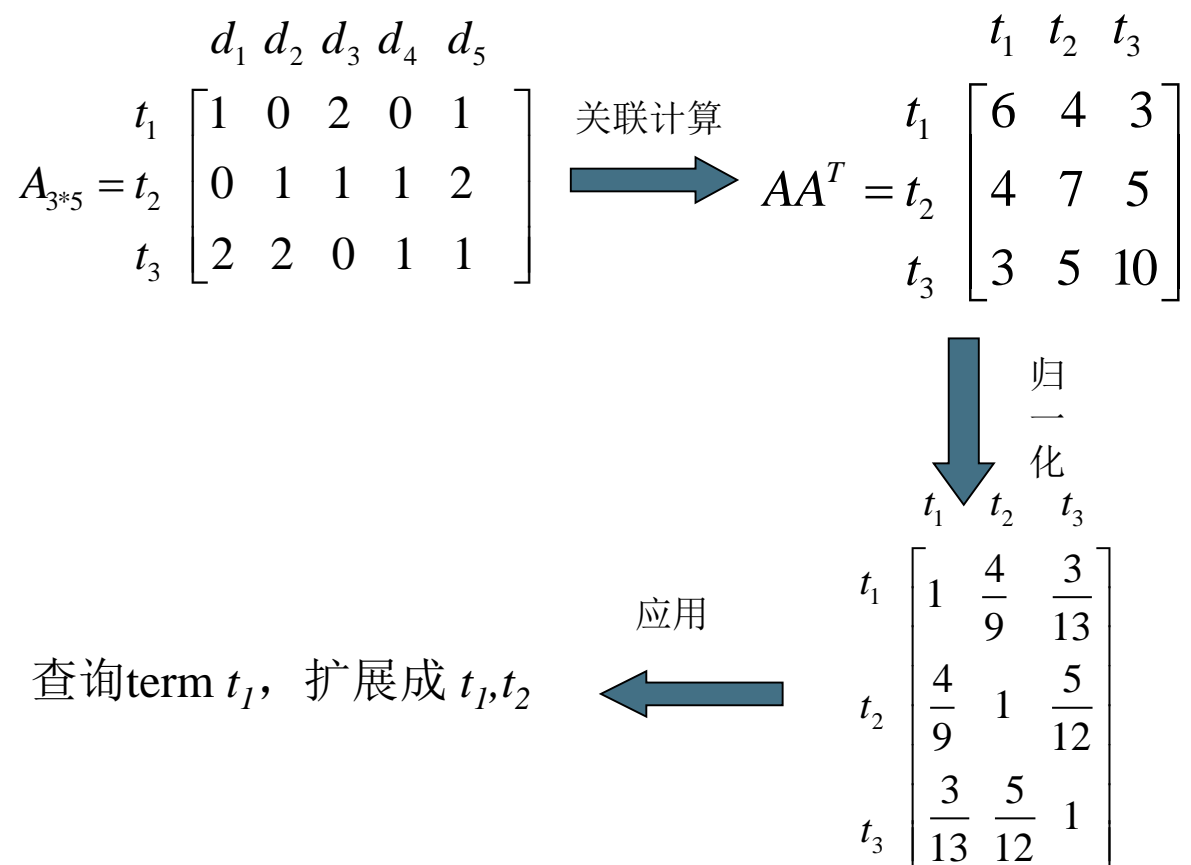
关联簇(2)

- $c_{u,v}$ 实际上表示的是 t_u 、 t_v 在局部文档中共现 (Co-occur) 的频度
- 可以将 $c_{u,v}$ 进行归一化, 得到

$$c'_{u,v} = \frac{c_{u,v}}{c_{u,u} + c_{v,v} - c_{u,v}}$$

- 因此, 对于查询 q 的某个 term q_i , 可以选择在归一化关联矩阵其所在行中相似度较高的多个 term 进行扩展。

一个关联簇的例子



基于簇的查询扩展

- 簇中的不同term互称为邻居 (neighborhood)，或者称为搜索同义项，有别于语法意义上的同义词。
- 查询扩展方式：对于 q 中的每个term，都选择和该term最近的多个term进行扩充。
- 实验表明：度量簇（一种考虑了位置信息的term聚类）效果好于关联簇（没有考虑位置信息），说明位置信息是有用的。

思考题：

除了关联簇的方法，还有哪些方法可以获得词汇之间的语义相似度？

基于自动局部分析的查询扩展

- 基于局部聚类(local clustering)的方法
- 基于局部上下文分析(local context analysis, LCA)的方法←

基本思想

- 局部聚类的缺点：计算的是 q 中每个term和所有term之间的相似度，而不是计算 q 和所有term的相似度。
- 应该将 q 看成一个整体！
- LCA的思想：在局部文档中计算出和查询 q 最相近的term进行扩展。
- LCA是UMass的Jinxi Xu于1996年提出的。本质上说LCA是融合了局部分分析和全局分析的方法。

LCA的三个步骤

- 第一步，将所有文档都进行分段(比如300字节一段)，并将每个段落看成检索对象 (local context)，用原始查询 q 检索，返回和 q 最相似的 n 个段落 (passage)
- 第二步，计算这 n 个段落中的每个概念 c (通常就是term)和 q 的相似度 $sim(q, c)$
- 第三步，选择 sim 值最高的 m 个概念加入到原始查询中。其中加入的概念的权重为 $1-0.9 * i/m$ ， i 为其在 m 个概念中的排序序号。原始查询的term的权重设置为一个较大的值，比如2。

q 和 c 的相似度计算

- 首先定义 c 和某个term k_i (q 中的词) 的相似度, 其中 $pf_{i,j}$ 、 $pf_{c,j}$ 分别表示在第 j 个段落中 k_i 及 c 的出现次数

$$f(c, k_i) = \sum_{j=1}^n pf_{i,j} * pf_{c,j}$$

- 然后定义 c 和 q 的相似度

$$sim(q, c) = \prod_{k_i \in q} (\delta + \frac{\log(f(c, k_i) \times idf_c)}{\log n})^{idf_i}$$

- 其中, idf_i 、 idf_c 分别表示基于段落计算的 k_i 和 c 的 idf

$$idf_i = \max(1, \frac{\log_{10} N / np_i}{5}) \quad idf_c = \max(1, \frac{\log_{10} N / np_c}{5})$$

LCA的使用

- δ 是用于平滑的常数，常常取近0.1的值。
- $sim(q,c)$ 可以看成是利用TF-IDF进行相似度计算的一个变种。
- LCA在不同的文档集合上表现的效果不一致，因此，常常要在使用前做相应调整(调整公式)。

全局分析的思路

- 全局分析是利用全部文档信息计算term之间的相似度。这种相似度计算不依赖于 q ，因此可以事先算好(比如生成一部全局term相似度矩阵)。在此基础上，计算term和查询 q 之间的相似度。

传统全局分析 vs. 现代全局分析

- 传统全局分析利用所有文档集合来计算 q 中每个term和所有term之间的相似度。直到1990年初，全局分析都被认为是非常失败的一种方法。
- 现代全局分析利用所有文档集合来计算查询 q 和所有term之间的相似度。1990年以后，现代全局分析代替了传统全局分析，并表现出良好的效果。

基于相似词典的查询扩展

- 定义 itf （注意区别 idf ）
- 设 N 为所有文档数目， t 为整个文档集中的 term 数目， $f_{i,j}$ 为 term k_i 在文档 d_j 中的频度， t_j 为文档 d_j 中的不同 term 数目，则可以定义所谓 **文档 d_j 的逆 term 频率 itf_j** （可以和逆文档频率 idf 进行类比）

$$itf_j = \log \frac{t}{t_j}$$

idf 是为了计算文档的向量表示，向量空间是词集合， idf 衡量了每个词的语义区分度
 itf 是为了计算词的向量表示，向量空间是文档集合， itf 衡量了每篇文档的语义区分度

Term之间的相似度计算(1)

- 对于所有N篇文档，考虑其矩阵表示

$$A_{m \times N} = \begin{matrix} & d_1 & d_2 & \dots & d_N \\ \begin{matrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{matrix} & \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1N} \\ w_{21} & w_{22} & \dots & w_{2N} \\ \vdots & \vdots & \dots & \vdots \\ w_{m1} & w_{m2} & \dots & w_{mN} \end{bmatrix} \end{matrix}$$

- 每个 $w_{i,j}$ 表示的是 $[k_i, d_j]$ 对应的权重，可以如下计算

$$w_{i,j} = \frac{[0.5 + 0.5 \frac{f_{i,j}}{\max_j(f_{i,j})}] itf_j}{\sqrt{\sum_{l=1}^N [0.5 + 0.5 \frac{f_{i,l}}{\max_l(f_{i,l})}]^2 itf_l^2}}$$

Term之间的相似度计算(2)

- 将上述矩阵的第 i 行看成term k_i 的一个向量表示 \vec{k}_i
- 计算term k_u 和 k_v 之间的相似度，可以采用内积计算方法，至此，可以得到term相似度矩阵。其中的 u 行 v 列为：

$$c_{u,v} = \vec{k}_u \bullet \vec{k}_v = \sum_{\forall d_j} w_{u,j} \times w_{v,j}$$

查询 q 和term之间的相似度计算

- 将 q 向量化，将 q 看成一篇文档，对于 q 中的每个term k_i ，可以利用前面计算 $[k_i, d_j]$ 权重的公式计算权重 $w_{i,q}$ ，从而得到 q 的向量表示：

$$\vec{q} = \sum_{k_i \in q} w_{i,q} \vec{k}_i$$

- 于是可以计算 q 和任意term k_v 之间的相似度

$$sim(q, k_v) = \vec{q} \bullet \vec{k}_v = \sum_{k_u \in q} w_{u,q} \times k_u \bullet k_v = \sum_{ku \in q} w_{u,q} \times c_{u,v}$$

利用sim进行查询扩展

- 选择sim值最高的r个term加入到原始查询中得到新查询 q' ，新加入的term k_v 的权重设置为：

$$w_{v,q'} = \frac{sim(q, k_v)}{\sum_{k_u \in q} w_{u,q}}$$

- 实验表明，基于这种相似词典的方法能够提高检索的效果。

基于外部资源的查询扩展

主要使用的信息：

- 同义词或近义词词典 ([thesaurus](#))
 - 人工构建
 - 自动构建
- 检索日志

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
 - 中文同义词词典: 同义词词林
 - 英文同义词词典: WordNet

基于人工词典的扩展样例: PubMed

The screenshot displays the PubMed search interface. At the top, the NCBI logo is on the left, the PubMed logo is in the center, and the National Library of Medicine (NLM) logo is on the right. Below the logos, a navigation bar includes links for PubMed, Nucleotide, Protein, Genome, Structure, PopSet, and Taxonomy. The search bar contains the text 'Search PubMed for cancer' with 'Go' and 'Clear' buttons. 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 for About Entrez, Text Version, Entrez PubMed Overview, Help | FAQ, Tutorial, New/Noteworthy, E-Utilities, PubMed Services, Journals Database, MeSH Browser, Single Citation, and Matchbox. The main content area shows the 'PubMed Query:' section with the query: `("neoplasms"[MeSH Terms] OR cancer[Text Word])`. At the bottom, there are 'Search' and 'URL' buttons.

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)，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

查询扩展的例子

YAHOO! SEARCH

Web | [Images](#) | [Video](#) | [Audio](#) | [Directory](#) | [Local](#) | [News](#) | [Shopping](#) | [More »](#)

[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.


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.com](#)

[The Palms Casino Resort, Las Vegas](#)
Low price guarantee at the **Palms** Casino resort in Las Vegas. Book...
[lasvegas.hotelscorp.com](#)

Y! 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](#)

常用查询扩展的方法

- 基于相关反馈的查询扩展
- 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)

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
- 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)