中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

信息检索入门
Introduction to Information Retrieval

中国科学院计算技术研究所
王斌 骆卫华

2006.5

内容

- 信息检索的基本概念
- 信息检索的基本流程
- 信息检索的评价方法
- 信息采集
- 信息分析及索引
- 信息检索模型及其他相似度计算方法
- 查询扩展及相关反馈

内容

- 信息检索的基本概念←
- 信息检索的基本流程
- 信息检索的评价方法
- 信息采集
- 信息分析及索引
- 信息检索模型及其他相似度计算方法
- 查询扩展及相关反馈


information retrieval - Google 搜索 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹
地址(D) http://www.google.com/search?sourceid=navclient&hl=zh-CN&ie=UTF-8&rls=GGLD,GGLD:2006-09,GGLD:zh-CN&q=information+retrieval
转到 链接
Google ormation retrieval 搜索 14 个已拦截 拼写检查 选项 information retrieval
登录
Google
网页 图片 资讯 论坛 网页目录 更多 »
information retrieval
搜索 高级搜索 使用偏好
搜索所有网页 搜索所有中文网页 搜索简体中文网页
查询
网页
约有117,000,000项符合information retrieval的查询结果，以下是第1-10项。 （搜索用时 0.41 秒）
建议： 可直接输入拼音， Google会自动提示最符合的中文关键词 [细节]
information retrieval 的图书搜索结果
Information Retrieval - 作者： Mounia Lalmas - 352 页
Readings in Information Retrieval - 编者： Karen Sparck Jones, Peter Willett - 587 页
Visual Information Retrieval - 作者： Alberto del Bimbo - 270 页
Information Retrieval - [翻译此页 BETA]
An online book by CJ van Rijsbergen, University of Glasgow.
www.dcs.gla.ac.uk/Keith/Preface.html - 7k - 网页快照 - 类似网页
Information Retrieval - [翻译此页 BETA]
Online text of a book by Dr. CJ van Rijsbergen of the University of Glasgow covering advanced topics in information retrieval.
www.dcs.gla.ac.uk/~iain/keith/ - 5k - 网页快照 - 类似网页
Modern Information Retrieval - [翻译此页 BETA]
A recent IR book, covering algorithms, implementation, query languages, user interfaces, and multimedia...
www.sims.berkeley.edu/~hearst/irbook/ - 9k - 网页快照 - 类似网页
Information Retrieval Research - SearchTools Topics - [翻译此页 BETA]
Research on Information Retrieval, Library Experience and other academic work related to web site search tools.
www.searchtools.com/info/info-retrieval.html - 22k - 网页快照 - 类似网页
UMASS Amherst: Center for Intelligent Information Retrieval - [翻译此页 BETA]
University of Massachusetts research lab focused on efficient access to large, heterogeneous, distributed, text and multimedia databases.
ciir.cs.umass.edu/ - 6k - 网页快照 - 类似网页
SIGIR: Information Retrieval - [翻译此页 BETA]
"Addresses issues ranging from theory to user demands in the application of computers to the acquisition,...
www.acm.org/sigir/ - 类似网页
相关的结果
完毕
Internet

信息检索

- Information Retrieval(IR)：从文档集合中返回满足用户需求的信息
  - 例1：返回与信息检索相关的网页→搜索引擎(Search Engine, SE)
  - 例2：毛主席的生日是哪天？→问答系统(Question Answering, QA)
  - 例3：返回联想PC的型号、配置、价格等信息→信息抽取(Information Extraction, IE)
  - 例4：订阅有关NBA的新闻→信息过滤(Information Filtering)、信息推荐(Information Recommending)
- 狭义的IR通常是指Information Search，广义的IR包含非常多的内容(SE, QA, IE, …)

信息检索和数据库检索

	信息检索	数据库检索
检索对象	无结构、半结构数据 如网页、图片……	结构化数据 如：员工数据库
检索方式	通常是近似检索 如：每个结果有相关度得分	通常是精确检索 如：姓名==“李明”
检索语言	主要是自然语言 如：查与超女相关的新闻	SQL结构化语言

近年来，两种检索已经逐渐融合，边界越来越不明显。

信息检索的基本概念

- 用户需求(Information Need，IN)
  - 严格地说，IN存在于用户的内心，但是通常用文字来描述，如 查找与2006世界杯相关的新闻，通常也称为主题(Topic)
  - IN提交给检索系统时称为查询(Query)，如 2006 世界杯，一个IN可以对应多个Query
- 文档(Document)
  - 可以是文本、图像、视频、语音文件等
- 文档集合(Collection)
  - 所有待检索的文档构成的集合

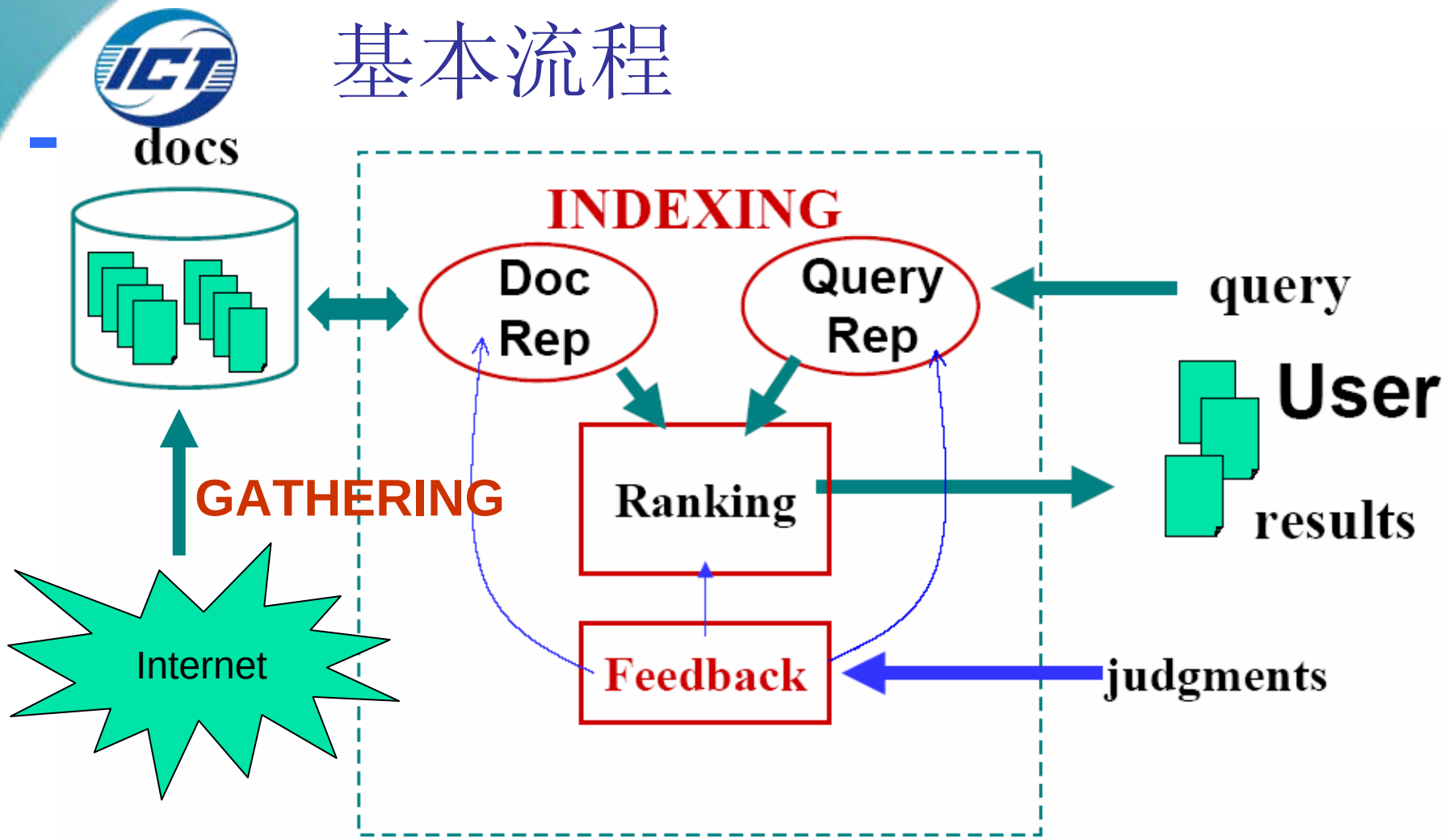
相关度(Relevance)

- 相关度目前也没有统一的定义，简单地认为是查询和文档的匹配相似度得分
- 形式上说，相关度是一个函数R，输入是查询Q、文档D和文档集合C，返回的是一个实数值
  - R=f(Q,D,C)
- 信息检索就是给定一个查询Q，从文档集合C中计算每篇文档D与Q的相关度并排序(Ranking)。
- 相关度通常只有相对意义，对一个Q，不同文档的相关度可以比较，而对于不同的Q的相关度不便比较
- 相关度的输入信息可以更多，比如用户的背景信息、用户的查询历史等等
- 现代信息检索中相关度不是唯一度量，如还有：重要度、权威度、新颖度等度量。
  - Google中据说用了上百种排名因子

内容

- 信息检索的基本概念
- 信息检索的基本流程←
- 信息检索的评价方法
- 信息采集
- 信息分析及索引
- 信息检索模型及其他相似度计算方法
- 查询扩展及相关反馈

基本流程


docs
INDEXING
Doc Rep
Query Rep
query
User
GATHERING
Ranking
results
Internet
Feedback
judgments

流程介绍

- 信息采集(Information Gathering)
  - 获取信息，通常是指从Internet上获取信息
- 信息标引(Information Indexing)
  - 将查询和文档表示成方便检索的某种方式
  - 信息标引通常也包括信息分析的组织
- 相似度计算和排序(Ranking)
- 相关反馈(Feedback)
  - 根据用户的交互重新构造查询进行检索
  - 可以不通过用户自动进行，称为伪相关反馈，比如假定前5篇文档相关

内容

- 信息检索的基本概念
- 信息检索的基本流程
- 信息检索的评价方法←
- 信息采集
- 信息分析及索引
- 信息检索模型及其他相似度计算方法
- 查询扩展及相关反馈

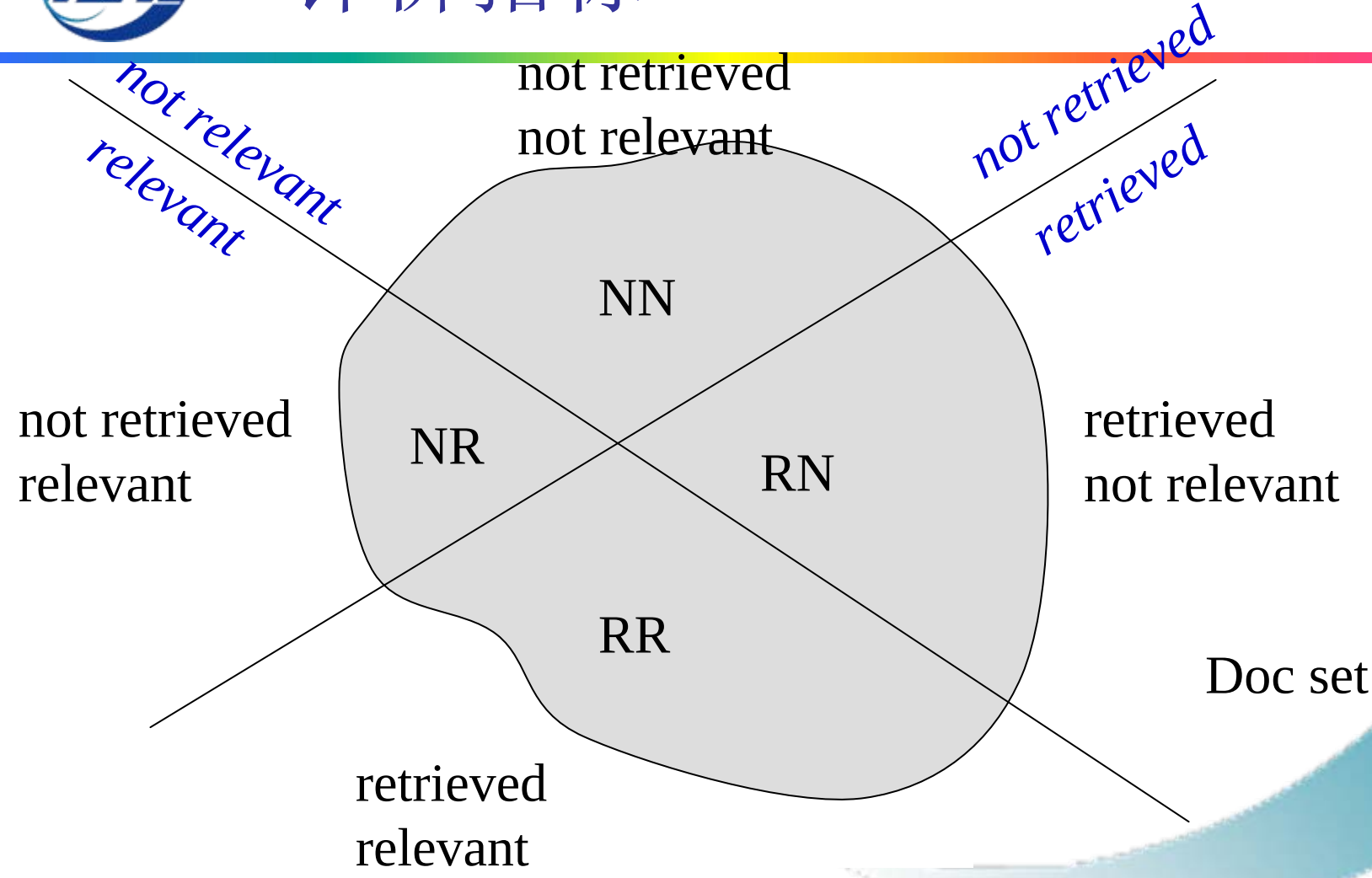
评价什么？

- 效率 (Efficiency)—可以采用通常的评价方法
  - 标引和检索算法的速度
  - 索引的更新能力
  - 存储开销
  - 并行或分布计算能力
- 效果 (Effectiveness)
  - 找到多少相关文档
  - 遗漏了多少相关文档
  - 找到的结果中非相关文档有多少

如何评价(效果)?

- 相同的文档集合，相同的主题集合，相同的评价指标，不同的检索系统进行比较。
  - **The Cranfield Experiments**, Cyril W. Cleverdon, Cranfield College of Aeronautics, 1957 –1968 (hundreds of docs)
  - **SMART System,** Gerald Salton, Cornell University, 1964-1988 (thousands of docs,)
  - **TREC(Text REtrieval Conference)**, Donna Harman, National Institute of Standards and Technology (NIST), 1992 - (millions of docs, 100k to 7.5M per set, training Q's and test Q's, 150 each)

评价指标


not relevant
relevant
not retrieved
not relevant
not retrieved
retrieved
NN
not retrieved
relevant
NR
RN
retrieved
not relevant
RR
Doc set
retrieved
relevant

评价指标

- 召回率(Recall): RR/(RR + NR)，返回的相关结果数占实际相关结果总数的比率，也称为查全率
- 正确率(Precision): RR/(RR + RN)，返回的结果中真正相关结果的比率，也称为查准率
- 一个例子：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, Precision=80/200
- 两个指标分别度量了系统的某个方面

评价指标

- 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5个相关结果，某系统排序返回的相关文档的位置分别是第1，第2，第5，第10，第20位，则AP=(1/1+2/2+3/5+4/10+5/20)/5
  - 插值的AP:在召回率分别为0,0.1,0.2,…,1.0的十个点上的正确率求平均
- MAP(Mean AP)：对所有查询求平均
- Precision@N：在第N个位置上的正确率，P@10, P@20对大规模搜索引擎非常有效

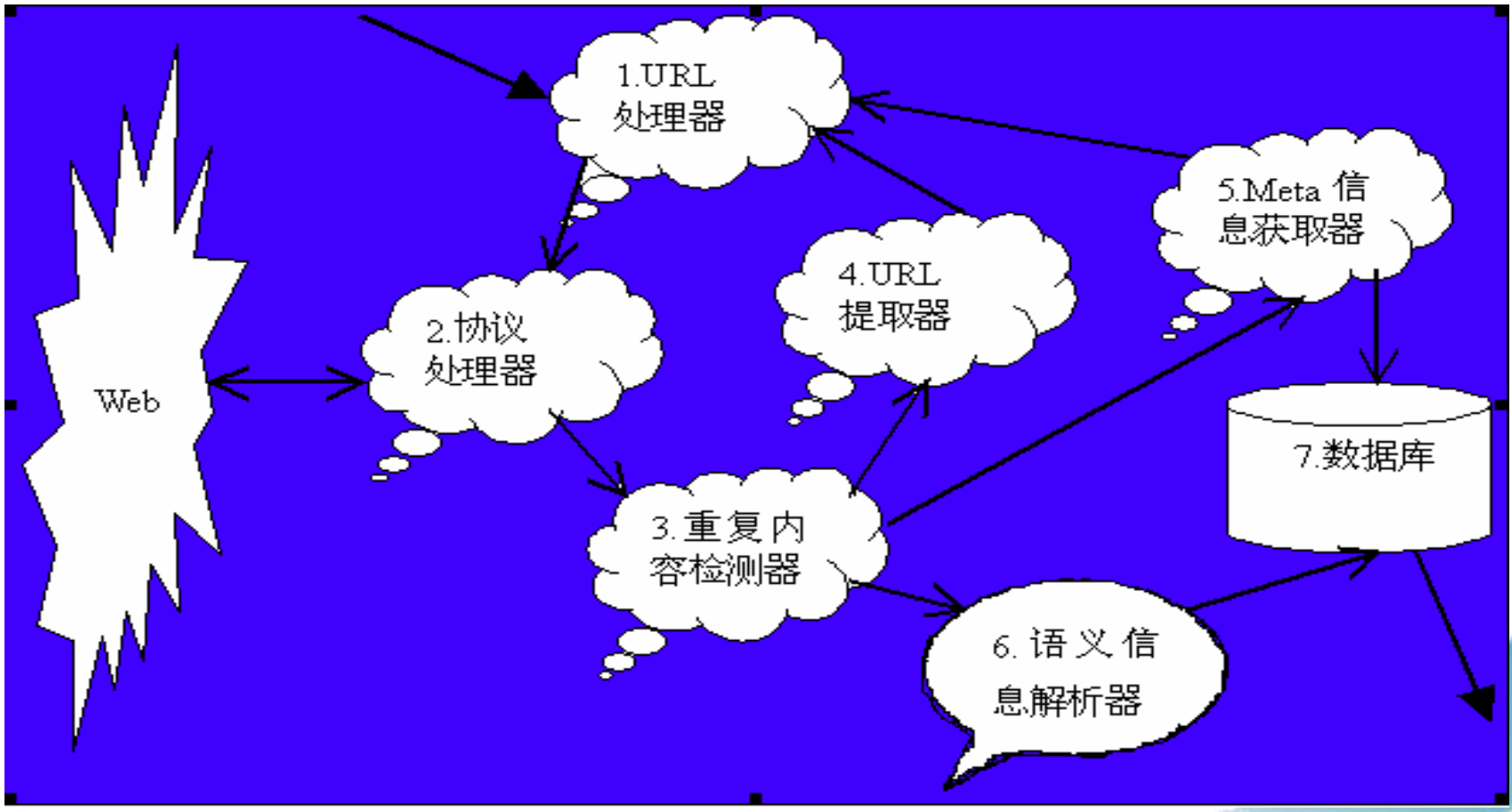
内容

- 信息检索的基本概念
- 信息检索的基本流程
- 信息检索的评价方法
- 信息采集←
- 信息分析及索引
- 信息检索模型及其他相似度计算方法
- 查询扩展及相关反馈

信息采集的概念

- 主要是指通过Web页面之间的链接关系从Web上自动获取页面信息,并且随着链接不断向所需要的Web页面扩展的过程
- 实际上是图的遍历过程
  - 通过种子页面或站点(Seed)，获取更多的链接，将它们作为下一步种子，循环
  - 这个过程永远不会结束！

信息采集的基本结构


1.URL
处理器
2.协议
处理器
Web
3.重复内
容检测器
4.URL
提取器
5.Meta 信
息获取器
6.语义信
息解析器
7.数据库

研究的问题

- 遍历算法：
  - 深度优先 vs 广度优先
  - 剪枝方法
- 页面更新
  - 更新周期
- 快速重复检测
  - 重复URL
  - 重复页面
- 动态页面采集
- 工程实现：并行、吞吐率

内容

- 信息检索的基本概念
- 信息检索的基本流程
- 信息检索的评价方法
- 信息采集
- 信息分析及索引←
- 信息检索模型及其他相似度计算方法
- 查询扩展及相关反馈

信息分析

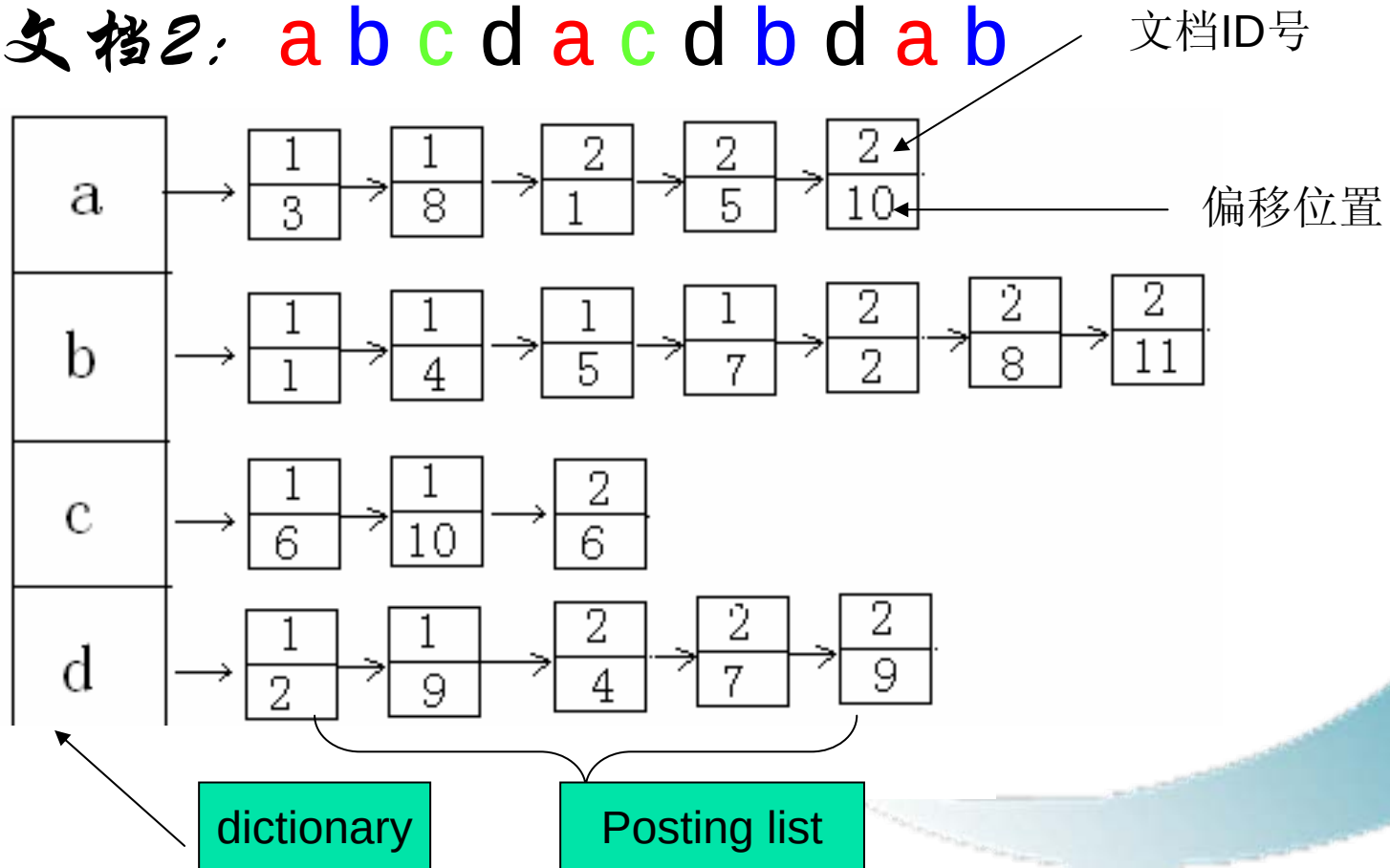
- 信息分析是对原始数据的预处理
  - 格式分析与转换(html/xml/doc/pdf/rtf)
  - 语种识别、编码识别与转换(GB/BIG5/Unicode)
  - 噪声数据的清洗
  - 冗余数据的处理
  - 信息编号

信息索引(1)

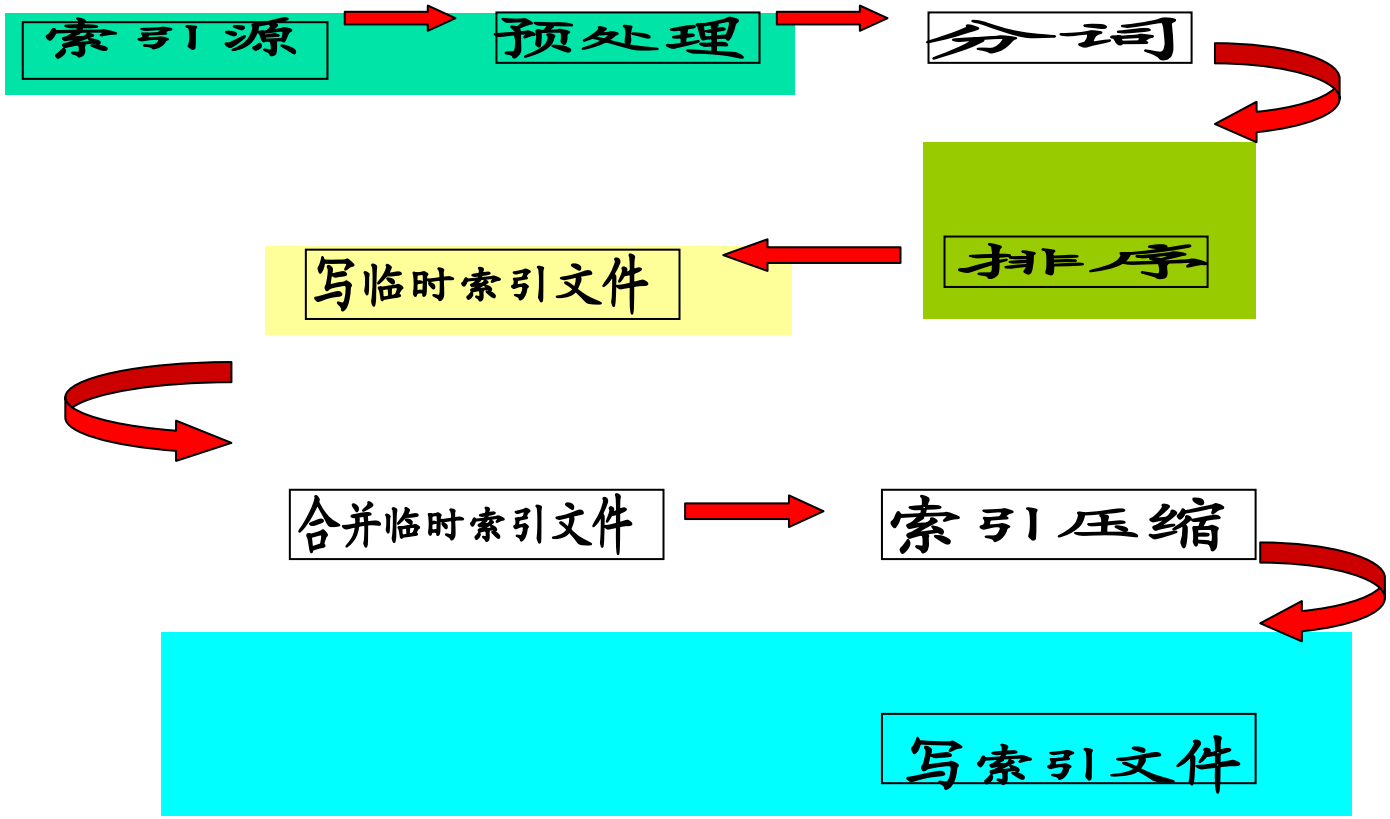
- 为加快搜索速度，建立特定的数据结构
  - 不可能是逐个文档扫描(太慢)
  - 倒排表、后缀树、签名表等等
- 大规模海量数据的索引常常用倒排表结构
  - Inverted file
  - 所有的搜索引擎都用倒排表
  - 速度最快

倒排索引示例

- 文档1： b d a b b c b a d c
- 文档2： a b c d a c d b d a b



建立倒排索引的大致框架



分词

- 对于中文，分词的作用实际上是要找出一个个的索引单位
- 例子：李明天天都准时上班
- 索引单位
  - 字：李 明 天 天 都 准 时 上 班
    - 索引量太大，查全率百分百，但是查准率低，比如查“明天” 这句话也会出来
  - 词：李明 天天 都 准时 上班
    - 索引量大大降低，查准率较高，查全率不是百分百，而且还会受分词错误的影响，比如上面可能会切分成：李 明天 天都 准时 上班

英文词根还原

- 进行词根还原：
  stop/stops/stopping/stopped→stop
  - 好处：减少词典量；坏处：按词形查不到，词根还原还可能出现错误
- 不进行词根还原：
  - Stopped→sto+ppe+d
  - 好处：支持词形查询；坏处：增加词典量

停用词消除

- 停用词(stop words)是指那些出现频率高但是无重要意义，通常不会作为查询词出现的词，如“的”、“地”、“得”、“都”、“是”等等
  - 消除：通常是通过查表的方式去除，去除的好处----大大较少索引量，坏处----有些平时的停用词在某些上下文可能有意义

排序

- 排序就是对分词产生的（文档号，单词号，出现位置）三元组按照单词号排序，单词号相同的项再按照文档号排序，单词号和文档号都相同的再按照出现位置排序。
- 排序是为了方便写入临时索引文件。

排序举例（1）

- 文档1
  - b d a b b c b a d c
- 文档2
  - a b c d a c d b d a b

排序举例（2）

- 对文档1分析产生的三元组如下：
- b d a b b c b a d c
- （文档号，单词号，位置）

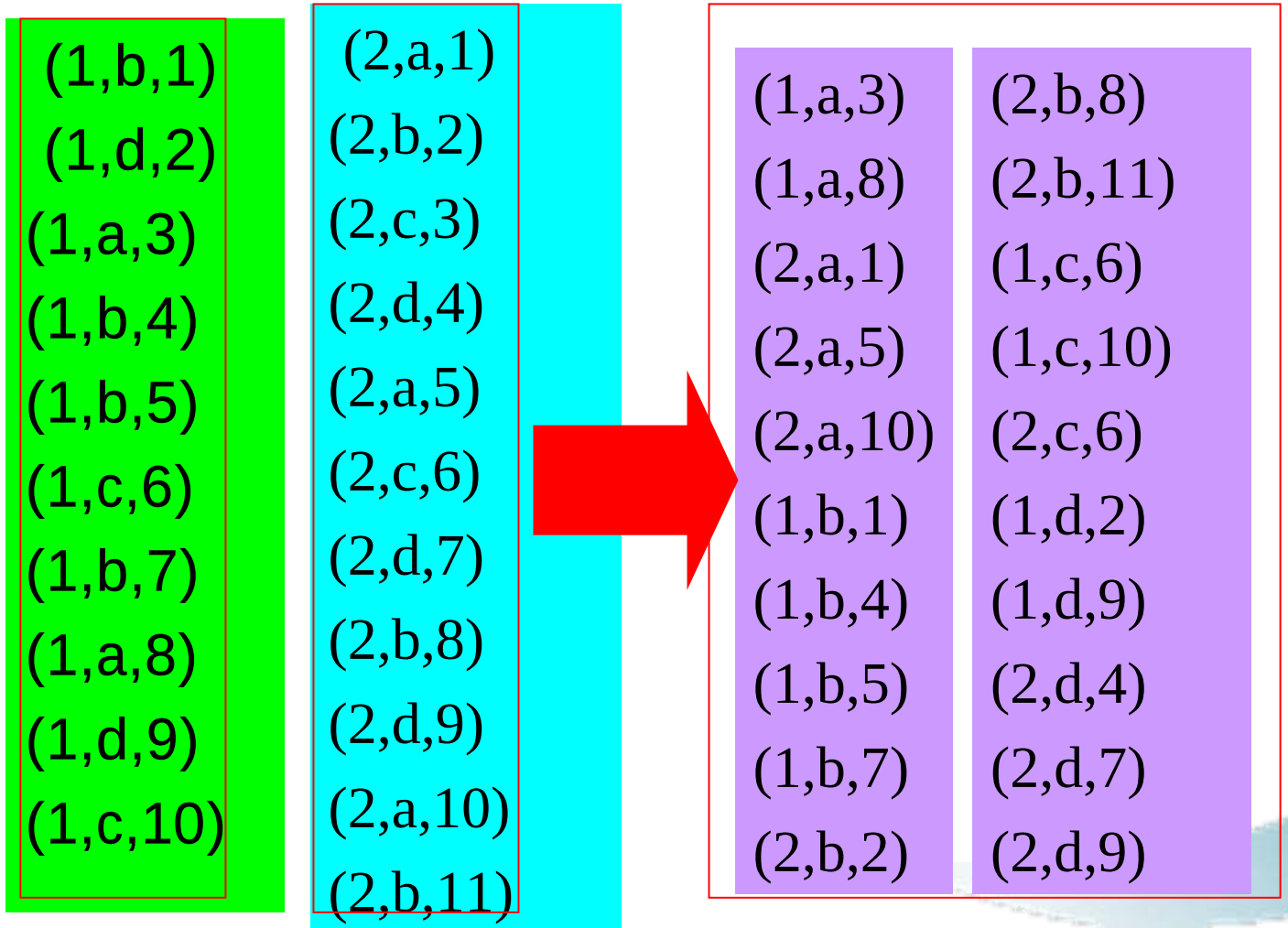
(1,b,1) (1,d,2) (1,a,3)
(1,b,4) (1,b,5) (1,c,6)
(1,b,7) (1,a,8) (1,d,9)
(1,c,10)

排序举例（3）

- 对文档2分析产生的三元组如下：
  - a b c d a c d b d a b
- (文档号，单词号，位置）

(2,a,1) (2,b,2) (2,c,3)
(2,d,4) (2,a,5) (2,c,6)
(2,d,7) (2,b,8) (2,d,9)
(2,a,10) (2,b,11)

排序举例（4）


(1,b,1)
(1,d,2)
(1,a,3)
(1,b,4)
(1,b,5)
(1,c,6)
(1,b,7)
(1,a,8)
(1,d,9)
(1,c,10)
(2,a,1)
(2,b,2)
(2,c,3)
(2,d,4)
(2,a,5)
(2,c,6)
(2,d,7)
(2,b,8)
(2,d,9)
(2,a,10)
(2,b,11)
(1,a,3)
(1,a,8)
(2,a,1)
(2,a,5)
(2,a,10)
(1,b,1)
(1,b,4)
(1,b,5)
(1,b,7)
(2,b,2)
(2,b,8)
(2,b,11)
(1,c,6)
(1,c,10)
(2,c,6)
(1,d,2)
(1,d,9)
(2,d,4)
(2,d,7)
(2,d,9)

写入临时索引文件

(1,a,3)	(2,b,8)
(1,a,8)	(2,b,11)
(2,a,1)	(1,c,6)
(2,a,5)	(1,c,10)
(2,a,10)	(2,c,6)
(1,b,1)	(1,d,2)
(1,b,4)	(1,d,9)
(1,b,5)	(2,d,4)
(1,b,7)	(2,d,7)
(2,b,2)	(2,d,9)

合并多个临时索引文件

- 一批产生一个索引文件
- 多批之间通过合并产生一个大的临时文件
- 如果不合并，可以建立通过文档进行分割的分布式索引

文档1，2，…，m

文档m+1，m+2，…，2m

- 也可以合并以后，按照词典进行分割

词1，2，…，n

词n+1，n+2，…，2n

索引压缩

- 需不需要压缩？
  - 要压缩：索引大小通常和原始文本大小相当，不压缩可能会耗费大量存储开销
  - 不压缩：硬盘很便宜，数据量不是特别大，而且不需要解压缩的消耗

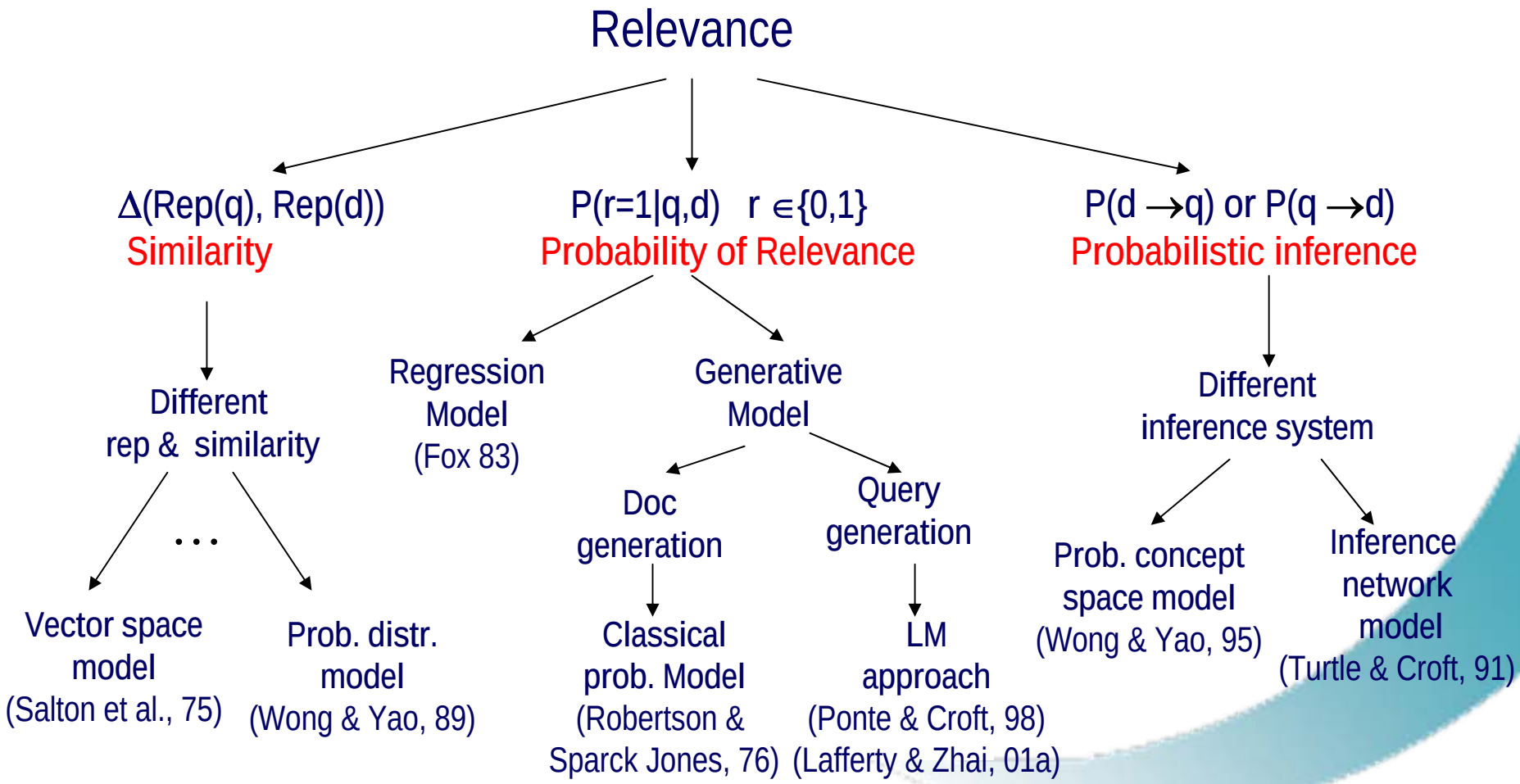
内容

- 信息检索的基本概念
- 信息检索的基本流程
- 信息检索的评价方法
- 信息采集
- 信息分析及索引
- 信息检索模型及其他相似度计算方法←
- 查询扩展及相关反馈

基本分类

- 信息检索模型是指如何对查询和文档进行表示，然后对它们进行相似度计算的框架和方法。本质上是对相关度建模。
- 基本分类：
  - 布尔模型：Boolean，可以看成VSM的一种特例
  - 向量空间模型：Vector Space Model
  - 概率IR模型：Probabilistic Model
  - 基于统计语言建模的IR模型：Statistics Language Model IR，可看成概率模型的一种

Relevance 概念(引自 Chengxiang Zhai IR教程)



布尔模型

- 布尔模型：
  - 查询为布尔表达式： 2006 and 世界杯
  - 匹配过程就是布尔表达式匹配，匹配上得分为1，否则为0
  - 类似于传统数据库检索，是精确匹配
  - 优点：简单
  - 缺点：不能近似匹配，一般用户构造查询不是很容易，造成结果过多或者过少
  - 现代很多搜索引擎中仍然使用布尔模型的思想

向量空间模型

- 查询和文档都转化成标引项(Term)及其权重组成的向量表示
  - 康奈尔大学 Salton 1970年代提出并倡导，原型系统SMART
  - 查询：(<2006,1>,<世界杯,2>)
  - 文档：文档A (<2006,1>,<世界杯,3>,<德国,1>,<举行,1>)，文档B (<2002,1><世界杯,2>,<韩国,1>,<日本,1>)
  - 查询和文档进行向量的相似度计算：夹角余弦或者内积
  - 优点：简洁直观
  - 缺点：标引项之间的独立性假设与实际不符

向量空间模型

- 标引项(Term)的选择：词、短语、N-gram或者某种语义单元
- 权重计算：
  - Term的频率: TF，TF越高权重越高
  - Term的文档频率：DF越高区分度越低，因此权重也越低，逆文档频率(Inverse DF)
  - 文档的长度：长度归一化
  - 常用的公式(Pivoted Normalization)

$$\sum_{w \in q \cap d} \frac{1 + ln(1 + ln(c(w,d)))}{(1-s) + s\frac{|d|}{avdl}} \cdot c(w,q) \cdot ln\frac{N+1}{df(w)}$$

概率检索模型

- 概率检索模型是通过概率的方法将查询和文档联系起来，P(R=1|Q,D)
- 概率模型包括一系列模型
  - 回归模型(Regression Model)：假设文档和查询的相关概率是多个特征的加权组合函数($P=a_1*f_1+a_2*f_2+..+a_n*f_n$)，然后通过对训练集合进行回归的方式计算每个特征$f_i$的权重$a_i$
  - 二元独立概率模型(Binary Independence Relevance)：伦敦城市大学Robertson及剑桥大学Sparck Jones 1970年代提出，代表系统OKAPI
  - BIR：由每个Term在相关文档和不相关文档中的概率P(t|相关)、P(t|不相关)求得文档和查询相关的概率或某个比值，如：R=P(D|相关)/P(D|不相关)，P(D,x)可以转化为所有Term的出现概率的乘积
  - BIR模型举例：对于查询 2006 世界杯，我们可以计算出现相关文档出现 世界杯的概率是 0.2, 出现2006的概率是0.1，出现在不相关文档的概率分别是0.02和0.05。则可以计算某篇文档的R值。
  - BIR模型更流行
- 优点(BIR)：理论上具有可解释性
- 缺点(BIR)：计算时通常引入标引项独立性假设，很多参数需要估计

概率检索模型

- 由于对于每个查询，无法事先得到相关文档集和不相关文档集，所以必须进行估计
- 有多种估计方法
- 最流行的OKAPI BM25公式：

$$\sum_{w \in q \cap d} \left(ln \frac{N - df(w) + 0.5}{df(w) + 0.5} \times \frac{(k_1 + 1) \times c(w,d)}{k_1((1-b) + b\frac{|d|}{avdl}) + c(w,d)} \times \frac{(k_3 + 1) \times c(w,q)}{k_3 + c(w,q)} \right)$$

统计语言建模IR模型

- 麻省大学Bruce Croft于1998年提出，包括一系列模型，代表系统Lemur
  - 查询似然模型：把相关度看成是每篇文档对应的语言下生成该查询的可能性
    - 类比：作者A和作者B写的文章用词风格很不相同，可以统计用词的概率(语言)，然后对应一篇新的文章，判断是A写的还是B写的
  - 翻译模型：假设查询经过某个噪声信道变形成某篇文章，则由文档还原成该查询的概率(翻译模型)可以视为相关度
  - KL距离模型：查询对应某种语言，每篇文档对应某种语言，查询语言和文档语言的KL距离作为相关度度量
- 优点：理论上具有解释性，不依赖于独立性假设
- 缺点：数据稀疏性，需要参数估计

其他相似度计算方法

- Google 的PageRank：与查询无关
  - 受启发于 文献引用，越多越重要的文献引用的文献越重要
  - WEB上的链接关系看成引用
- IBM的HITS算法：与查询相关
  - 每篇文章具有两个值：authority 和 hub
  - 通过递归计算文章的这两个值
- 不是传统的相关度概念
- 一般不单独使用，和内容相关度融合使用

内容

- 信息检索的基本概念
- 信息检索的基本流程
- 信息检索的评价方法
- 信息采集
- 信息分析及索引
- 信息检索模型及其他相似度计算方法
- 查询扩展及相关反馈←

查询扩展

- 对用户的查询进行扩充：比如用户输入 计算机，我们扩充 一个词 电脑
- 同义词扩展：
  - 同义词词典
  - 通过统计构造的同义词词典
- 相关词扩展：
  - 相关词：“2006世界杯” 与“德国”
  - 基于全局分析的查询扩展：对文档集合进行分析得到某种相关词典
- 查询重构：对用户的初始查询进行修改(可以是加词、减词，或者对于向量模型表示的初始查询进行权重的修改等等)，是比查询扩展更泛的一个概念

相关反馈

- 指根据用户对初始检索结果的干预来重新生成查询或者修改模型参数等等
  - 伪相关反馈：系统假定一些相关的结果，并根据这些结果来进行返回
  - 相关反馈是一种手段，目的可以是查询扩展或者重构，也可以是模型的调整
  - 基于伪相关反馈和局部分析进行查询重构：根据某些文档中的信息来对查询进行重构

小结

- 信息检索的基本概念
- 信息检索的基本流程
- 信息检索的评价方法
- 信息采集
- 信息分析及索引
- 信息检索模型及其他相似度计算方法
- 查询扩展及相关反馈

参考教材

- Baeza-Yates, R. & B. Ribeiro-Neto. eds. Modern Information Retrieval. ACM Press, 1999
- Witten, Ian et al. Managing Gigabytes. Orlando, FL: Morgan Kaufmann Publishers Incorporated, 1999

谢谢！

wangbin@ict.ac.cn