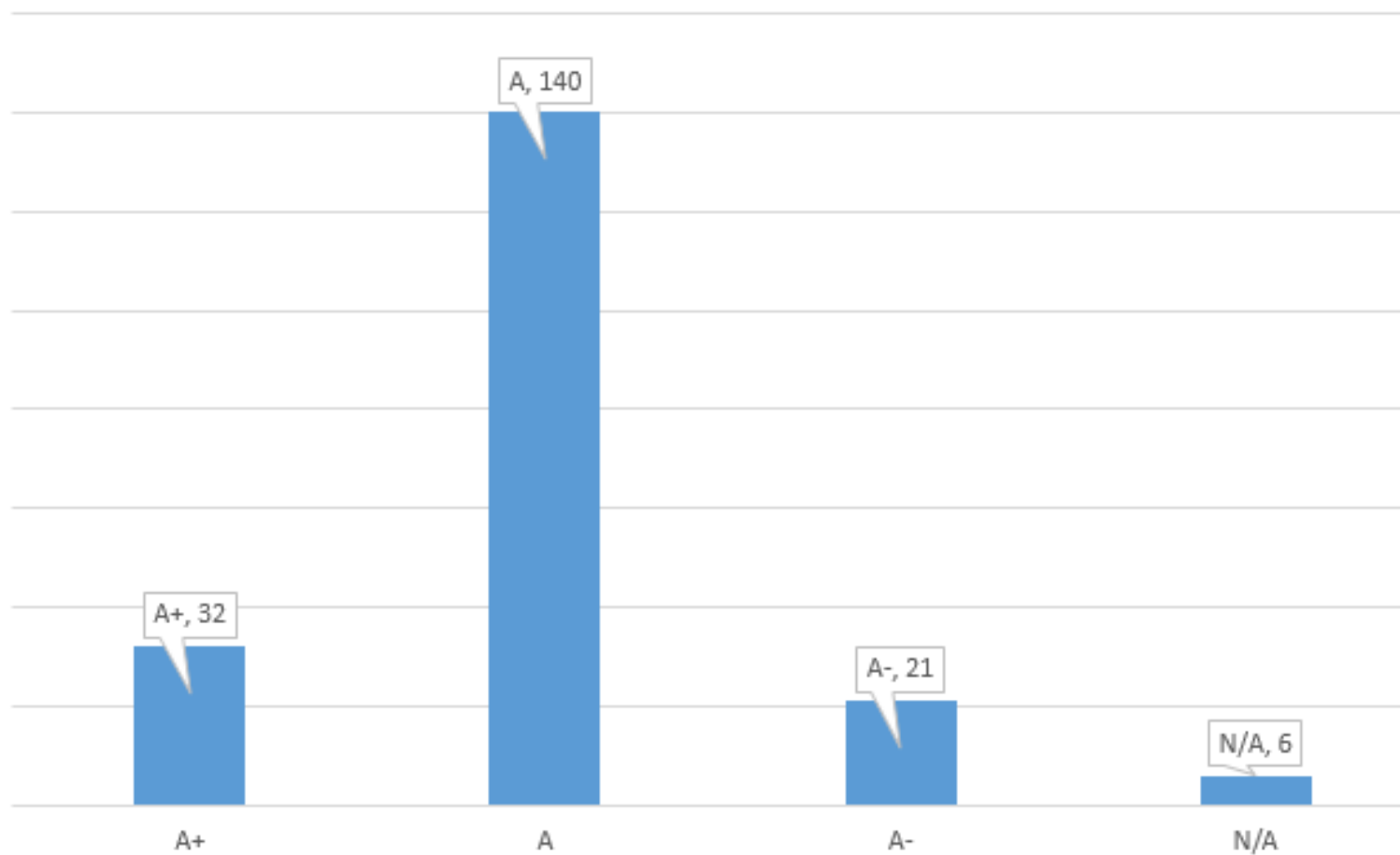


# 互联网搜索与排序

## Web Search and Ranking

徐 君

# 第一次课堂作业总结



# 提纲

- 互联网搜索介绍
- 传统相关性排序模型
- 搜索结果多样化排序
- 总结

# 信息检索

- 信息检索(Information Retrieval, IR)是指从大规模的非结构化数据集中(通常指文本文档)寻找满足用户信息需求的过程
- 互联网搜索引擎是目前最常见的信息检索系统，但信息检索不局限于互联网搜索：
  - 企业搜索(如SharePoint Search)
  - 特定领域文档搜索(Scholar, Patent等)
  - 桌面搜索、Email搜索

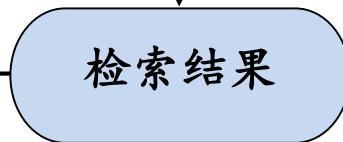
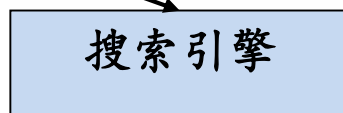
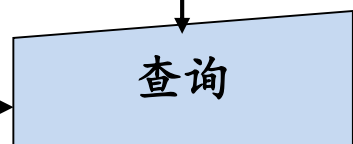
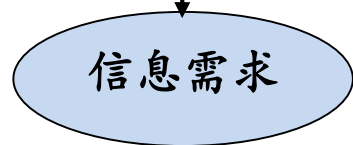
# 关于几个关键词

- **寻找信息**：与构造新的信息内容(如统计归纳)不同，信息检索只负责提供已有的信息给用户
- **非结构化数据**：与数据库中关系数据不同，非结构化数据不容易被计算机处理
- **信息需求**：通常通过查询词进行表达
- **大规模数据**：例如互联网网页、企业内部网数据等，数据量大，处理数据的方法需要足够高效且可扩展

# 对信息检索系统的基本假设

- 静态文档集合
  - 假设在用户搜索的时刻，文档集合不发生变化
- 检索目的
  - 从文档集合中检索出与用户的信息需求**相关**的文档，从而帮助用户完成某一特定**任务**

# 搜索概念模型



Get rid of mice in a politically correct way

Info about removing mice without killing them

how trap mice alive

Search



# 互联网搜索引擎发展

Archie FAQ  
(1990)  
精确FTP文件名  
搜索



提供简单目录搜索



(1995)  
支持自然语言搜索  
和高级搜索语法



(1998)

World Wide Web  
Wanderer  
(1993)  
第一个网络爬虫程序



(1994)  
全文搜索引擎



(1995)  
Inktomi公司, 抓取索引1千  
万页/天, 储存用户搜索喜好



(1999)  
Fast公司, 利用ODP自动  
分类改善搜索



(1993)  
网站主动提交检索信息



(1994)  
网页自动摘要



(1996)  
自然语言提问, 优先  
提供答案



(2000)  
搜索结果自动聚类



(1993)  
分析字词关系  
概念搜索



infoseek  
(1994)  
网页自动摘要,  
同时提供网页目  
录等其他服务



(1997)  
第一个中文搜索引擎



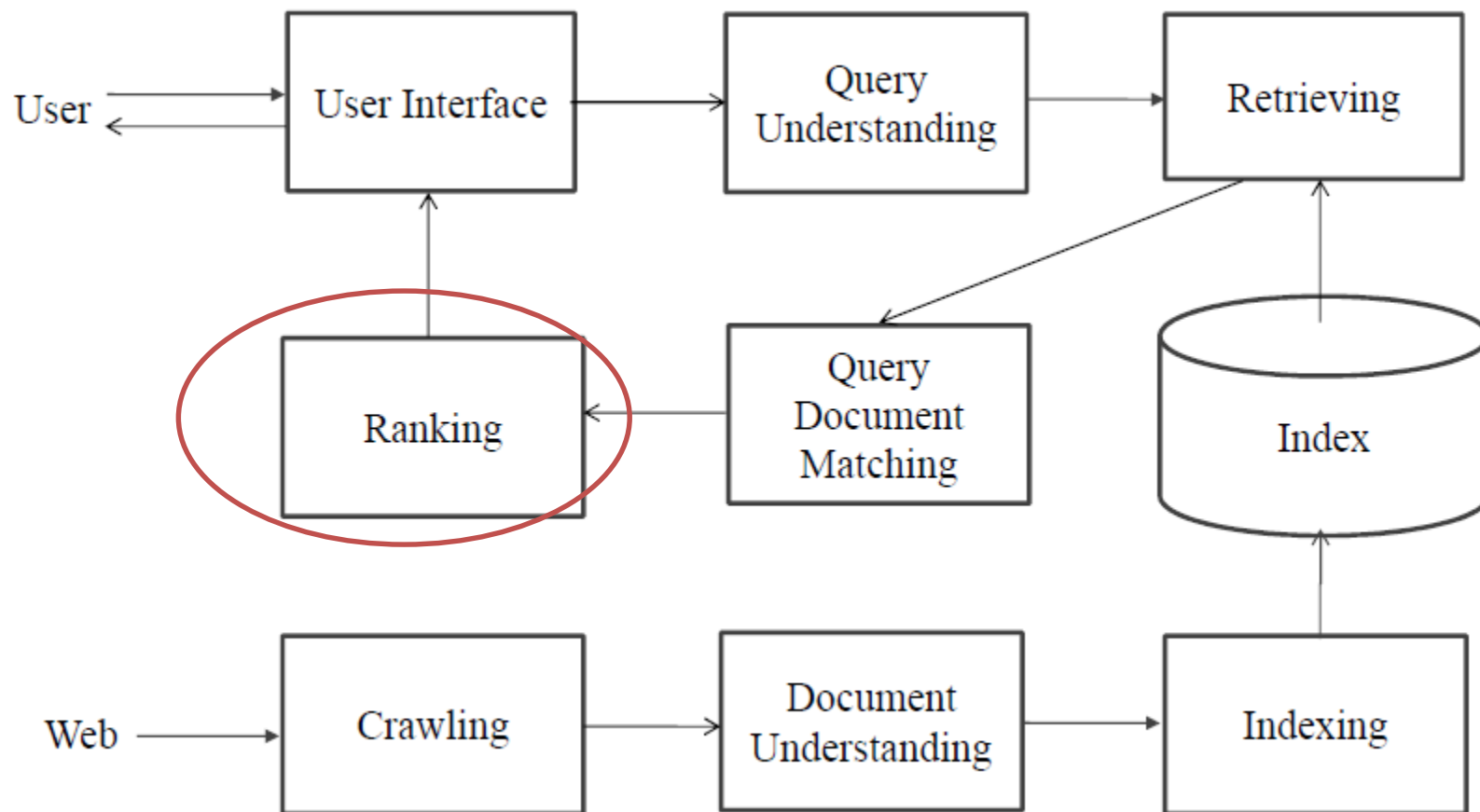
(2000)  
目前为止最成功的中  
文搜索引擎



# 互联网搜索引擎发展

- 第一代（1994—1998）
  - 基于语法的查询-内容匹配 (syntactic matching)
- 第二代（1998—约2008）
  - 不仅仅考虑网页内容与查询的匹配(beyond “on-page” content)
  - 同时考虑链接分析、用户点击路径等
- 第三代（2008—约2015）
  - 结果页面不仅仅显示网页链接（Beyond 10 blue links）
  - User intension, short cut, rich content
- 第四代
  - 移动搜索？
  - 个性化搜索？

# 搜索引擎主要模块



# 排序：搜索结果的展示手段

- 传统展示：显示所有结果集合
  - 文档太多：难以浏览
  - 文档太少：找不到满意结果
- 排序
  - 按照相关度从上往下排序
  - 辅助展示手段：（动态）摘要与飘红



# 排序的准则

- 在不同的搜索应用中有不同的排序准则
- 明确的排序准则
  - 时间（如学术搜索、Email搜索、新闻搜索）
  - 引用量(学术搜索)
  - 评论数、成交量、下载量 (商品搜索、apps搜索)
  - .....
- 模糊的排序准则
  - 相关度
  - 重要性

# 按照时间排序

Title	+ Add	More	1-20	Cited by	Year
Modeling Parameter Interactions in Ranking SVM Y Zhang, J Xu, Y Lan, J Guo, M Xie, Y Huang, X Cheng Proceedings of the 24th ACM International on Conference on Information and ...					2015
A Probabilistic Model for Bursty Topic Discovery in Microblogs X Yan, J Guo, Y Lan, J Xu, X Cheng Twenty-Ninth AAAI Conference on Artificial Intelligence				2	2015
Post Processing of Ranking in Search J Xu					2015
Next Basket Recommendation with Neural Networks S Wan, Y Lan, P Wang, J Guo, J Xu, X Cheng					2015
Learning Word Representations by Jointly Modeling Syntagmatic and Paradigmatic Relations F Sun, J Guo, Y Lan, J Xu, X Cheng the 53rd Annual Meeting of the Association for Computational Linguistics and ...				4 *	2015
Learning Maximal Marginal Relevance Model via Directly Optimizing Diversity Evaluation Measures L Xia, J Xu, Y Lan, J Guo, X Cheng The 38th annual international ACM SIGIR conference on Research and ...				1 *	2015
Learning Hierarchical Representation Model for Next Basket Recommendation P Wang, J Guo, Y Lan, J Xu, S Wan, X Cheng 38th annual international ACM SIGIR conference on Research and development ...					2015
Semantic Matching in Search J Xu					2014
Query expansion for web search J Xu, H Li US Patent 8,898,156					2014

# 按照引用量排序

<input type="checkbox"/> Title	<input type="checkbox"/> + Add	<input type="checkbox"/> More	1-20	Cited by	Year
<input type="checkbox"/> Adarank: a boosting algorithm for information retrieval				456	2007
J Xu, H Li Proceedings of the 30th annual international ACM SIGIR conference on ...					
<input type="checkbox"/> Letor: Benchmark dataset for research on learning to rank for information retrieval				364	2007
TY Liu, J Xu, T Qin, W Xiong, H Li Proceedings of SIGIR 2007 workshop on learning to rank for information ...					
<input type="checkbox"/> Adapting ranking SVM to document retrieval				347	2006
Y Cao, J Xu, TY Liu, H Li, Y Huang, HW Hon Proceedings of the 29th annual international ACM SIGIR conference on ...					
<input type="checkbox"/> LETOR: A benchmark collection for research on learning to rank for information retrieval				168	2010
T Qin, TY Liu, J Xu, H Li Information Retrieval 13 (4), 346-374					
<input type="checkbox"/> Directly optimizing evaluation measures in learning to rank				88	2008
J Xu, TY Liu, M Lu, H Li, WY Ma Proceedings of the 31st annual international ACM SIGIR conference on ...					
<input type="checkbox"/> Ranking definitions with supervised learning methods				57	2005
J Xu, Y Cao, H Li, M Zhao Special interest tracks and posters of the 14th international conference on ...					
<input type="checkbox"/> Regularized latent semantic indexing				52	2011
Q Wang, J Xu, H Li, N Craswell SIGIR'11, 685-694					
<input type="checkbox"/> Using SVM to extract acronyms from text				25	2007
J Xu, Y Huang Soft Computing 11 (4), 369-373					

# 按照其它特定统计量排序

Baidu 音乐

新闻 网页 贴吧 知道 音乐 图片 视频 地图 百科 文库

许巍

百度一下

经典老歌 南山南 儿歌 周杰伦 小苹果 音乐 TFBOYS 大 一次就好

首页 歌手 分类 榜单 MV 歌单

热门歌曲

搜索“许巍”，找到相关歌曲共209首。

歌曲(209) 歌手(1) 专辑(8) 歌词(112) 酷我(176)

 **许巍**  
单曲: 138首 专辑: 6张  
[播放 Ta 的热门歌曲](#) [收听 Ta 的电台](#)

☐ 全部 [播放选中歌曲](#) [+ 加入播放列表](#)

<input type="checkbox"/>	01	 第三极 中央电视台纪录片《第三极》主题曲 影视原声	许巍	38,889人听过	<a href="#">▶</a> <a href="#">+</a> <a href="#">≡</a> <a href="#">□</a> <a href="#">♥</a>
<input type="checkbox"/>	02	 Opening+空谷幽兰 现场	许巍		<a href="#">▶</a> <a href="#">+</a> <a href="#">≡</a> <a href="#">□</a> <a href="#">♥</a>
<input type="checkbox"/>	03	 救赎之旅+鼓Solo 现场	许巍		<a href="#">▶</a> <a href="#">+</a> <a href="#">≡</a> <a href="#">□</a> <a href="#">♥</a>
<input type="checkbox"/>	04	 故乡	许巍		<a href="#">▶</a> <a href="#">+</a> <a href="#">≡</a> <a href="#">□</a> <a href="#">♥</a>
<input type="checkbox"/>	05	 救赎之旅+鼓Solo 现场	许巍		<a href="#">▶</a> <a href="#">+</a> <a href="#">≡</a> <a href="#">□</a> <a href="#">♥</a>
<input type="checkbox"/>	06	 青鸟I	许巍		<a href="#">▶</a> <a href="#">+</a> <a href="#">≡</a> <a href="#">□</a> <a href="#">♥</a>
<input type="checkbox"/>	07	 开场 天鹅之旅 现场	许巍		<a href="#">▶</a> <a href="#">+</a> <a href="#">≡</a> <a href="#">□</a> <a href="#">♥</a>

# 网页搜索排序准则

- 相关性排序
  - 模糊的排序准则，如何精确定义？
- 研究者们试图从查询和文档性质，以及它们中的词的共现关系，计算查询-文档的相关度
  - 共现次数(次数越多越相关)
  - 词的重要性
  - 文档长度
  - 文档重要性(微软主页和苹果主页谁更重要？)



# 提纲

- 互联网搜索介绍
- 传统相关性排序模型
- 搜索结果多样化排序
- 总结

# 文档处理

- 将自然语言的文本处理为计算机容易处理的格式，如词-文档矩阵
- 单词可能出现错误拼写，具有多种形式
  - 单复数: car, cars; foot, feet; mouse, mice
  - 时态: go, went; say, said
  - 形容词副词: active, actively; rapid, rapidly
- 不同的语言有不同的处理方式
  - 如: 中文需要分词，英文分词比较简单
- 文档处理是IR的第一步，直接影响搜索结果

# 何为文档

- 格式上，如果文档不是纯文本格式，则需要转换为纯文本
  - PDF、HTML、Word → Text
  - 图像、格式信息将丢失
- 文档为可检索的基本单元，如何定义？
  - 《网络数据挖掘》PDF包含了很多章节，是分为多个文档还是看成一个文档？
  - 互联网上的课件，每一页幻灯片都被做成了一个单独的网页，是否需要组合成一个文档？
  - 没有一致的答案。

# 分词

- 将句子分解成词序列

*Two households, both alike in dignity, in fair Verona, where*

Two	households	both	alike	in	dignity	in	fair	Verona	where
-----	------------	------	-------	----	---------	----	------	--------	-------

- 最基本的方式(英文)
  - 去除标点符号
  - 按照分隔符(空格)分开
- 中文
  - 成为自然语言处理中一个重要的课题
  - 目前精度已经达到可用的程度

# 分词中的一些问题

- 语言中总会出现一些特殊情况
  - boys' → boys vs. can't → can t
  - <http://www.bigdatalab.ac.cn> and [junxu@ict.ac.cn](mailto:junxu@ict.ac.cn)
  - *co-ordinates* vs. *good-looking man*
  - straight forward, white space, Los Angeles, hot dog
  - Compounds:  
Lebensversicherungsgesellschaftsangestellter

morphemes can be used as stand-alone words. German is an *agglutinative* language, which forms compound words like *Lebensversicherungsgesellschaftsangestellter* (life insurance company employee). Old English was an agglutinative language like German,

# 停用词

- 出现非常频繁但又不承载具体意义的词，可能需要将它们从系统中去除或者特殊处理
  - a the and or as be am is are by from for
- 如何去除
  - 显式: 停用词表
  - 隐式: 去除集合中最频繁的k个词
- 去除停用词表的效果
  - 极大提高系统效率(停用词几乎出现在所有文档中)
- 带来的问题（自然语言多变）
  - To be or not to be
  - with or without you
  - The The
- 现在的倾向：不去除或者尽量少去除

# 词干处理(Stemming)

- 很多次在信息检索中可以当成一个词看待
  - 单复数/名词形容词副词/时态
- 暴力手段:将词尾去掉
  - ponies => poni, individual => individu
- 词干未必是一个词, 不同意义相近的词被映射到同一个词干
- 英文词干处理工具
  - Porter (<http://tartarus.org/martin/PorterStemmer/>)
  - Krovetz

# 词干处理效果

*Two households, both alike in dignity,  
In fair Verona, where we lay our scene,  
From ancient grudge break to new mutiny,  
Where civil blood makes civil hands unclean.  
From forth the fatal loins of these two foes*

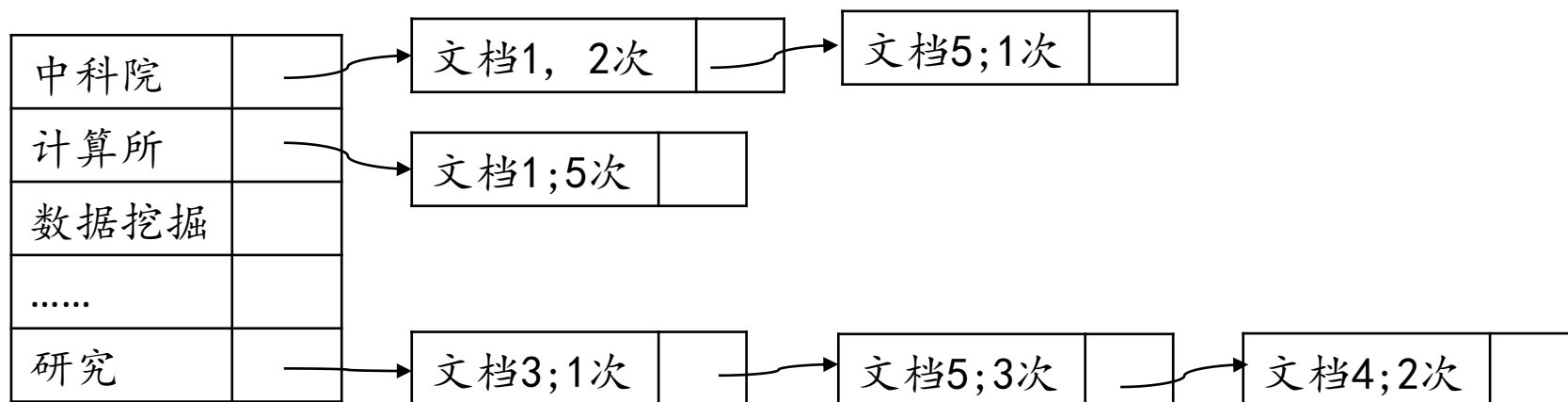


*Two household, both alik in digniti,  
In fair Verona, where we lay our scene,  
From ancient grudg break to new mutini,  
Where civil blood make civil hand unclean.  
From forth the fatal loin of these two foe*



# 文档表达——倒排索引

	检索词1	检索词2	...	检索词 M
文档1	1			3
文档2		2		
...				
文档N		5		7



# 传统信息检索模型

- 布尔模型(非排序)
- 向量空间模型
- BM25
- 语言模型(Language Models for Information Retrieval, LM4IR)

# 布尔检索模型

- 布尔变量：表示一个词是否出现在一个文档中
- 布尔运算符：AND, OR, NOT
- 布尔查询：变量与运算符的组合
  - Brutus AND Caesar AND NOT Calpurnia
  - NOT ((Duncan AND Macbeth) OR (Capulet AND Montague))
- 查询结果
  - 非排序
  - 满足查询的所有文档集合

# 文档表达

- Word-Doc矩阵表示单词是否出现在文档中
  - 列：文档中出现过哪些单词
  - 行：单词出现在哪些文档中
  - 查询处理：首先选择行(过滤掉绝大部分无关文档)，然后将布尔查询应用于每一个相关的列(文档)

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

# 布尔模型应用

- 曾经广泛应用于商业系统中
  - 稳定、简单可控、易于理解
- 现在很多系统依然依赖布尔模型
  - Outlook Email搜索
  - 图书馆图书搜索
  - Patent搜索

# 举例：图书检索系统



快速检索 | 高级检索

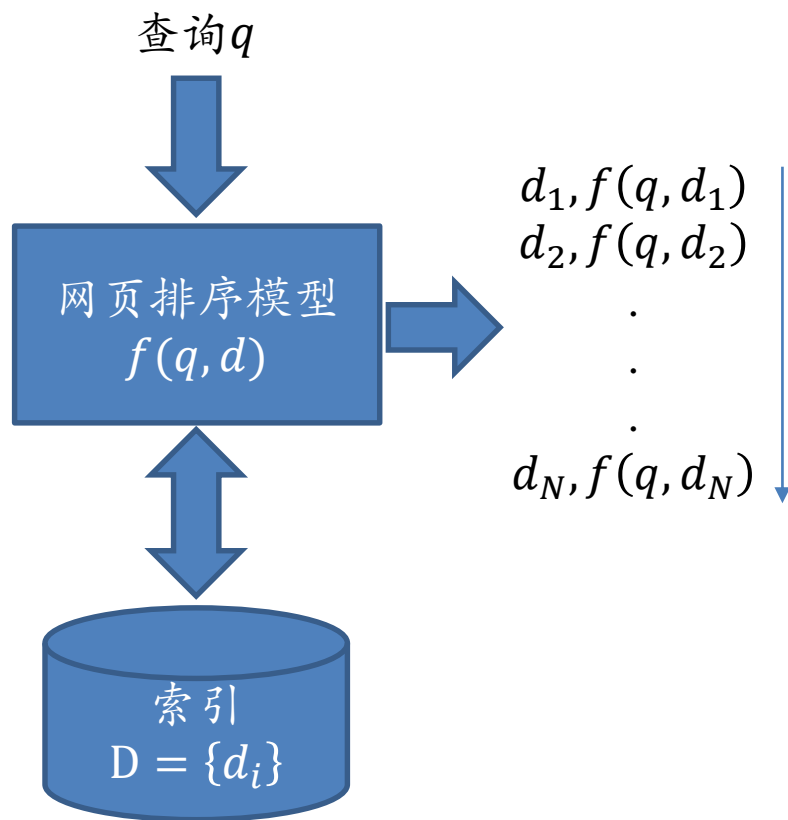
逻辑	检索项	检索词
	书名 ▼	Dreamweaver
并且 ▼	书名 ▼	ASP
并且 ▼	主题词 ▼	
出版年代从 2004年 ▼ 到 2007年 ▼		
排序	出版日期 ▼	降序 ▼
每页显示 10 ▼ 条记录		

检索 重填

选择检索范围

- 检索结果为无序的文档集合
  - 条件过松：文档太多无法一一人工检查
  - 条件过严：返回空集

# 更好的解决方案:排序



CCIR 2015

网页

新闻

贴吧

知道

音乐

图片

视频

地图

文库

更多»

## [第二十一届全国信息检索学术会议\(CCIR2015\)](#)

CCIR2015青年学者讲坛(8月24日)时间安排及学者介绍 Tutorial A(14:00-15:10):How to generate a good word embedding? 主讲人:刘康, 刘康, 博士, 中科院自动化所...

[www.ccir2015.com/](#) - 百度快照 - 评价

## [第二十一届全国信息检索学术会议\(CCIR2015\)征文通知](#)

第二十一届全国信息检索学术会议(CCIR2015)开始征稿了,会议官方网站为<http://www.ccir2015.com>,欢迎大家踊跃投稿参会,投稿截止日期为2015年4月30日,更多详...

[www.cs.sdu.edu.cn/getN...](#) - 百度快照 - 评价

## [...和程序委员会主席 - 第二十一届全国信息检索学术会议\(CCIR2015\)](#)

2015年4月1日 - 2015-04-01 21:23最新消息:CCIR2015指导委员会经过讨论确定:大会主席为清华大学的马少平教授。程序委员会主席为山东大学的马军教授。特向两位主席...

[ccir2015.com/...jsp?id...](#) - 百度快照 - 评价

## [第二十一届全国信息检索学术会议\(CCIR2015\)](#)

赞助形式与标准 赞助单位 top ©2015 第二十一届全国信息检索学术会议(CCIR2015) 版权所有 技术支持:凡科建站电脑版 在线留言 在线地图...

[m.ccir2015.com/](#) - 百度快照 - 评价

## [管理科学与工程学院](#)

2014年5月8日 - 2 全国信息检索学术会议(CCIR) 由中国计算机学会(CCF)和中国中文信息学会...各刊发表会议“全篇论文”截止日期2015年5月。大会论文投稿咨询: email:...

[einfo.dufe.edu.cn/ind....](#) - 百度快照 - 80%好评

# 更好的解决方案：排序

- 检索排序模型的先驱



Karen Spärck Jones



Stephen Robertson



Keith van Rijsbergen



# 向量空间模型

## (Vector Space Model, VSM)

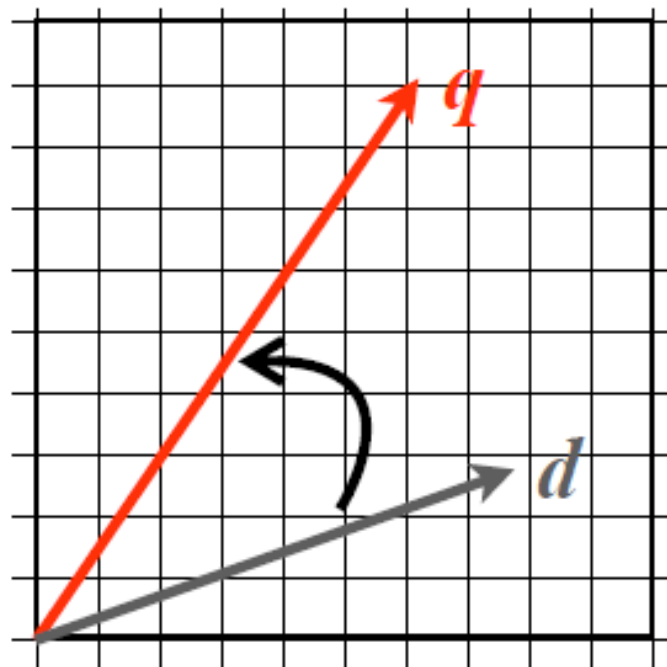
- 将查询字符串表达为带权重的tf-idf向量(查询向量)
- 类似，将文档字符串表达为带权重tf-idf 向量(文档向量)
- 计算查询向量和文档向量的余弦相似度
- 将文档按照其与查询的相似度分值从大到小进行排序
- 返回前K个(e.g.,  $K = 10$ )文档并展示给用户

# 什么是tf-idf ?

- 如何计算文档/查询中的词权重 $d_i$ 和 $q_i$ ?
- $\text{tf-idf}(w, d)$ : 衡量某一个词在文档中的重要性
  - $\text{tf}(w, d)$ : term frequency, 词 $w$ 在文档 $d$ (查询)中出现的次数。  
 $\text{tf}(w, d)$ 越大, 对文档 $d$ 而言 $w$ 越重要
  - $\text{df}(w)$ : document frequency, 在整个数据集合中, 包含 $w$ 的文档个数。 $\text{df}(w)$ 越大,  $w$ 越不重要。极端情况,  $w$ =“的”, 有可能在每一个文档中都出现(停用词)。注意 $\text{df}(w)$ 与文档 $d$ 没有直接关系
  - $\text{idf}(w)$ : inverse document frequency,  
 $\text{idf}(w) = \log \frac{N}{\text{df}(w)}$ , 其中 $N$ 为整个集合中文档数目
  - $\text{tf-idf}(w, d) = \text{tf}(w, d) * \text{idf}(w)$
- 词袋(bag of words)假设: 不考虑词在查询(文档)中出现的位置和顺序

# VSM

$$\begin{aligned} \text{sim}(\mathbf{q}, \mathbf{d}) &= \frac{\mathbf{q} \cdot \mathbf{d}}{\|\mathbf{q}\| \|\mathbf{d}\|} \\ &= \frac{\sum_{i=1}^{|\mathcal{V}|} \mathbf{q}_i \mathbf{d}_i}{\sqrt{\sum_{i=1}^{|\mathcal{V}|} \mathbf{q}_i^2} \sqrt{\sum_{i=1}^{|\mathcal{V}|} \mathbf{d}_i^2}} \\ &= \frac{\mathbf{q}}{\|\mathbf{q}\|} \cdot \frac{\mathbf{d}}{\|\mathbf{d}\|} \end{aligned}$$



# BM25

- BM25 “Best Match 25”
  - 在Okapi检索系统中开发
  - 在TREC竞赛中逐步完善
  - 是信息检索中最广为人知的排序模型之一

Foundations and Trends® in  
Information Retrieval  
Vol. 3, No. 4 (2009) 333–389  
© 2009 S. Robertson and H. Zaragoza  
DOI: 10.1561/15000000019

**now**  
the essence of knowledge

## **The Probabilistic Relevance Framework: BM25 and Beyond**

By Stephen Robertson and Hugo Zaragoza

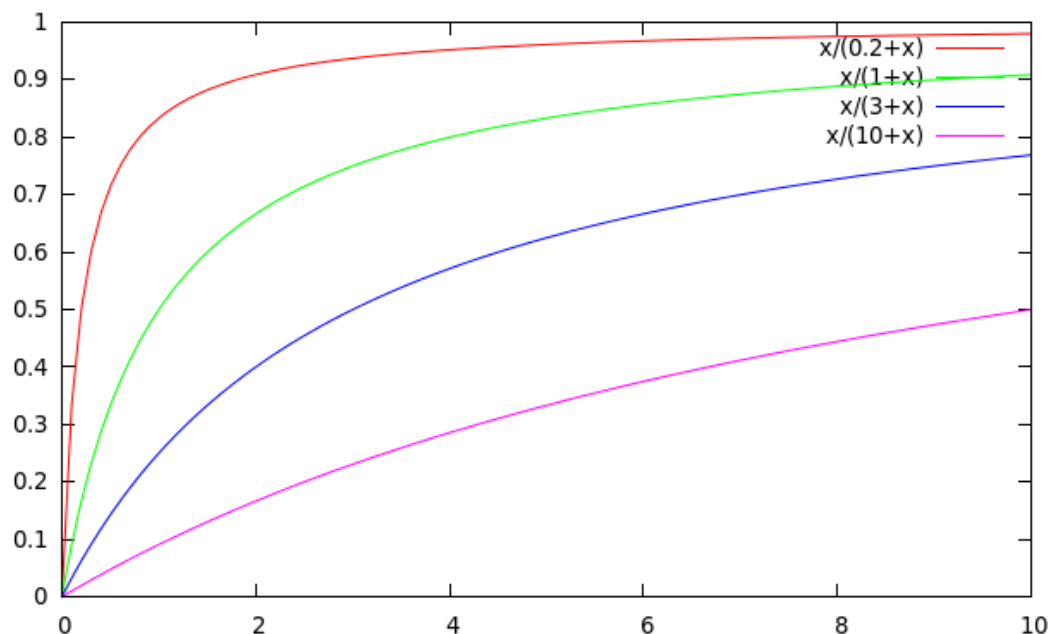
# BM25

$$BM25 = \sum_{i \in q} \log \frac{N}{df_i} \cdot \frac{(k_1 + 1)tf_i}{k_1((1 - b) + b \frac{dl}{avdl}) + tf_i}$$

- $avdl$ : 集合中平均文档长度
- $k_1$ : 控制因tf的增大最终排序值的速度
  - $k_1 = 0$ : 二值模型, 只反映词是否出现, 不考虑出现次数
  - $k_1$  无穷大: 反映真正的tf值
- $b$ : 控制文档长度归一化程度
  - $b = 0$ : 不考虑文档长度对最终分值的影响
  - $b = 1$ : 考虑文档长度平均文档长度的相对值
- 经验值:  $k_1 = 1.2 \sim 2$ ,  $b = 0.75$

# BM25分值与 $tf$ 的关系

$$\frac{tf}{k_1 + tf}$$



- 饱和函数(saturation function)

- 相对 $tf_i$ 为单调增函数
- 增长迅速饱和
- 参数 $k_1$ 控制饱和速度( $k_1$ 越大, 饱和越慢)

$$BM25 = \sum_{i \in q} \log \frac{N}{df_i} \cdot \frac{(k_1 + 1)tf_i}{k_1((1 - b) + b \frac{dl}{avdl}) + tf_i}$$

# 文档长度归一化

- 文档长度定义：

$$dl = \sum_i tf_i$$

- 文档长度归一化部分

$$B = \frac{dl}{\sum_i dl} (1 - b) + b \frac{dl}{\sum_i dl}, \quad 0 \leq b \leq 1$$

- $b = 1$  : 全文档长度归一化
- $b = 0$  : 不进行归一化

$$BM25 = \sum_{i \in q} \log \frac{N}{df_i} \cdot \frac{(k_1 + 1)tf_i}{k_1((1 - b) + b \frac{dl}{\sum_i dl}) + tf_i}$$

# 语言模型

- 语言模型(language model): 单词序列上的概率分布  $P(w_1, w_2, \dots, w_m)$ 
  - 给定一个文档集合, 如何计算?
- Unigram language model
  - 词袋假设:  $P(w_1, w_2, \dots, w_m) = P(w_1)P(w_2) \dots P(w_m)$
  - 估计每一个词的出现概率:  $P(w) = \frac{\#w \text{ in collection}}{\# \text{ all words in collection}}$

Terms	Probability
a	0.1
world	0.2
likes	0.05
we	0.05
share	0.3
...	...

$$\sum_{w \in V} P(w) = 1$$

需要估计的参数数目:  
词表中单词数  $|V|$  (10万~100万)



# 语言模型(续)

- N-Gram language model
  - 在给定上文的条件下, 估计每一个词的出现概率 $P(w_n | w_{n-1}, w_{n-2}, \dots, w_1)$
  - $P(w_1, w_2, \dots, w_m) = P(w_1)P(w_2 | w_1) \dots P(w_m | w_{m-1} \dots w_1)$
- Bi-gram language model
  - 马尔科夫假设  $P(w_n | w_{n-1}, w_{n-2}, \dots, w_1) = P(w_n | w_{n-1})$
  - $P(w_1, w_2, \dots, w_m) = P(w_1)P(w_2 | w_1)P(w_3 | w_2) \dots P(w_m | w_{m-1})$



Unigram

Bigram

# Bi-Gram Language Model

- 给定训练文档集合，统计给定一个词之后，其它词出现的概率

$$P(w_i|w_j) = \frac{\#(w_j w_i) \text{ in collection}}{\# w_j \text{ in collection}}$$

- 需要估计的参数数目： $|V| \times |V| + |V|$

$w_i$	Probability
a	0.1
world	0.2
likes	0.05
we	0.05
share	0.3
...	...

$$\sum_{w_i \in V} P(w_i | \text{"is"}) = 1$$

$w_j = \text{"is"}$

$w_i$	Probability
a	0.01
world	0.25
likes	0.15
we	0.01
share	0.02
...	...

$$\sum_{w_i \in V} P(w_i | \text{"that"}) = 1$$

$w_j = \text{"that"}$

# 数据稀疏问题

- 对于某些稀有词，在训练集合中出现次数非常少，Unigram模型对其概率的估计不准确
  - 如：“希格斯玻色子”可能在数据集合中仅仅出现<10次
- 在Bigram情况下，问题更加严重
  - 在给定“希格斯玻色子”条件下，估计 $P(w|“希格斯玻色子”)$ ，最多只有10个单词可以估计到有效概率，其它全为0
- 很多情况下，我们不希望出现0概率
  - $P(w_1, w_2, \dots, w_m) = P(w_1)P(w_2|w_1) \dots P(w_m|w_{m-1} \dots w_1)$
  - 任意一项等于0将导致整体的概率为0，意味着句子中任意出现一个稀有词将毁掉所有的计算
- 解决方案：平滑化（将在后面详细介绍）

# 语言模型应用

- 语音识别

- 目标:输入语音向量序列A(sequence of acoustic vectors)，输出对应的单词序列W

- 建模: $P(W|A) = \frac{P(A|W)P(W)}{P(A)}$

- 寻找使得 $P(W|A)$ 最大的单词序列

$$W^* = \operatorname{argmax}_W P(A|W)P(W)$$

- 稀疏导致的问题

- 含有稀有词的句子概率为0

语音语言转  
化概率

语言模型：语言  
本身出现概率

# 语言模型应用

- 中文输入法联想
  - 输入：用于已输入的前K个单词序列
  - 目标：预测用户将要输入的单词
  - 建模：高阶语言模型

$$P(w_n | w_{n-1} \cdots w_{n-K})$$

- 训练数据：用户历史输入记录
- 展示：按照上述K-gram语言模型对词进行排序，取top-N
- 数据稀疏将导致概率为0，无法联想到新词



Foundations and Trends® In  
Information Retrieval  
2:3 (2008)

# Statistical Language Models for Information Retrieval

A Critical Review

ChengXiang Zhai

语言模型应用:信息检索

# 用于信息检索的语言模型

## Language models (LMs) for IR

- 利用每一个待排序的文档 $d$ 训练一个语言模型，去生成用户输入的查询 $q$
- 处理流程:
  1. 定义生成模型的细节
  2. 估计模型参数 $P(w|d)$ （为每一个文档估计一个模型）
  3. 平滑化（防止零概率）
  4. 将文档对应的生成模型应用于查询，计算生成概率
  5. 按照生成概率将文档排序，取top N展现给用户

# 用于信息检索的语言模型

- 给定查询 $q$ 和一个文档 $d$ ，对文档的打分为 $P(d|q)$
- 应用贝叶斯公式 $P(d|q) = \frac{P(q|d)P(d)}{P(q)}$ 
  - $P(q)$ : 对所有文档都一样，可以忽略
  - $P(d)$ : 文档的先验，比如重要度等
  - $P(q|d)$ : 文档与查询的匹配程度
- 对 $P(q|d)$ 用语言模型进行估计
  - Unigram假设  $P(q|d) = P(q_1q_2, \dots, q_M|d) = \prod_{i=1}^M P(q_i|d)$
  - 对 $P(q_i|d)$ 的估计:  $P(q_i|d) = \frac{tf(q_i, d)}{|d|}$



# 参数平滑(Smoothing)

- 查询q: 中国\_科学院\_大学
- 文档d: 科学院\_大学\_计算机\_学院
- 零概率问题
  - $P(q|d)=P(\text{中国}|d)P(\text{科学院}|d)P(\text{大学}|d)=0 * 0.25 * 0.25 = 0$
  - 原因: 查询词“中国”未在文档中出现
  - 造成后果: 其它所有词的贡献都被抹掉, 不符合IR现实需求
- 参数平滑化: 使得每一个在字典中出现的词(即使其没有在文档d中出现)都有一定正概率(劫富济贫)

# 平滑化方法：混合平滑模型

- 基于整个文档集合估计出一个“背景”语言模型  $P(w|C) = \frac{tf(w,C)}{|C|}$ 
  - 假设字典中的每一个词在至少一个文档中出现过，因此  $P(w|C) > 0$
- 基于当前文档  $d$  估计出文档语言模型  $P(w|d) = \frac{tf(w,d)}{|d|}$ 
  - $P(w|d)$  稀疏，对于不出现在  $d$  中的  $w$ ， $P(w|d) = 0$
- 线性插值
$$P_{mix}(w|d) = \lambda P(w|d) + (1 - \lambda)P(w|C)$$

# 平滑化方法：狄里克莱平滑

- Dirichlet平滑

$$P_{dir}(w|d) = \frac{tf(w, d) + \mu P(w|C)}{|d| + \mu}$$
$$= \frac{tf(w, d)}{|d| + \mu} + \frac{\mu P(w|C)}{|d| + \mu}$$

- 一般设置  $\mu = 100 \sim 200$

- 直观解释

- 文档d中还有 $\mu$ 个位置未被观测到(增长后文档长度为 $|d| + \mu$ )
  - 这 $\mu$ 个位置被字典中所有的词瓜分，瓜分的比例为其在整个文档集合中出现的比例(注意一个位置可以被多个词一起瓜分，每个词出现次数可以小于1)。

# 举例

- $D=\{d1, d2\}$ , Query  $q$ : **Michael Jackson**
  - $d1$  : Jackson was one of the most talented entertainers of all time
  - $d2$ : Michael Jackson anointed himself King of Pop
  - $P(q|d1)=\frac{0}{11} * \frac{1}{11} = 0$ ,  $P(q|d2)=\frac{1}{7} * \frac{1}{7} = \frac{1}{49}$ ,  $P(\text{Michael}|C)=\frac{1}{18}$ ,  $P(\text{Jackson}|C)=\frac{2}{18}$
- 混合模型( $\lambda = 0.5$ )

$$P_{mix}(q|d1) = \left( \frac{0}{11} * \frac{1}{2} + \frac{1}{18} * \frac{1}{2} \right) * \left( \frac{1}{11} * \frac{1}{2} + \frac{2}{18} * \frac{1}{2} \right) \approx 0.003$$

$$P_{mix}(q|d2) = \left( \frac{1}{7} * \frac{1}{2} + \frac{1}{18} * \frac{1}{2} \right) * \left( \frac{1}{7} * \frac{1}{2} + \frac{2}{18} * \frac{1}{2} \right) \approx 0.013$$

- 狄里克莱平滑( $\mu = 5$ )

$$P_{dir}(q|d1) = \left( \frac{0 + \frac{5}{18}}{11 + 5} \right) \left( \frac{1 + 5 * \frac{2}{18}}{11 + 5} \right) = 0.0017$$

$$P_{dir}(q|d2) = \left( \frac{1 + \frac{5}{18}}{7 + 5} \right) \left( \frac{1 + 5 * \frac{2}{18}}{7 + 5} \right) = 0.0138$$

# 传统排序模型总结

- 估计用户输入文本 $q$ 与文档文本 $d$ 之间的相关度(relevance)
  - 总体相关度是每一个查询词相关度的和
  - 考虑因素包括
    - 词频 $tf$ : 查询词在文档中的出现次数; 次数越多越相关, 但是会趋于饱和(saturate function)
    - 词的文档频率 $df$ : 衡量词的重要性,  $df$ 越大, 词越不重要, 极端情况为停用词
    - 文档长度 $dl$ : 长的文档会削弱相关度

# 排序评价指标

# 评价目的

- 比较不同模型、不同参数设置的优劣，为模型和参数选择提供依据
  - 在线评价
    - 上线应用→搜集用户行为→评价
    - 需要系统和真实用户，代价高、周期长，体现用户真实体验，常作为上线前最后的比较和评估
  - 离线评价
    - 标注数据→应用模型得到排序→评价
    - 可在相同数据上重复对比不同模型
- 本次课程关注离线评价指标

# 标注数据 (TREC)

```
<top>

<num> Number: 451
<title> What is a Bengals cat?

<desc> Description:
Provide information on the Bengal cat breed.

<narr> Narrative:
Item should include any information on the
Bengal cat breed, including description, origin,
characteristics, breeding program, names of
breeders and catteries carrying bengals.
References which discuss bengal clubs only are
not relevant. Discussions of bengal tigers are
not relevant.

</top>
```

```
451 0 WTX003-B26-240 0
451 0 WTX003-B26-249 1
451 0 WTX003-B26-252 0
451 0 WTX003-B26-263 0
451 0 WTX003-B31-203 0
451 0 WTX004-B07-355 0
451 0 WTX004-B10-130 0
451 0 WTX004-B11-116 0
451 0 WTX004-B20-66 0
451 0 WTX004-B22-112 0
451 0 WTX004-B22-42 0
451 0 WTX005-B15-324 0
451 0 WTX005-B18-28 0
451 0 WTX005-B42-245 0
451 0 WTX006-B19-257 0
451 0 WTX006-B20-23 0
451 0 WTX006-B20-31 0
451 0 WTX006-B20-41 0
451 0 WTX006-B20-48 0
451 0 WTX006-B21-236 0
451 0 WTX006-B50-26 0
451 0 WTX007-B25-24 0
451 0 WTX007-B35-206 0
451 0 WTX007-B42-101 0
451 0 WTX007-B42-124 0
451 0 WTX007-B50-87 0
451 0 WTX008-B17-23 0
451 0 WTX008-B26-172 0
451 0 WTX008-B37-10 2
451 0 WTX008-B38-114 0
451 0 WTX008-B39-477 0
451 0 WTX008-B39-479 0
451 0 WTX008-B39-480 0
451 0 WTX008-B40-124 0
451 0 WTX009-B02-590 0
```

<http://trec.nist.gov/data/t9.web.html>



# 标注数据以及评价指标

- 标注数据三元组  $(q, d, r)$ 
  - $r$ : 人工相关度标签
  - 二值相关度: 0表示不相关, 1表示相关
  - 多级相关度: 2相关、1部分相关、0不相关;  
或者5级: Bad, Fair, Good, Excellent, Perfect
- 常用排序评价指标
  - P@K
  - MAP
  - NDCG

# 排序评价指标

- 基于二值相关度标签
  - Precision@K (P@K)
  - Mean Average Precision (MAP)
  - Mean Reciprocal Rank (MRR)
- 基于多值相关度标签
  - Normalized Discounted Cumulative Gain (NDCG)

# Precision at K ( $P@K$ )

- 设置一个排序位置K
- 计算前K个位置相关的文档所占百分比
- 忽略排在K个位置之外的文档
- 举例：绿色—相关；红色—不相关

–  $P@1 = 1/1$



–  $P@2 = 1/2$



–  $P@3 = 2/3$



–  $P@4 = 2/4$

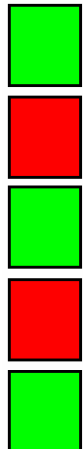


–  $P@5 = 3/5$



# Mean Average Precision (MAP)

- 考虑出现过相关文档的位置
  - $K_1, K_2, \dots K_R$
- 分别计算位置  $K_1, K_2, \dots K_R$  的  $P@K$
- Average Precision = average of  $P@K$

• 例如:  
$$avgP = \frac{1}{3} \cdot \left( \frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) \approx 0.76$$

- MAP: 在所有的测试查询上计算 avgP 的平均值

# MAP

- 是信息检索中广泛使用的评价准则
  - 从文档角度考虑问题，如果一个相关的文档没有出现，则对应的相关度贡献为0
  - **MAP**:M是在所有查询上的平均，每一个查询的重要性相同
  - **MAP**假设用户在提交一个查询后，希望看到多个相关的文档

# 多级别标注

**YAHOO!** Web Images Video Local Shopping More ▼

Toyota safety Search Options ▼

Search Pad  
SearchScan - On

108,000,000 results for **Toyota safety**:

Show All

Toyota  
Motor Trend  
CarsDirect  
Shopping Sites

Also try: [toyota safety ratings](#), [toyota safety recall](#), [More...](#)

**Sponsored Results**

**Toyota Recall**  
Toyota Takes Care of its Customers. Read the FAQs at **Toyota.com**.  
[www.Toyota.com/Recall](#)

**Toyota Safety**  
& Latest Prices. Free Info. **Toyota** Research, Reviews.  
[www.Toyota.Edmunds.com](#)

**TOYOTA | Car Safety Innovation and Technology**  
Toyota home page for car **safety** and car technology Prius model.  
[www.safetytoyota.com](#) - [Cached](#)

**Toyota home page for car safety and car technology ...**  
We are presenting **Toyota's safety** technologies for cars. We clearly explain about car **safety** and car technology using movies and more.  
[www.safetytoyota.com/en-gb](#) - [Cached](#)

**Toyota Safety Ratings - Toyota Safety Features - Motor Trend ...**  
MotorTrend offers **Toyota safety** ratings, comprehensive auto **safety** reports, and more. View a all of the standard **Toyota safety** features. ...  
[motortrend.com/new\\_cars/07/toyota/safety\\_ratings/index.html](#) - 149k - [Cached](#)

**Toyota Motor Europe Corporate Site Safety**  
Our approach. **Toyota** believes that all stakeholders in the road **safety** equation share a responsibility to reduce the frequency of road accidents. ...  
[www.toyota.eu/Safety](#) - [Cached](#)

**[PDF] pdf European Safety Brochure 2005**  
4047k - Adobe PDF - [View as html](#)  
not guarantee that all accidents or injuries will be avoided when driving a **Toyota** and/or Lexus brand motor vehicle equipped with the **safety** systems ...  
[www.toyota.no/Images/Safety\\_Brochure\\_tcm308-344461.pdf](#)

**Toyota - Star Safety System**  
Star **Safety** System ... **Toyota** Mobility Program. Careers. Contact Us. Home. contact us. site map. your privacy rights. legal terms. **Toyota** Newsroom. sign up for info ...  
[www.toyota.com/vehicle/demo/star/safety.html](#) - 58k - [Cached](#)

**Sponsored Results**

**Safety for a Toyota**  
Research **Safety** Ratings and Reviews For New Car at Kelley Blue Book.  
[www.kbb.com](#)

**Toyota Safety**  
Find **Toyota Safety** dealers, new cars, prices, and photos.  
[www.NewCars.org](#)

**Toyota Safety**  
**Toyota safety** Discount Prices Save Money Shopping Online Today.  
[www.sma.com](#)

**Safety Toyota**  
Explore 5,000+ Pro Sports Choices. Save On Safety Toyota.  
[BaseballGear.Shopzilla.com](#)

[See your message here...](#)

**fair**  
**fair**  
**Good**  
**Bad**

# Discounted Cumulative Gain (DCG)

- 两个假设:
  - **文档的相关性标注**: 文档与查询的相关度可以被分为多个级别, 高度相关的文档比一般相关的文档产生更高的效应(utility)
  - **文档的位置**: 相关的文档被排序的位置越靠后, 因为被用户看到的可能性越小, 其产生的效应(utility)越低。

# Discounted Cumulative Gain

- Gain: 一个文档对用户产生的Gain与其与查询的相关度有关
  - 例如  $Gain = 2^{\text{label}} - 1$ : Bad-0分, Fair-1分, Good-3分, Excellent-7分, Perfect-15分
- Discounted Cumulative Gain:
  - Cumulative Gain: 多个文档对用户产生的Gain总量为它们Gain的总和
  - Discounted Cumulative Gain: 考虑到用户从上往下阅读的习惯, 按照位置对每个文档的Gain进行打折, 再进行求和
  - 打折方法:  $1/\log_2(rank + 1)$



# 只考虑前N个文档：DCG@N

- CG@N

- 假设前N个文档的Gain为  $r_1, r_2, \dots, r_N$

- $CG = r_1 + r_2 + \dots + r_N$

- DCG@N

- $DCG = r_1/\log_2 2 + r_2/\log_2 3 + r_3/\log_2 4 + \dots + r_N/\log_2 (N+1)$

# Normalized DCG (NDCG)

- DCG@N的缺陷：取值范围不确定，最优排序的 $DCG@N \neq 1$
- Normalized Discounted Cumulative Gain (NDCG) at N
  - 利用最优排序对DCG@N进行归一化
  - 最优排序为按照用户标注对文档进行排序（可能不止一个最优排序，如交换两个标注值一样的文档位置）
- NDCG在互联网公司应用广泛
  - 微软必应搜索
  - 百度搜索
  - 搜狗搜索

# NDCG计算举例

4 个文档:  $d_1, d_2, d_3, d_4$

位置	用户标注		排序函数1		排序函数2	
	文档排序	Gain	文档排序	Gain	文档排序	Gain
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	NDCG <sub>GT</sub> =1.00		NDCG <sub>RF1</sub> =1.00		NDCG <sub>RF2</sub> =0.9652	

$$DCG_{GT} = DCG_{RF1} = \left( \frac{2}{\log_2 2} + \frac{2}{\log_2 3} + \frac{1}{\log_2 4} + \frac{0}{\log_2 5} \right) = 3.7619$$

$$DCG_{RF2} = \left( \frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{2}{\log_2 4} + \frac{0}{\log_2 5} \right) = 3.6309$$

# 传统排序模型总结(个人看法)

- BM25是使用最多(?)的传统排序模型
  - 简单、易于计算
  - 在不同的实验环境中表现极为稳定，对参数不敏感，常作为IR实验的baseline
  - BM25扩展BM25F：假设文档存在多个域  
S. Robertson et al., Simple BM25 extension to multiple weighted fields. CIKM 2004.
- Language models for IR
  - 概率化建模，由文档模型生成查询
  - 平滑方法的选择:较多选择狄里克莱平滑
- 评价方法
  - 传统文本检索，标注为两类，P@N和MAP使用较多
  - 互联网搜索，标注为多类，NDCG@N应用广泛

# 提纲

- 互联网搜索介绍
- 传统相关性排序模型
- 搜索结果多样化排序
- 总结

# 相关性排序模型的特点

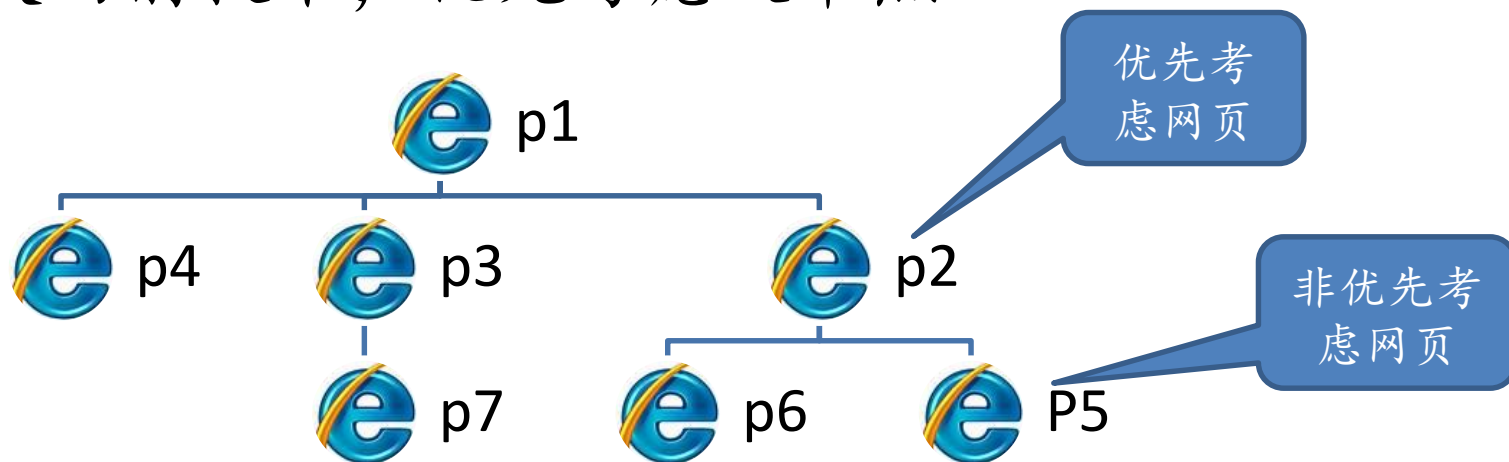
- 打分函数
  - 文档独立打分
    - 分值与位置无关
    - 分值与其他文档无关
  - 单一打分函数 $f(q, d)$ 
    - 只依赖于查询和待打分文档本身
- 排序过程
  - 按打分分值从大到小排序

# 相关性排序模型的局限性

- 打分时不考虑网页将要显示的位置，但是
  - 不同网页将显示在不同的排序位置
  - 排序较高的网页可能影响排序较低网页的效应(搜索结果多样化)
- 网页独立打分
  - 网页间存在多种关系(相似、链接上下级等)
  - 多个网页将显示在一个结果页面中

# 话题提取(Topic Distillation)

- Generate a short top-N list, even when a very large set of on-topic documents are available [Craswell & Hawking, TREC '03]
- 在相关的前提下，优先考虑父节点



需要考虑网页间的链接和目录关系



# 搜索结果多样化

## (Search Result Diversification)

- 搜索结果多样化
  - 用最少的文档覆盖最多的子话题(subtopic)
- 查询 “programming language”

“好” 的结果	“不好” 的结果
Java	Java
C++	Java
Python	Java

需考虑网页间的相似度

# 为何关注多样化排序?

Google jaguar

网页 图片 新闻 地图 视频 更多 搜索工具

找到约 443,000,000 条结果 (用时 0.36 秒)

**捷豹中文官方网站 - jaguar.com.cn**  
www.jaguar.com.cn/  
全铝航空科技,轻量化2.0升i4涡轮增压发动机或3.0升V6机械增压发动机,更显灵动从容

相关搜索: jaguar汽车价格 jaguar 价格

**捷豹汽车\_最具生命力的英伦汽车-Jaguar中国官方网站**  
www.jaguar.com.cn/  
历经时代变迁,捷豹(jaguar)始终致力于为汽车机械赋予灵魂,有无可复制的生命力的极致奢华。捷豹汽车,创新精神、诱人设计、卓越性能,全新

**【捷豹 Jaguar】汽车最新资讯 图片 视频 新车**  
newcar.xcar.com.cn/  
捷豹汽车: 爱卡汽车  
车图片,捷豹汽车视频,以  
捷豹XF - 捷豹XJ - 捷豹

**Jaguar**  
www.jaguar.com.cn/  
Official worldwide  
specific markets

**jaguar的**  
八柱一陈欧  
中  
中

**【捷**  
car a  
凤凰  
价格

**【进**  
car b  
进口  
进口

**【捷**  
car b  
捷豹  
1922

**捷豹 - 维基百科，自由的百科全书**  
zh.wikipedia.org/zh-cn/捷豹 转为简体网页  
捷豹汽车有限公司（英语：Jaguar Cars Limited）是英国的一家豪华汽车生产商，总部起初座落于英格兰考文垂的布兰兰，后迁至考文垂的惠特利。1922年成立之初制造...

**捷豹汽车|捷豹汽车报价及图片-网上车市**  
product.cheshi.com, 汽车大全  
捷豹汽车的历史源远流长，可以追溯到1922年威廉·里昂斯爵士创造出第一辆摩托车跨斗之时。1932年，“捷豹(Jaguar)”的名字首次随着一款完全独自设计制造的...

**捷豹 - 汽车 - 网易**  
product.auto.163.com/brand/1711.html  
捷豹(jaguar)是塔塔汽车集团旗下品牌，品牌起源于英国。捷豹品牌热门车型包括捷豹XF、捷豹XJ、捷豹F-TYPE、捷豹XKR等。网易汽车为您提供捷豹全车型、最新...

Diverse user  
interests

Google 克里米亚公投

网页 地图 新闻 图片 视频 更多 搜索工具

找到约 3,920,000 条结果 (用时 0.16 秒)

**2014年克里米亚归属公投 - 维基百科，自由的百科全书**  
zh.wikipedia.org/zh-cn/2014年克里米亚归属公投 转为简体网页  
2014年克里米亚归属公投，是克里米亚自治共和国政府於2014年3月16日發起的一場公投，讓克里米亞選民決定是否從烏克蘭獨立，並且加入俄羅斯聯邦。参与者包含...

背景 - 日期 - 公投选项 - 烏克蘭政府「中止公投決議」

**联合国大会通过决议称克里米亚公投无效 - 国际中心**  
news.ifeng.com  
2014年3月28日  
16

**克里米亚公投 - 国际新闻，乌克兰**  
news.sina.com.cn  
2014年3月28日 - 决议还说，16日在克里米亚自治共和国和塞瓦斯托波尔市举行的全民公投“无效”，“不能成为改变克里米亚自治共和国和塞瓦斯托波尔市地位的基础”。

**联合国大会决议称克里米亚公投无效 - 新闻中心 - 央视网(cctv.com)**  
news.cntv.cn/special/video/Crimea/index.shtml  
克里米亚全民公投于3月17日凌晨结束，民调显示：九成以上选民支持加入俄罗斯联邦。奥巴马强调公投违反乌克兰宪法，永远都不会得到美国和国际社会承认。

**乌克兰局势最新消息：克里米亚独立公投提前至3月30日举行 ...**  
www.guancha.cn/europe/2014\_03\_02\_209886.shtml  
2014年3月2日 - 乌克兰局势最新消息：综合外媒报道，乌克兰克里米亚自治共和国议会2月27日决定，今年5月就克里米亚的地位举行公民投票。但克里米亚总理阿克...

**联合国大会通过决议：克里米亚公投无效 - 新闻中心 - 中国网**  
news.china.com.cn, 新闻中心, 国际, 国际滚动  
2014年3月28日 - 乌克兰克里米亚自治共和国16日举行全民公投，近97%的投票者赞成克里米亚加入俄罗斯。18日，俄总统普京在克里姆林宫向克里米亚及塞瓦斯托波...

**克里米亚公投，解读中国的弃权票（陈破空）**  
www.rfa.org/mandarin/pinglun/chenpokong/ke-03182014114959.html  
2014年3月18日 - 3月16日，克里米亚举行公投，结果：近97%的投票者支持克里米亚脱离乌克兰、加入俄罗斯。俄罗斯总统普京（Vladimir Putin）迅速与克里米亚地区...

**俄总理首访公投后克里米亚俄决定划为经济特区 - 国际 - 环球网**  
world.huanqiu.com, 国际新闻, 独家  
2014年4月1日 - 梅德韦杰夫访问克里米亚，他是在公投并入俄罗斯后第一个访问该地区的俄罗斯领导人。

Reducing  
redundant

# 顺序文档选择过程

- 基于当前已排序(选择)文档，选择下一个文档
- 贪心算法：每次选择当前情况下最优文档
  - 输入:待排序文档集合 $D = \{d_1, \dots, d_N\}$ ，已选文档集合 $S = \phi$
  - 1. 为 $D$ 中的所有文档进行打分 $f(q, d, S)$ ，分数综合考虑相关性和多样化
  - 2. 选择分数最大的文档 $d$
  - 3.  $D \leftarrow D \setminus \{d\}, S \leftarrow S \cup \{d\}$
  - 4. 如 $D = \phi$ ，退出
  - 5. 转step 1
  - 6. 返回 $S$ 及文档加入 $S$ 的顺序

# 最大化边缘相关度

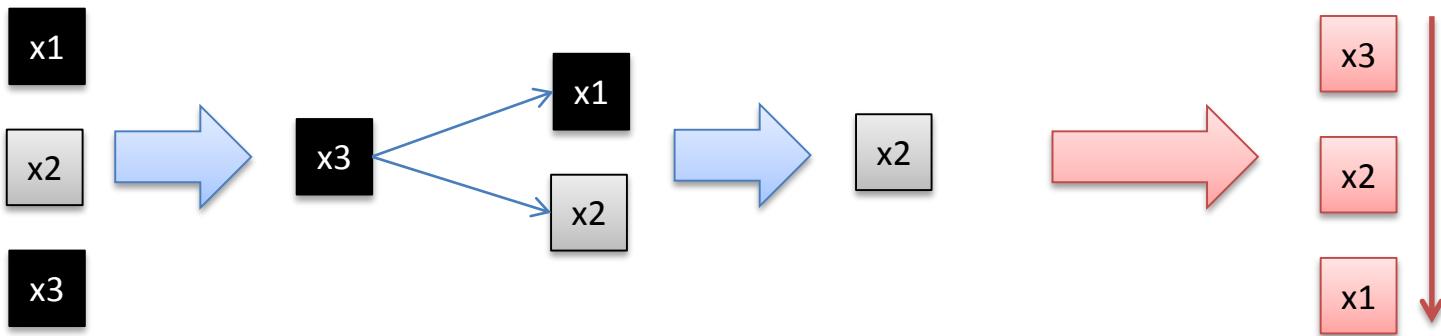
## Maximal Marginal Relevance (MMR)

- 搜索结果多样化中常用的方法
  - 打分：综合考虑查询-文档相关度与文档-文档关系(相似度)

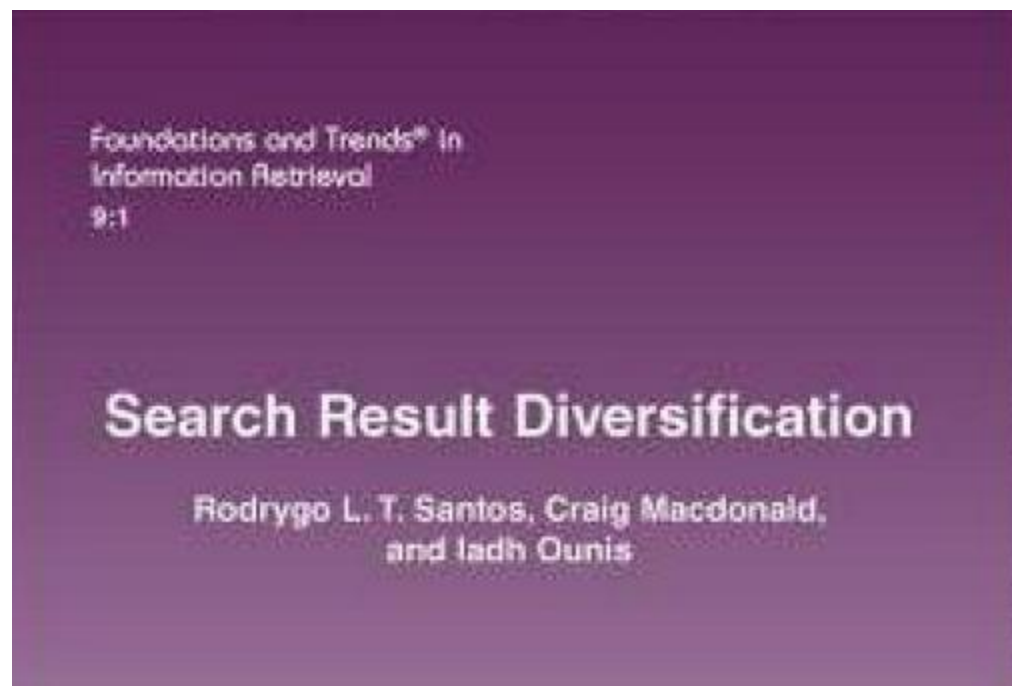
$$\text{MMR} \stackrel{\text{def}}{=} \text{Arg max}_{D_i \in R \setminus S} [\lambda \text{Sim}_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} \text{Sim}_2(D_i, D_j)]$$

相关度                      文档相似度

- 排序：顺序文档选择，最大化边缘相关度(maximal marginal relevance, MMR) [Carbonell & Goldstein, SIGIR '98]
  - 建模用户从上往下查看网页的习惯



# 更多搜索结果多样化工作



Rodrygo L. T. Santos, Craig Macdonald and Iadh Ounis (2015), "Search Result Diversification", Foundations and Trends® in Information Retrieval: Vol. 9: No. 1, pp 1-90.

# 提纲

- 互联网搜索介绍
- 传统相关性排序模型
- 搜索结果多样化排序
- 总结

# 本次课总结

- **排序**是互联网搜索中的核心问题
- 本次课：互联网搜索中的相关性排序
  - 相关性排序：考虑查询词在文档中出现的情况，构造打分函数，经典模型包括VSM、BM25、LMIR
  - 搜索结果多样化排序：突破独立性假设，综合考虑相关性和文档-文档间相似度
- 本次课介绍的排序模型基于专家知识构造出的打分函数
- 下一次课：利用监督学习的方法学习排序函数(排序学习, learning to rank)

**谢谢！**