Ch X

评价与相关性

- 评价的概念与意义
- 信息检索评价
- 评价的基本条件
- 缓冲池
- 统一测评

基本原理

- 文档集的划分
- 基本评价指标
  - P Precison 准确率/查准率
  - R Recall ratio 查全率/召回率
  - Fallout ratio 非相关检出率
  - P-R 关系
  - F值 调和平均值
- 单值评价指标
  - MAP Mean Average Precision 平均准确率均值
  - p@10
  - RP R-Precision

评价与相关性

评价的概念与意义

1. 评价：发现和收集关于某种活动的数据，从中判断该项活动的质量及达到预期目标程度的行为。简单地说，评价就是对系统的价值和效率进行测评
2. 信息检索系统的评价：根据给定的指标体系，采用一定的方法和程序，对信息检索系统的功能、特性和运营状况进行评测，或对有关假设、预期效益、性能值进行验证，以确定系统达到了何种水平、投入成本是否值得、是否可以改进和如何改进，乃至系统是否应生存下去

3. 信息检索评价的意义
    a. 了解已有检索系统的功能，找出缺陷并改进
    b. 比较各种检索系统的优劣
    c. 提高效率和效益
    d. 有助于新的检索系统的设计
    e. 丰富信息检索的理论

信息检索评价

1. 数据检索是“确定性”的；信息检索是“相关性”的
2. 信息检索中的“相关性”主要是指检索系统针对用户的信息需求从文档集合中检出的文档与用户需求之间的一种匹配关系

评价的基本条件

1. 一个文档集合C：系统将从该集合中按照用户查询检出相关文档
2. 一组用户查询$\{q_1, q_2, ..., q_n\}$：每个用户查询q_i描述了用户的信息需求
3. 对应每个用户查询的标准相关文档集$\{R_1, R_2,..., R_n\}$：该集合可由人工方式构造
4. 一组评价指标：这些指标反映系统的检索性能。通过比较系统实际检出的结果文档集和标准的相关文档集，对它们的相似性进行量化，得到这些指标值

缓冲池

1. 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
2. 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
3. 这种做法被验证是可行的，在TREC会议等多种测试中被广泛采用

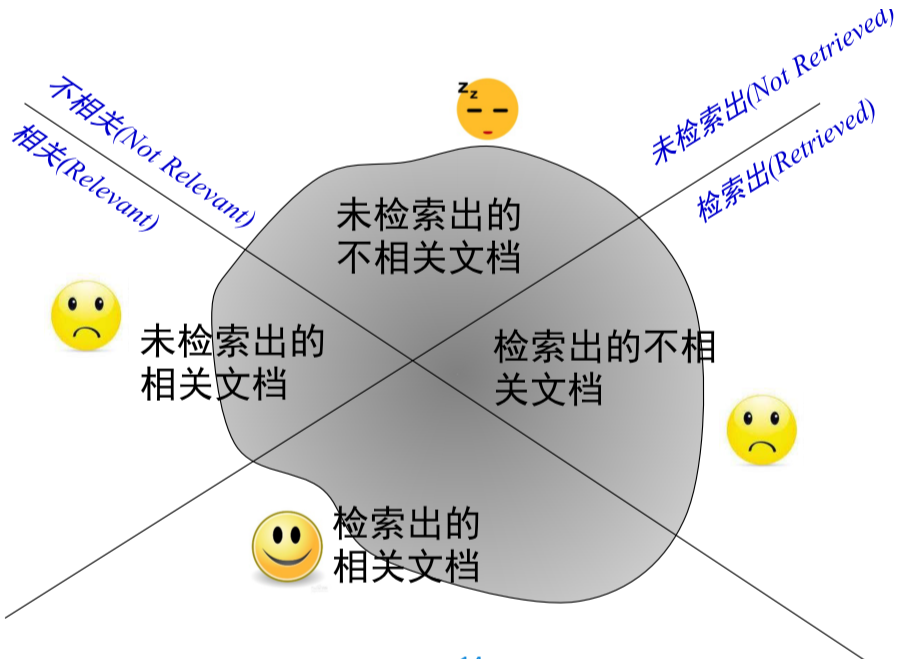
统一测评

1. 同一个算法在不同的数据条件下得到的结果差异很大
2. 没有统一的测试方法和共同的数据集合，几乎不可能比较不同算法

3. 数据采集需花费很大的人力物力，而由政府学术机构或者学术团体组织的开放技术评测，可以为科研提供一种统一的、普遍认可的评价基准和大型测试集，节省了各个研究者重复采集数据而造成的重复付出，对整个领域的科学研究和技术进步起到很大的推动作用
4. 通过技术评测可以提出新的研究问题

基本原理

文档集的划分

1. 

基本评价指标

P Precison 准确率/查准率

1. 检出的相关信息数量与检出的信息总量的比率

R Recall ratio 查全率/召回率

1. 检出的信息数量与检索系统中相关信息总量之间的比率

Fallout ratio 非相关检出率

1. 检出的非相关信息数量与系统中的非相关信息总量的比率

P–R 关系

1. 返回了大多数相关文档，但是包含很多垃圾
2. 返回最相关的文本，但是漏掉了很多相关文本

F值 调和平均值

1. 将准确率和召回率加权平均的评价方法
2. $$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2 \times P \times R}{P + R}$$

单值评价指标

MAP Mean Average Precision 平均准确率均值

1. 单个查询的平均准确率是逐个考察排序中每个新的相关文档，然后对其准确率值进行平均后的平均值
2. 查询集合的平均准确率是每个查询的平均准确率AP的平均值
3. 反映系统在全部查询上性能的单值指标
4. 系统检索出来的相关文档位置越靠前， MAP就可能越高
5. 如果系统没有返回相关文档，则MAP默认为0
6. $$MAP = \frac{1}{r}\sum_{i=1}^{r} \frac{i}{\text{第}i\text{个相关文档的位置}}$$

p@10

1. 系统对于查询返回的前10个结果的准确率
2. 对于搜索引擎系统来讲，由于没有一个搜索引擎系统能够保证搜集到所有的网页，所以召回率很难计算，因而准确率成为目前的搜索引擎系统主要关心的指标
3. 考虑到用户在查看搜索引擎结果时，往往希望在第一个页面（如10个结果）就找到自己所需的信息，因此P@10能比较真实有效地反映在真实应用环境下所表现的性能

RP R-Precision

1. 单个查询的R准确率是指检索出R篇文档时的准确率
2. R是当前检索中相关的文档总数
3. 查询集合中所有查询的R准确率是每个查询的R准确率的平均值
4. $$R-\Pr ecision=\frac{\text{前}R\text{篇文档中实际相关文档数}}{R}$$