信息检索与数据挖掘

第6章 检索的评价

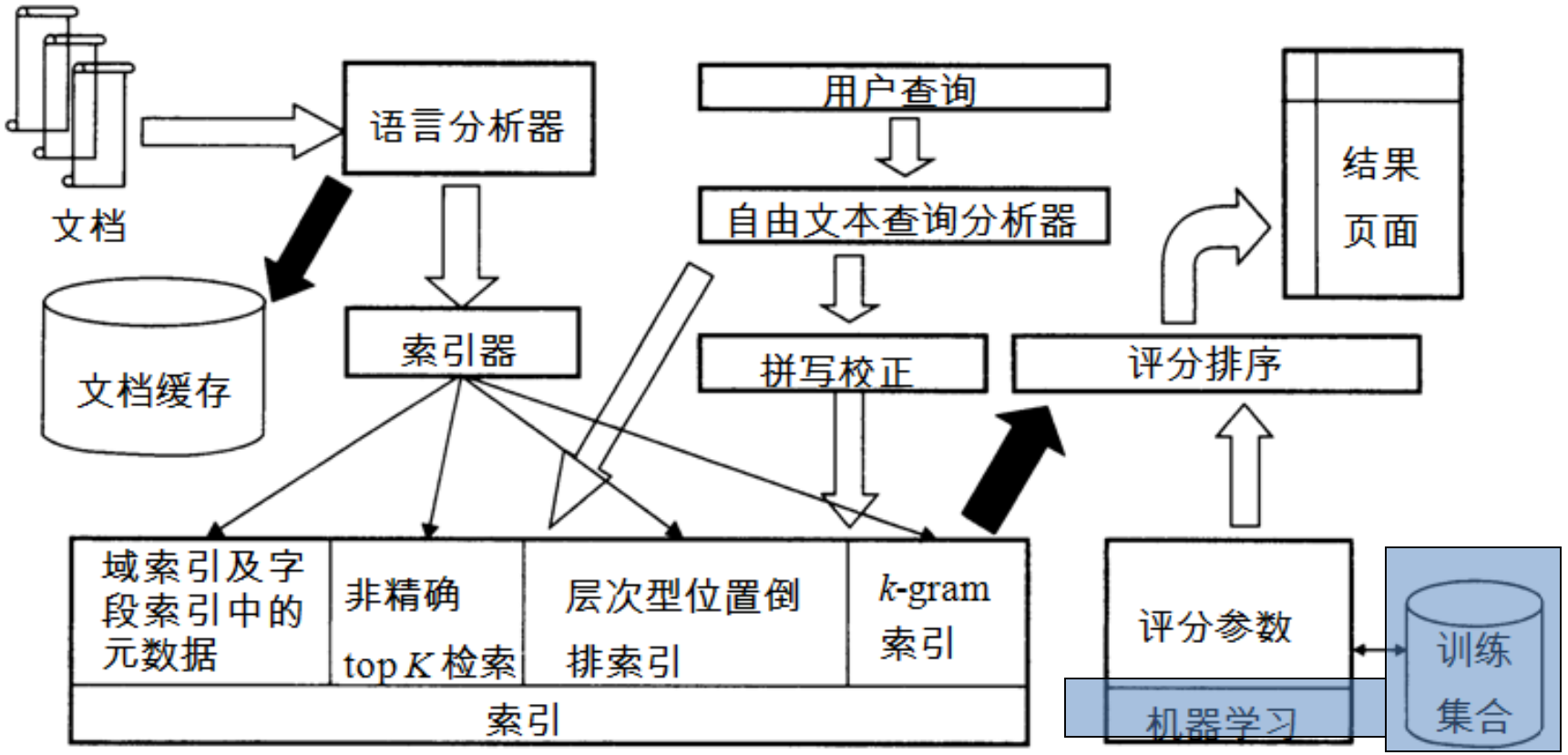
课程内容

- 第1章　绪论
- 第2章 布尔检索及倒排索引
- 第3章 词典查找及扩展的倒排索引
- 第4章 索引构建和索引压缩
- 第5章 向量模型及检索系统
- **第6章 检索的评价**
- 第7章 相关反馈和查询扩展
- 第8章 概率模型
- 第9章 基于语言建模的检索模型
- 第10章 文本分类
- 第11章 文本聚类
- 第12章 Web搜索
- 第13章 多媒体信息检索
- 第14章 其他应用简介

提纲

1. 上一讲回顾
2. 检索系统的评价概述
3. 无序检索结果的评价
4. 有序检索结果的评价
5. 为IR系统构建测试集
6. 检索结果的展示

搜索系统组成


文档
语言分析器
用户查询
自由文本查询分析器
结果
页面
文档缓存
索引器
拼写校正
评分排序
域索引及字
段索引中的
元数据
非精确
top K 检索
层次型位置倒
排索引
k-gram
索引
索引
评分参数
机器学习
训练
集合

综合评分

- 已经介绍的评分函数有余弦相似度、静态得分、近邻性等。

- 如何将这些评分组合才是最优的？
- 通用方法——机器学习

机器学习有下面几种定义：“机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能”。“机器学习是对能通过经验自动改进的计算机算法的研究”。“机器学习是用数据或以往的经验，以此优化计算机程序的性能标准。”一种经常引用的英文定义是：A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

提纲

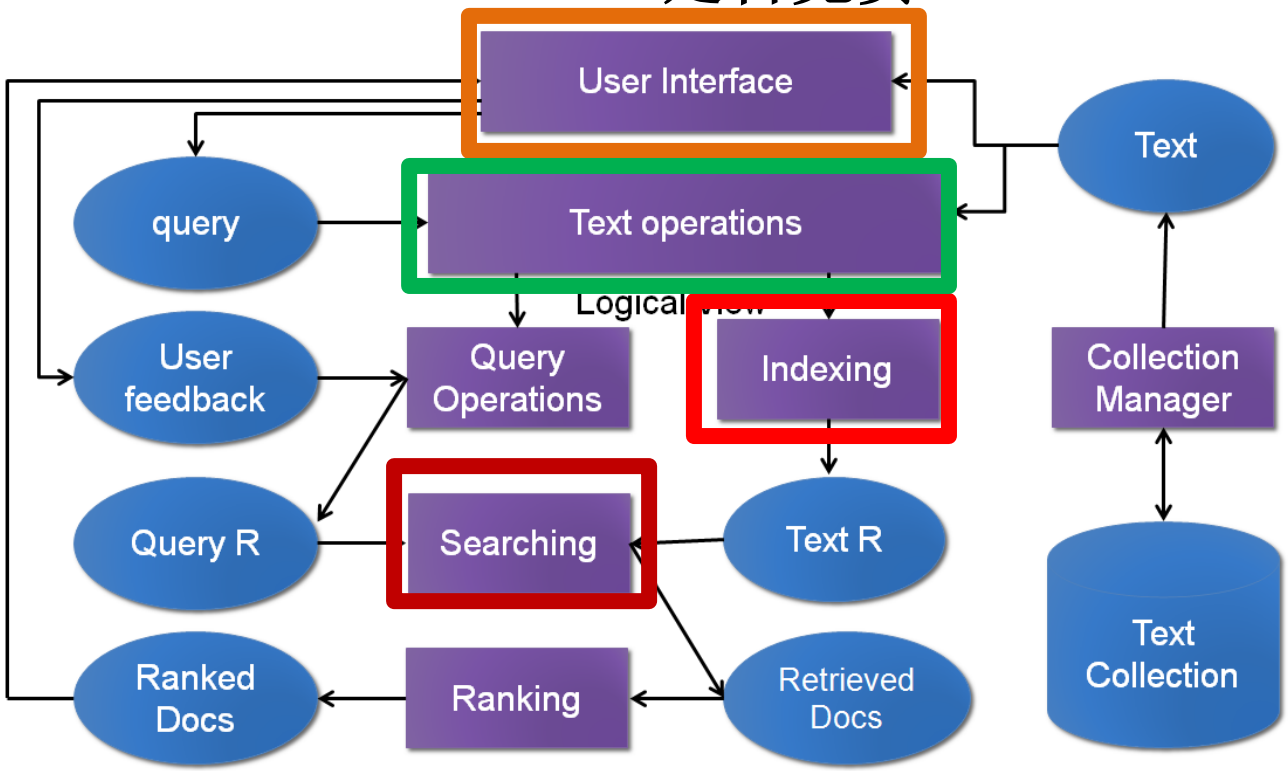
1. 上一讲回顾
2. 检索系统的评价概述
3. 无序检索结果的评价
4. 有序检索结果的评价
5. 为IR系统构建测试集
6. 检索结果的展示

为什么要评价IR？

- 前面各章介绍了信息检索系统设计中的各种方法。怎样才能知道其中哪些技术在哪些应用中有效？
  - 信息检索已经发展成为一门高度经验性的学科，需要在具有代表性的文档集上进行全面细致的评价，从而论证新技术的应用所带来的性能提升。
- 通过评价可以判断不同技术的优劣，不同因素对系统的影响，从而促进本领域研究水平的不断提高。
- 信息检索系统的目标是较少消耗情况下尽快、全面返回准确的结果。

搜索引擎的评价

- 建立索引的速度
  - 每小时索引的文档数量
  - 平均的文档大小
- 搜索的速度
  - 和索引大小相关
- 查询语言的表达能力
  - 是否能表达复杂的信息需求
  - 对复杂查询的处理速度
- 流畅和清晰的用户界面
- 是否免费？



搜索引擎的评价

- 上述的评价标准都是可以定量的
  - 我们可以测量速度或者索引大小

- 关键的评价标准：用户满意度
  - 用户满意度如何定义？
  - 搜索引擎响应速度和索引的覆盖范围是要考虑的因素
  - 但是如果结果不能让用户满意，响应速度再快，也是没有意义的
- 需要一种定量的方法来衡量用户满意度

如何用客观的 measurement 给出主观的满意度

用户满意度的衡量

- 关键问题：我们要使哪种用户满意？
  - 根据搜索服务的不同定位而异
- Web搜索引擎
  - 用户通过搜索引擎发现自己想要的东西，以后会继续使用这个搜索引擎
    - 可以统计用户的“回头率”
- 电子商务网站
  - 用户发现自己想要的东西，就会购买
    - 可以统计用户购买所花费时间，以及统计购买的用户占总的搜索的用户的百分比
- 企业：关心“用户的生产力”
  - 用户使用搜索引擎寻找信息，能节省多少时间？
  - 也需要考虑其他的准则：访问的安全性，访问的广度

满意度是很难衡量的

- 最通常的度量：搜索结果的相关度
  - 用搜索结果的相关度这个客观度量来替代对满意度的评估

- →如何衡量相关度？
- 衡量相关度需要3个要素：
  1. 评测文档集合
  2. 评测查询集合
  3. 对每个查询的每个返回文档做出“相关”或者“不相关”的评价（有些也可能不是二值的）

信息检索系统的评价

- 需要注意的是，信息需求用查询来表示，但相关性是相对于信息需求而言的，而不是相对于查询而言。

- 例如
  - 信息需求：在降低心脏病发作的风险方面，饮用红葡萄酒是否比饮用白酒更有效？
  - 查询：白酒 红酒 心脏病 有效
  - 在对返回的文档进行评估时，应当考虑是否满足信息需求

标准测试集

- CRANFIELD Cranfield 测试集
- TREC（Text Retrieval Conference）
  - TREC - National Institute of Standards and Technology（美国国家标准技术研究所，NIST)长期维护了一个大规模的IR测试环境。评测文档集合包含路透社和其他文档集合。在这个框架下定义了很多任务，每个任务都有自己的测试集
- NTCIR 日本国立情报研究所的信息检索测试集
- CLEF 跨语言评价论坛
- Reuters 语料
- 20 Newsgroups

小结：IR系统评价

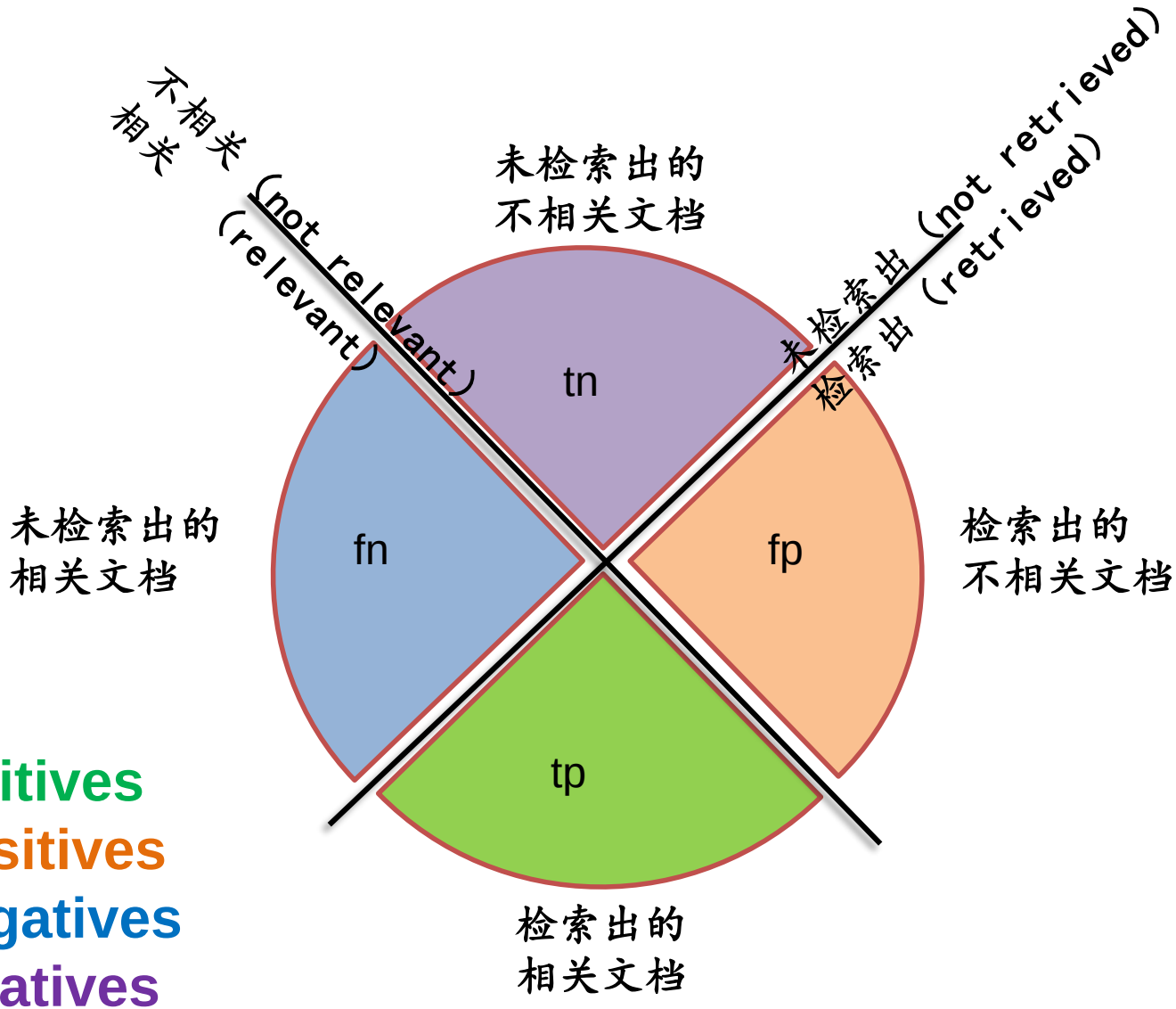
- 思路：用搜索结果的相关度这个客观度量来替代对满意度的评估

- 采用常规的方式来度量IR系统的效果，需要一个测试集（test collection），它由3 个部分构成：
  - (1) 一个文档集；
  - (2) 一组用于测试的信息需求集合，信息需求可以表示成查询；
  - (3) 一组相关性判定结果，对每个查询—文档对而言，通常会赋予一个二值判断结果——要么相关（relevant），要么不相关（nonrelevant）。

提纲

1. 上一讲回顾
2. 检索系统的评价概述
3. 无序检索结果的评价
4. 有序检索结果的评价
5. 为IR系统构建测试集
6. 检索结果的展示

按照文档“是否相关”“是否被检索出”划分

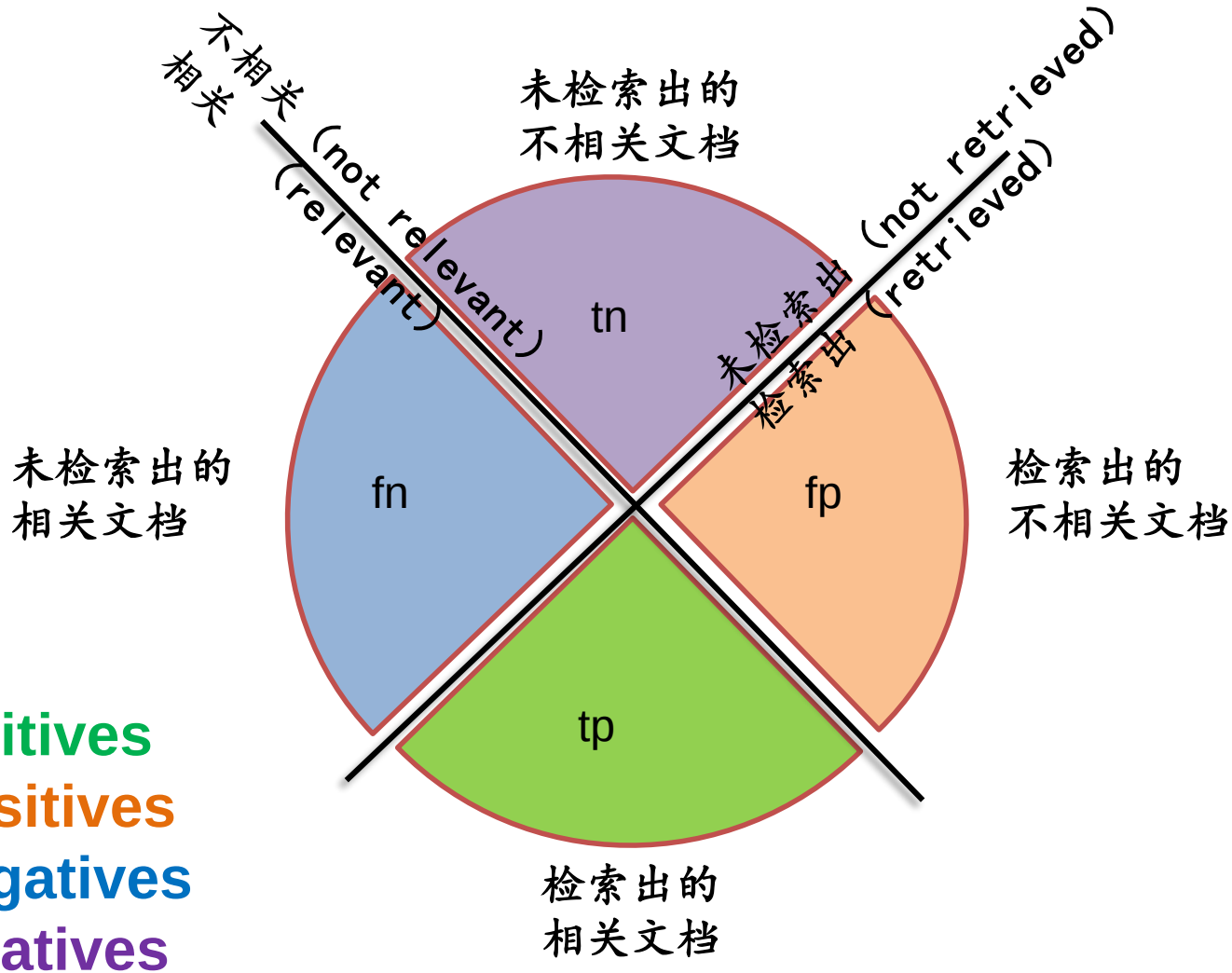


tp: true positives
fp: false positives
fn: false negatives
tn: true negatives

精确率

文档集中所有判断正确的文档所占的比例

精确率：(tp + tn) / (tp + fp + fn + tn)



tp: true positives

fp: false positives

fn: false negatives

tn: true negatives

精确率(Accuracy)指标

- 文档集中所有判断正确的文档所占的比例。
- 精确率：（tp + tn) / (tp + fp + fn + tn)

- 精确率是机器学习中模式分类的一个常用评价标准

但是它对信息检索的结果评价不是很有用。

绝大多数情况下，信息检索中的数据存在着极度的不均衡性，比如通常情况下，超过99.9%的文档都是不相关文档。这样的话，一个**简单地将所有的文档都判成不相关文档的系统就会获得非常高的精确率值**，从而使得该系统的效果看上去似乎很好。

人们使用搜索引擎，总是希望找到一些有用的信息，即使有些不相关的信息也是可以容忍的

查准率和查全率

- 查准率/正确率：返回的相关文档占返回文档总数的百分比
- 查全率/召回率：返回的相关文档占所有相关文档的百分比。

	Relevant	Nonrelevant
Retrieved	真正例 (true positives, tp)	伪正例 (false positives, fp)
Not Retrieved	伪反例 (false negatives,fn)	真反例 (true negatives,tn)

查准率/正确率Precision　　P = tp/(tp + fp)

查全率/召回率Recall　　R = tp/(tp + fn)

正确率和召回率示例

- 查询Q，本应该有100篇相关文档，某个系统返回200篇文档，其中80篇是真正相关的文档
- Recall=80/100=0.8
- Precision=80/200=0.4
- 结论：召回率较高，但是正确率较低

$$\text{Precision} = \frac{\text{返回结果中相关文档的数目}}{\text{返回结果的数目}} = P(\text{relevant} \mid \text{retrieved})$$

$$\text{Recall} = \frac{\text{返回结果中相关文档的数目}}{\text{所有相关文档的数目}} = P(\text{retrieved} \mid \text{relevant})$$

关于查准率和查全率的讨论

- “宁可错杀一千，不可放过一人”➔偏重查全率，忽视正确率。冤杀太多。
- 例如，判断是否有罪：
  - 如果没有证据证明你无罪，那么判定你有罪。
  - ➔查全率高，有些人受冤枉
  - 如果没有证据证明你有罪，那么判定你无罪。
  - ➔查全率低，有些人逍遥法外
- 不同的应用、不用的用户对两者的要求不一样
  - 垃圾邮件过滤：宁愿漏掉一些垃圾邮件，但是尽量少将正常邮件判定成垃圾邮件。
  - 有些用户希望返回的结果全一点，他有时间挑选；有些用户希望返回结果准一点。

查准率和查全率的问题

正确率（查准率）召回率（查全率）的融合

- 两个指标分别衡量了系统的某个方面，但是也为比较带来了难度，究竟哪个系统好？
  - 典型的Web 检索用户希望第一页的所有的结果都是相关的，也就是说他们非常关注高正确率，而对是否返回所有的相关文档并没有太大的兴趣。相反地，一些专业的搜索人士（如律师助手、情报分析师等）却往往重视高召回率，有时甚至宁愿忍受极低的正确率也要获得高的召回率。往往需要在这两个指标之间形成某种折衷。
- 解决方法：单一指标，将两个指标融成一个指标

一个综合评价准则：平衡F值（balanced F measure）

- F值是查准率和查全率的加权调和平均数

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

- 通常使用平衡的F_1值
  - β = 1 or α = ½

为什么要使用调和平均？

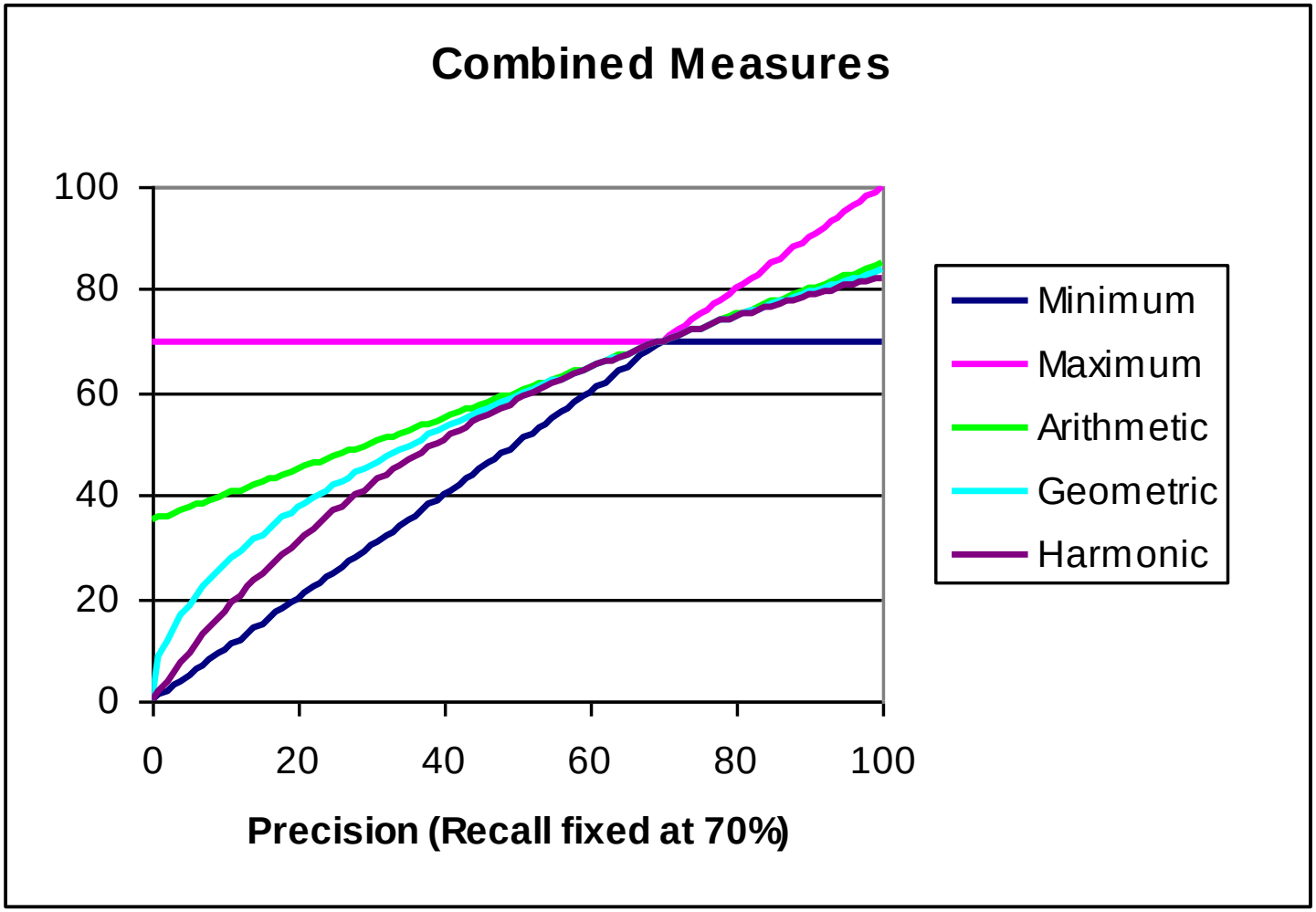
为什么使用调和平均计算F值

- 为什么不使用其他平均来计算F，比如算术平均
- 如果采用算术平均计算F值，那么一个返回全部文档的搜索引擎的F值就不低于50%，这有些过高。
- 做法：不管是P还是R，如果十分低，那么结果应该表现出来，即这样的情形下最终的F值应该有所惩罚
- 采用P和R中的最小值可能达到上述目的
- 但是最小值方法不平滑而且不易加权
- 基于调和平均计算出的F 值可以看成是平滑的最小值函数

调和平均数（harmonic mean）

调和平均比较“保守”：调和平均小于算数平均和几何平均

$F_{\beta=1}$ 和其他平均数的比较



查准率和查全率的问题

关于召回率（查全率）的计算

- 对于大规模语料集合，列举每个查询的所有相关文档是不可能的事情，因此，不可能准确地计算召回率
- 缓冲池(Pooling)方法：对多个检索系统的Top N个结果组成的集合进行人工标注，标注出的相关文档集合作为整个相关文档集合。这种做法被验证是可行的(可以比较不同系统的相对效果)，在TREC会议中被广泛采用。

查准率和查全率的问题

无序的缺陷

- 两个指标都是基于(无序)集合进行计算，并没有考虑序的作用
  - 举例：两个系统，对某个查询，返回的相关文档数目一样都是10，但是第一个系统是前10条结果，后一个系统是最后10条结果。显然，第一个系统优。但是根据上面基于集合的计算，显然两者指标一样。
  - 解决方法：引入序的作用

小结：无序检索结果的评价

- 为什么通常使用P、R、F而不使用精确率？
  - 由于和查询相关毕竟占文档集的极少数，所以即使什么都不返回也会得到很高的精确率。什么都不返回可能对大部分查询来说可以得到 99.99%以上的精确率
  - 用户希望找到某些文档并且能够容忍一定的不相关性，返回一些即使不好的文档也比不返回任何文档强

查准率/正确率Precision　　P = tp/(tp + fp)

查全率/召回率Recall　　R = tp/(tp + fn)

平衡F 值（balanced F measure）

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

提纲

引入序的作用

- **R-Precision：** 检索结果中，在所有相关文档总数位置上的准确率，如某个查询的相关文档总数为80，计算检索结果中**在前80篇文档的正确率**。

系统1，查询1	d3√	d6 √	d8	d10	d11
系统2，查询1	d6 √	d7	d2	d9 √	

- **在召回率(R)相同情形下，系统1和系统2的P不同**

评价排序后的结果

- P、R、F值都是基于集合的评价方法，它们都利用无序的文档集合进行计算。
  - →如果搜索引擎输出为有序的检索结果时，需要扩展。

- 对于一个特定检索词的有序检索结果
  - 系统可能返回任意数量的结果（=N）
  - 考虑Top k返回的情形(k=0, 1, 2, …, N)
  - 则每个k的取值对应一个R和P

- →可以计算得到查准率-查全率曲线

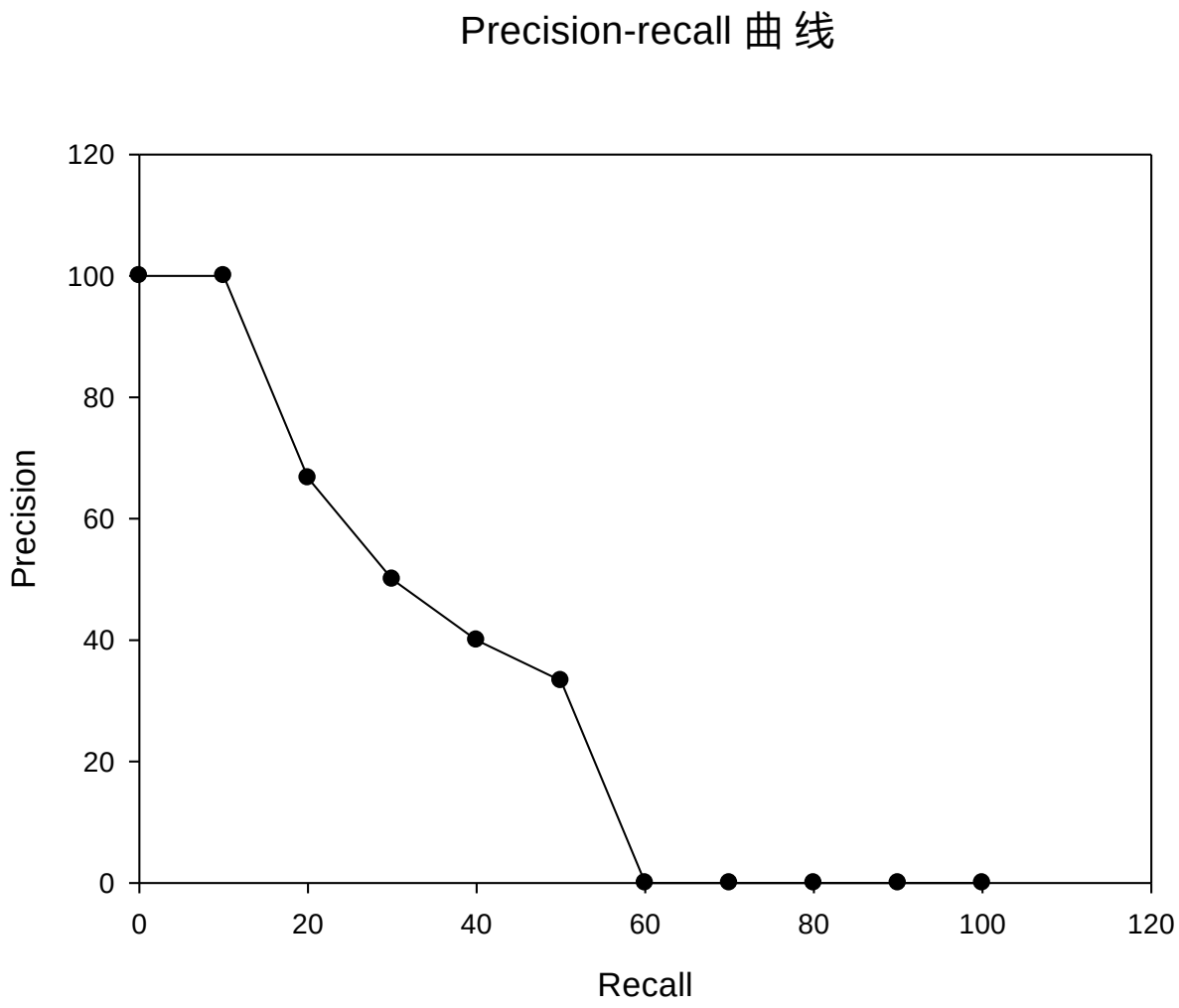
P-R曲线的例子

10个相关文档

- 某个查询q的标准答案集合为：
  Rq={d3,d5,d9,d25,d39,d44,d56,d71,d89,d123}
- **某个IR系统对q的检索结果如下**：

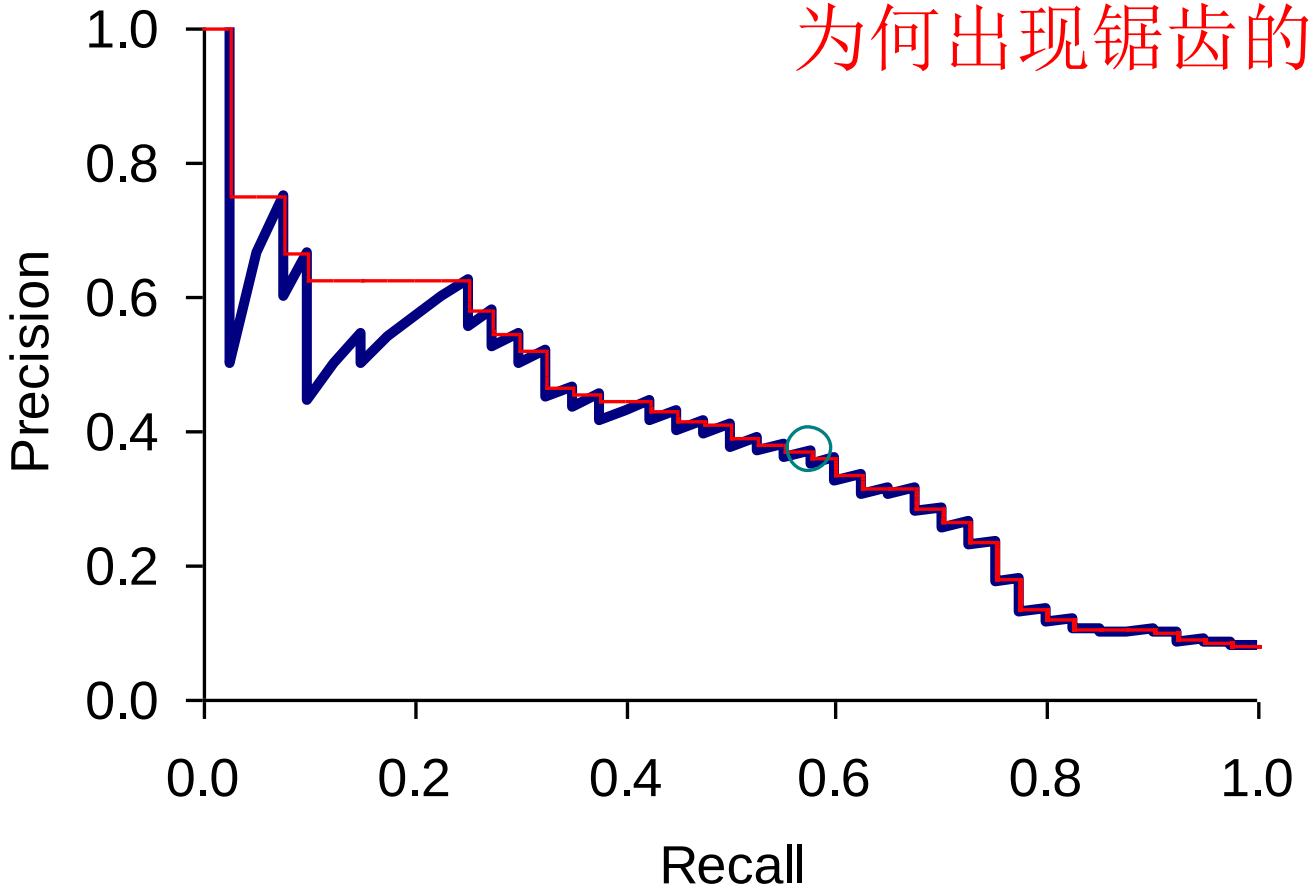
1. d123 R=0.1,P=1	6. d9 R=0.3,P=0.5	11. d38
2. d84	7. d511	12. d48
3. d56 R=0.2,P=0.67	8. d129	13. d250
4. d6	9. d187	14. d113
5. d8	10. d25 R=0.4,P=0.4	15. d3 R=0.5,P=0.33

上例的P-R曲线

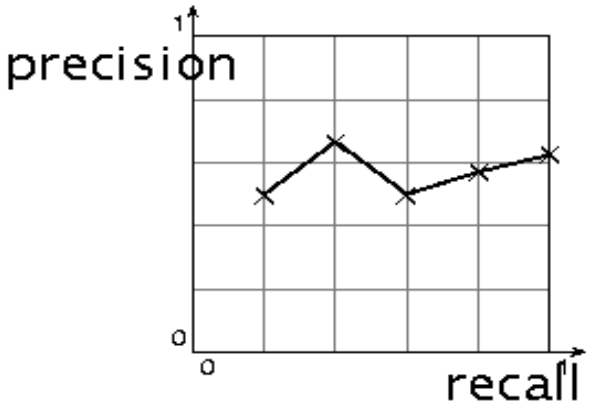

Precision-recall 曲 线
Precision
Recall
120
100
80
60
40
20
0
0
20
40
60
80
100
120

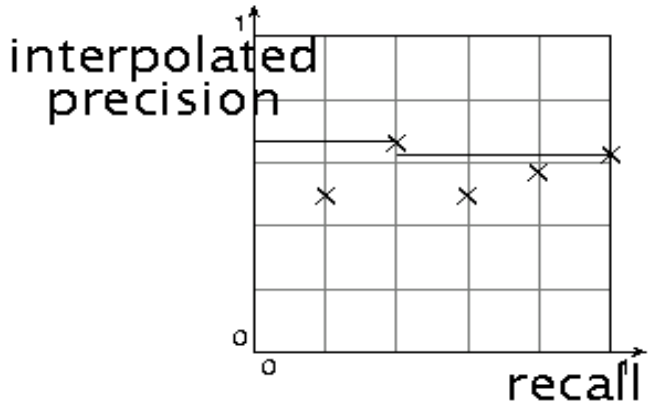
一般的P-R曲线

为何出现锯齿的形状？



插值查准率





原始的曲线常常呈现锯齿状（左图），这是很正常的。因为如果第（K+1）篇文档不相关，则查全率和前k篇文档的查全率是一样的，但是准确率降低了，所以曲线会下降。如果第（K+1）篇文档相关，则查全率和查准率都上升。如此就会出现锯齿状。

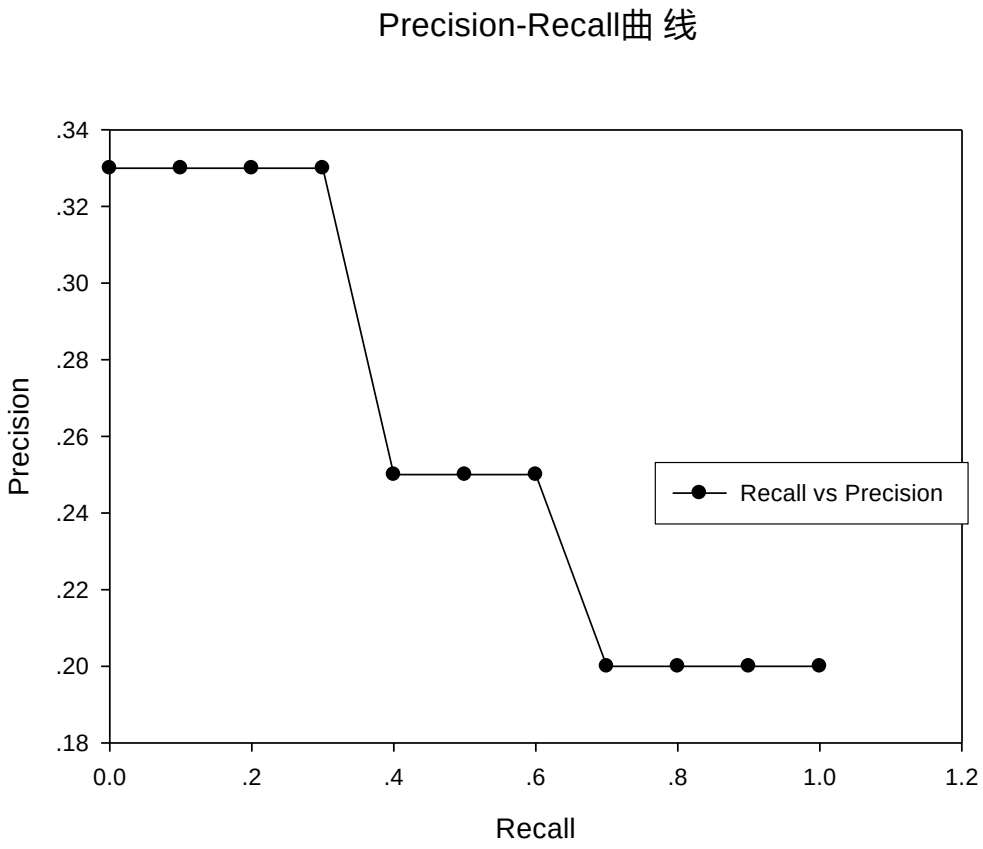
我们需要对去掉锯齿，进行平滑。采用**插值查准率**(interpolated precision)，记为p_{interp}

在查全率为r的位置的插值查准率，定义为查全率不小于r的位置上的查准率的最大值，即 $p_{\text{int }erp}(r) = \max_{r' \geq r} p(r')$ （见右图）

P-R 曲线的插值问题

- 对于前面的例子，假设Rq={d3,d56,d129}
  - **d3** R=1,P=0.2; **d56** R=0.33,P=0.33; **d129** R=0.66, P=0.25
- 不存在10%, 20%,…,90%的召回率点，而只存在33.3%, 66.7%, 100%三个召回率点
- 在这种情况下，需要利用存在的召回率点对不存在的召回率点进行插值(interpolate)
- 对于t%，如果不存在该召回率点，则定义t%为从t%到(t+10)%中**最大的正确率值**。
- 对于上例，0%,10%,20%,30%上正确率为0.33，40%~60%对应0.25，70%以上对应0.2

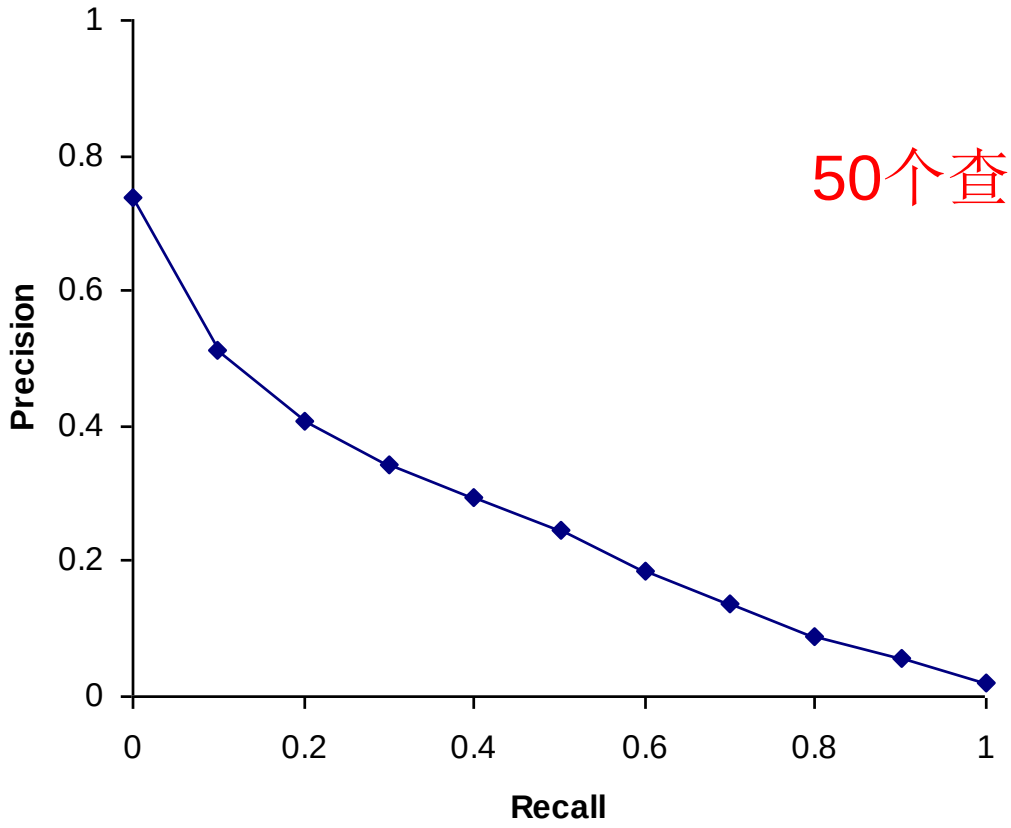
插值后的P-R曲线图


Precision-Recall曲 线
Precision
.34
.32
.30
.28
.26
.24
.22
.20
.18
0.0
.2
.4
.6
.8
1.0
1.2
Recall
Recall vs Precision

P-R曲线的优缺点

- 曲线图虽然好，但是评价标准如果能浓缩成一个数字，就更加清晰明了
  - **固定检索等级的查准率**
    - Precision@k：前k个结果的查准率
    - 对大多数的web搜索是合适的，因为用户看重的是在前几页中有多少好结果
    - 但是这种*平均的方式不好*，是通常所用指标中最不稳定的
  - **11点插值查准率**
    - 对每个信息需求，插值的正确率定义在0、0.1、0.2、…、0.9、1共十一个召回率水平上
    - 对于每个召回率水平，对测试集中每个信息需求在该点的插值正确率求算术平均。

典型的11点插值正确率-召回率平均曲线



50个查询的平均

更多的评价准则：AP

- 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AP**：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
  - **插值的AP**:在召回率分别为0,0.1,0.2,…,1.0的十一个点上的正确率求平均，等价于11点平均
  - **只对返回的相关文档进行计算的AP**
    AP=(1/1+2/2+3/5+4/10+5/20)/5，倾向那些快速返回结果的系统，没有考虑召回率

更多的评价准则：MAP

- 平均查准率均值 Mean Average Precision（MAP）
  - 在每个相关文档位置上查准率的平均值，被称为平均查准率（AP）
  - 对所有查询求平均，就得到平均查准率均值（MAP）

$$\mathrm{MAP}(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\mathrm{Precision}(R_{jk})$$

- 参数说明
  - Q为信息需求，$q_j \in Q$ 所对应的所有相关文档集合为$\{d_1, d_2, \cdots, d_{mj}\}$，$R_{jk}$是查询$q_j$的返回结果、该结果中包含$\{d_1, d_2, \cdots, d_k\}$而不含有$d_{k+1}$及以后的相关文档

更多的评价准则：R正确率

- R-Precision
  - 检索结果中，在所有相关文档总数位置上的准确率。如某个查询的相关文档总数为Rel，返回的结果中前|Rel|个中r个是相关文档，则R正确率是r/|Rel|。
  - **R正确率**能够适应不同的相关文档集的大小
    - 例：Rel=8；r=8。此时R正确率是1，但是P@20=0.4
  - 一个**完美**的系统的**R-precision=1**

更多的评价准则： GMAP

- GMAP(Geometric MAP)：TREC2004 Robust 任务引进
- 先看一个例子

系统	Topic	AP	Increase	MAP
系统A	Topic 1	0.02	-	0.113
	Topic 2	0.03	-	
	Topic 3	0.29	-	
系统B	Topic 1	0.08	+300%	0.107
	Topic 2	0.04	+33.3%	
	Topic 3	0.20	-31%	

- 从MAP来看，系统A好于系统B，但是从每个查询来看，3个查询中有2个Topic B比A有提高，其中一个提高的幅度达到300%
- 几何平均值 $$GMAP = \sqrt[n]{\prod_{i=1}^{n} AP_i} = \exp(\frac{1}{n}\sum_{i=1}^{n} \ln AP_i)$$
- 上面那个例子 $GMAP_a$=0.056， $GMAP_b$=0.086 $GMAP_a$<$GMAP_b$
- GMAP和MAP各有利弊，可以配合使用，如果存在难Topic时，GMAP更能体现细微差别

更多的评价准则： NDCG (Normalized Discounted Cumulative Gain，归一化折损累积增益)

- 每个文档不仅仅只有**相关**和**不相关**两种情况，**而是有相关度级别**，比如0, 1, 2, 3。

  我们可以假设，对于返回结果：

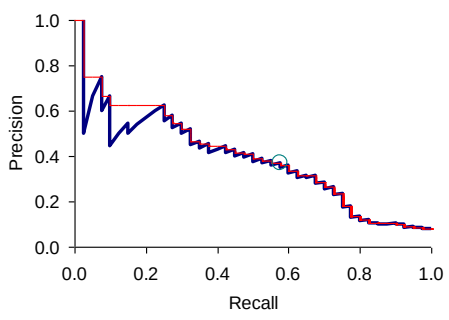
  - **相关度级别越高的结果越多越好**
  - **相关度级别越高的结果越靠前越好**

$$\mathrm{NDCG}(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)}-1}{\log_2(1+m)}$$

  - R（j,m）是**评价人员给出的文档d对查询j的相关性得分**，Z_{kj}是归一化因子，保证对完美系统NDCG的值为1，m是返回文档的位置

小结：有序检索结果的评价

- 现有评价体系远没有达到完美程度
  - 对评价的评价研究
  - 指标的相关属性(公正性、敏感性)的研究
  - 新的指标的提出(新特点、新领域)
  - 指标的计算(比如Pooling方法中如何降低人工代价？)



$$\mathrm{MAP}(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\mathrm{Precision}(R_{jk})$$

$$\mathrm{NDCG}(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}Z_{kj}\sum_{m=1}^{k}\frac{2^{R(j,m)}-1}{\log_2(1+m)}$$

提纲

1. 上一讲回顾
2. 检索系统的评价概述
3. 无序检索结果的评价
4. 有序检索结果的评价
5. 为IR系统构建测试集
6. 检索结果的展示

常用的测试集

Collection	*NDocs*	*NQrys*	*Size (MB)*	*Term/Doc*	*Q-D RelAss*
ADI	82	35			
AIT	2109	14	2	400	>10,000
CACM	3204	64	2	24.5	
CISI	1460	112	2	46.5	
Cranfield	1400	225	2	53.1	
LISA	5872	35	3		
Medline	1033	30	1		
NPL	11,429	93	3		
OSHMED	34,8566	106	400	250	16,140
Reuters	21,578	672	28	131	
TREC	740,000	200	2000	89-3543	» 100,000

从文档集合如何构建测试集

- 需要“用于测试的查询”和“相关性的判定”
  - **用于测试的查询**
    - 必须和测试文档集合有密切关系
    - 最好由领域的专家设计
    - 随机的查询并不好
  - **相关性的判定**
    - 人工判定耗时较长
    - 使用一组人进行判定是否是最好的方式？

用户判定的有效性

- 只有在用户的评定一致时，相关性判定的结果才可用；
- 如果结果不一致，那么不存在标准答案无法重现实验结果；
- 如何度量不同判定人之间的一致性？
- → Kappa 指标

相关性判定之间的一致性

- Kappa统计量
  - 衡量不同人做出的相关性判定之间的一致性
  - 对随机一致性比率的简单校正
- Kappa = [P(A) - P(E)] / [1 - P(E)]
- P(A) - 实际观察到的一致性判断比率
- P(E) - 随机情况下所期望的一致性判断的比率
- Kappa = 0 和随机判断的情况一样，Kappa = 1 不同人做出的相关性判定完全一致.
- *k在* [2/3, 1.0]时，判定结果是可以接受的
- 如果k值比较小，那么需要对判定方法进行重新设计

计算kappa统计量

通常采用边缘统计量（marginal statistics）来计算随机一致性比率

		第二个人的相关性判定		
		Yes	No	Total
第一个人的相关性判定	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

观察到的两个人的一致性判断比率

$P(A) = (300 + 70)/400 = 370/400 = 0.925$

边缘统计量

$P(nonrelevant) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$

$P(relevant) = (320 + 310)/(400 + 400) = 630/800 = 0.7878$

两个人的随机一致性比率

$P(E) = P(nonrelevant)^2 + P(\text{relevant})^2 = 0.2125^2 + 0.7878^2 = 0.665$

Kappa统计量

$\kappa = (P(A) - P(E))/(1 - P(E)) = (0.925 - 0.665)/(1 - 0.665) = 0.776$ (still in acceptable range)

大型搜索引擎的评价

- Web下召回率难以计算
- 搜索引擎常使用top *k*的正确率来度量, 比如, *k* = 10 . . .
- . . . 或者使用一个考虑返回结果所在位置的指标，比如正确答案在第一个返回会比第十个返回的系统给予更大的指标
- 搜索引擎也往往使用非相关度指标
  - 比如：第一个结果的点击率
  - 仅仅基于单个点击使得该指标不太可靠 (比如你可能被检索结果的摘要所误导，等点进去一看，实际上是不相关的) . . .
  - 当然，如果考虑点击历史的整体情况会相当可靠
- 举例: A/B 测试

A/B 测试

- 目标: 测试某个新引入的独立的创新点
- 先决条件：大型的搜索引擎已经在上线运行
- 方法：
  - 很多用户使用老系统，将一小部分(如 1%)流量被随机导向包含了创新点的新系统
  - 对新旧系统进行自动评价，并得到某个评价指标，比如判断第一个结果的点击率是否有提升
  - 于是，可以通过新旧系统的指标对比来判断创新点的效果
- 这也可能是大型搜索引擎最信赖的方法

SIGIR 2017 Honourable Mentions(最佳提名)

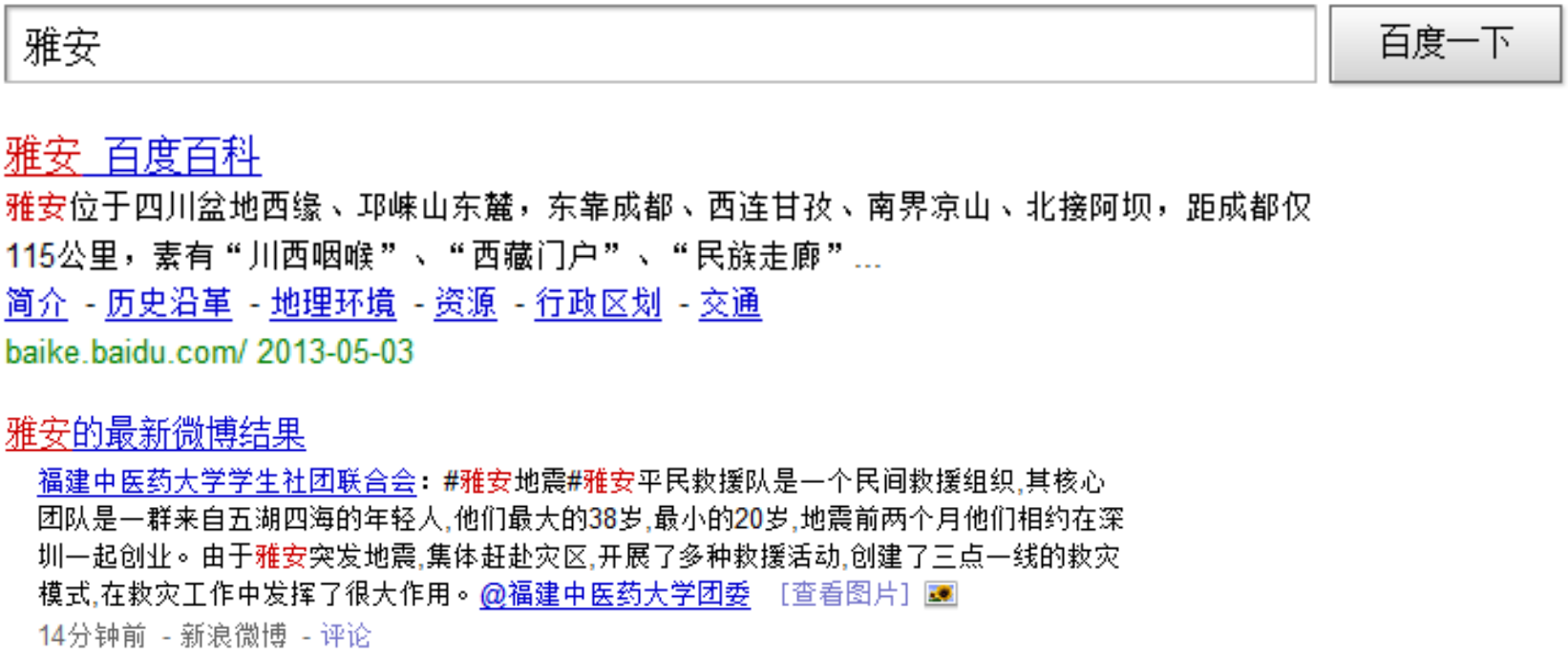
- IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models
  - Jun Wang (University College London), Lantao Yu (Shanghai Jiao Tong University), Weinan Zhang (Shanghai Jiao Tong University), Yu Gong (Alibaba Inc.), Yinghui Xu (Alibaba Inc.), Benyou Wang (Tianjin University), Peng Zhang (Tianjin University), Dell Zhang (Birkbeck, University of London)
- 评价指标设计一直是信息检索技术研究中的核心问题之一，而估计用户的期望收益与期望付出则是搜索用户行为模型的关键组成部分。受模型框架限制，当前几乎所有信息检索评价指标均无法做到同时将用户的期望收益和付出纳入会话终止条件的估计。针对这一问题，计算机系师生受流行电子游戏“Bejewed（中文名：宝石迷阵）”机制启发，设计了一个创新性的用户交互模型框架，将期望收益与付出因素重新建模，并把现有的绝大多数评价指标纳入这一框架的范畴。在真实用户行为数据上的实验表明，该框架比现有指标能够更好的预测用户满意程度。

提纲

1. 上一讲回顾
2. 检索系统的评价概述
3. 无序检索结果的评价
4. 有序检索结果的评价
5. 为IR系统构建测试集
6. 检索结果的展示

结果摘要

- 对与查询相关的检索结果排序后，我们可以展现一个列表
- 通常情况下，这个列表包含文档的标题和一段摘要



摘要

- 标题通常是从文档的元数据中自动抽取出来的
  - 这个描述信息非常重要，用户可以根据它来判断这个文档是不是相关
- 两种基本的摘要方法
  - 静态：一个文档的摘要是固定的，与查询无关
  - 动态：与查询相关。摘要说明了为什么这篇文档和查询相关

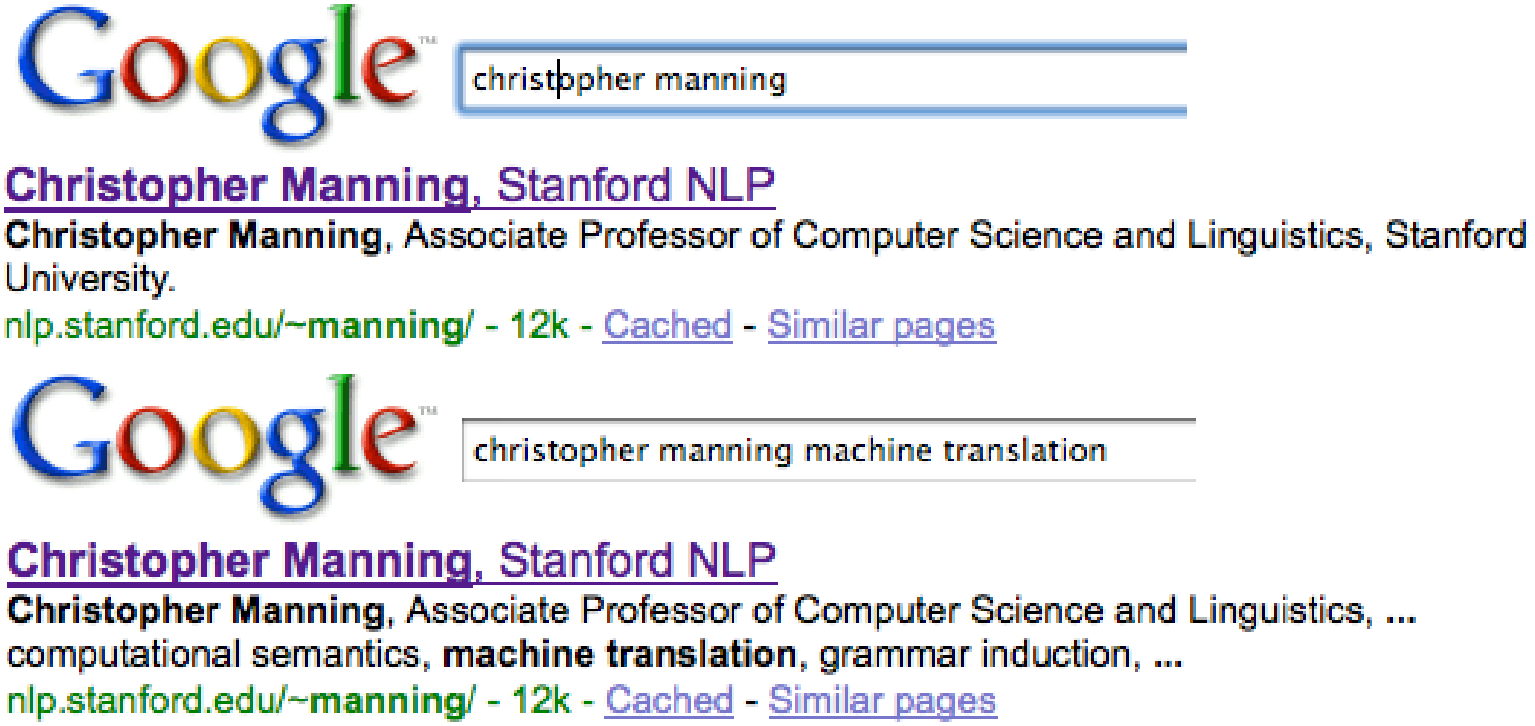
静态摘要

- 在典型的系统中，静态摘要是文档的一个子集
- 最简单的方法：文档的前若干个词汇
  - 在建立索引的时候就缓存好
- 更复杂的方法：从文档中抽取一些关键的句子
  - 用简单的自然语言处理的方法对句子进行打分

    用打分最高的几个句子组成摘要
- 最复杂的方法：用自然语言处理的方法合成摘要
  - 在IR系统中几乎不用

动态摘要

- 显示文档中包含查询词的一句或者几句文字
  - “KWIC”片段： Keyword in Context presentation



动态摘要

https://www.baidu.com/s?wd=%E5%A5%8B%E6%96%97%E7%9A%84

Bai du 百度 奋斗的人生 百度一下 百度首页 消息 设

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约20,100,000个 搜索工具

习近平:只有奋斗的人生才称得上幸福的人生_新华每日电讯

习近平指出,只有奋斗的人生才称得上幸福的人生。奋斗是艰辛的,艰难困苦、玉汝于成,没有艰辛就不是真正的奋斗,我们要勇于在艰苦奋斗中净化灵魂、磨砺意志...

www.xinhuanet.com/mrdx... V3 - 百度快照

奋斗的人生的最新相关信息

带着总书记的嘱托 | 奋斗成就精彩人生 新浪新闻 12小时前

从2002年进入现在的公司工作,到2012年12月当选十二届广东省人大代表,再到2018年1月当选十三届全国人大代表,米雪梅靠奋斗成就了人生的精彩。...

【幸福的奋斗者】一生之计在于勤 凤凰网 3天前

奋斗的人生,不会虚度 求是理论网 3月14日

越奋斗越幸福 刘晓平:种草种出"花样人生" 衡阳广电网 2天前

奋斗的人生才称得上幸福的人生 新浪 3月17日

人这一生为什么要努力? - 知乎

从外曾祖父母算起,我们家族四代八十多口人已经在美国奋斗了半个多世纪。我们人人都在努力,我们一代好过一代。 如果你的人生起点不高,不管有人为你走过人生百...

https://www.zhihu.com/question... V2 - 百度快照

人文社会类书籍 展开

人生不设限

NickVujicic

著励志类

人生就是奋斗

人生智慧图书

奋斗改变人生

成功法则图书

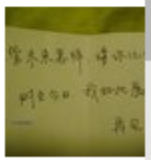
相关词汇 展开

信心

汉语词语诚心的意思

拼搏

拼死也要实现



唯美的句子

提供文章自由发表平台

动态摘要相关技术

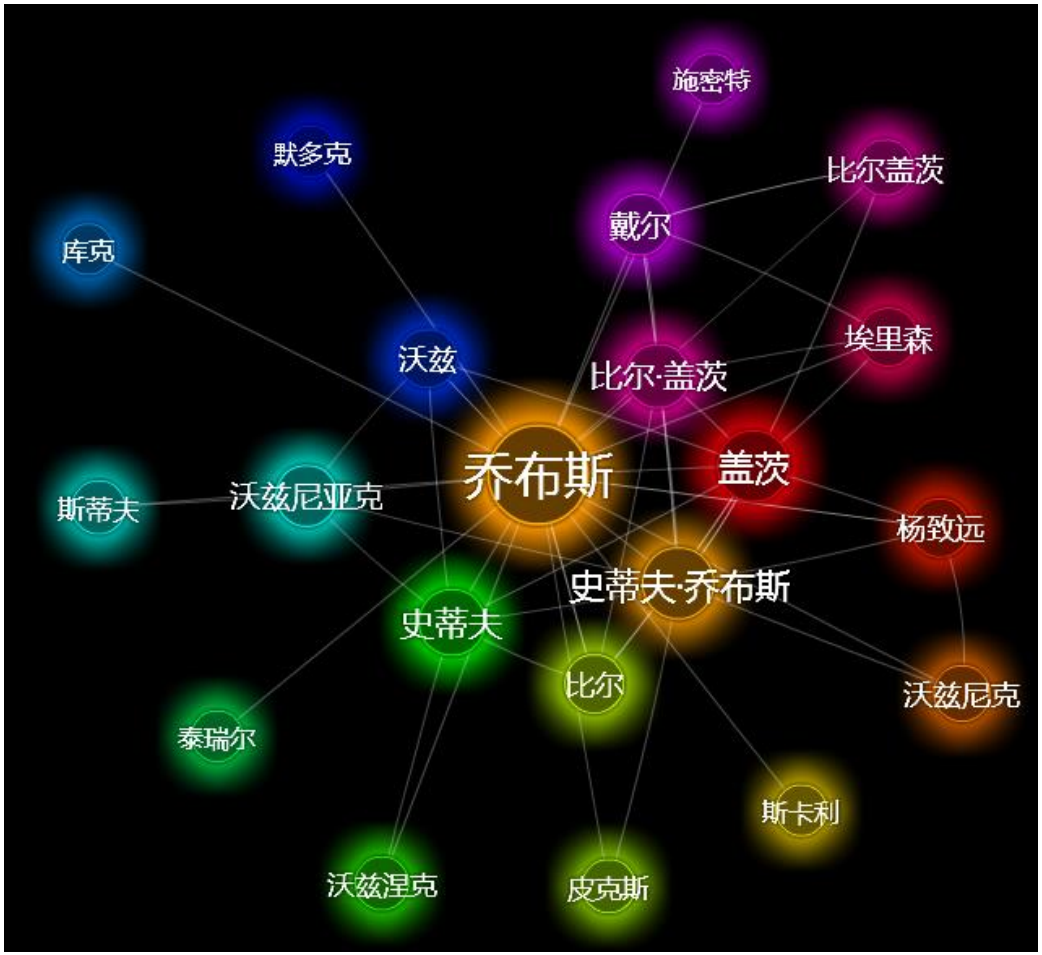
- 快速在文档中寻找包含查询词的“窗口”（范围）
- 根据查询对文档中上述窗口打分
  - 用多种特征，如窗口的长度，在文档中的位置，等等
  - 用一个打分函数融合多种特征
- 评价的挑战：对摘要的评价
  - 相关度的两两比较比单个文档的相关度判定简单

快捷链接

- 对导航性的查询，例如搜索“中国科学技术大学”，将会在页面上显示一些导航链接

其他的展现方式

- 人立方 http://renlifang.msra.cn/GuanxiMap.aspx



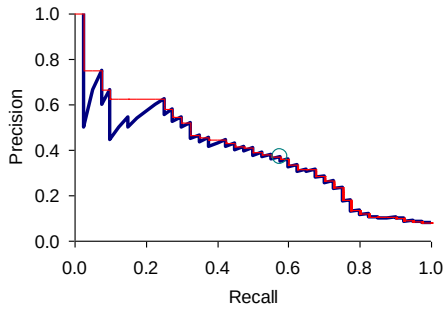
小结：检索的评价

- 无序检索结果的评价
  - P = tp/(tp + fp)
  - R = tp/(tp + fn)
  - 平衡F 值（balanced F measure）

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

- 有序检索结果的评价

$$\mathrm{MAP}(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\mathrm{Precision}(R_{jk})$$



$$\mathrm{NDCG}(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}Z_{kj}\sum_{m=1}^{k}\frac{2^{R(j,m)}-1}{\log_2(1+m)}$$

- 为IR系统构建测试集

$$kappa = \frac{P(A)-P(E)}{1-P(E)}$$

课后练习

- 习题8-7
- 习题8-10