

信息检索导论

An Introduction to Information Retrieval

第17讲 Web搜索

Web Search

授课人：古晓艳

中国科学院信息工程研究所/国科大网络空间安全学院

*改编自“An introduction to Information retrieval”网上公开的课件，地址 <http://nlp.stanford.edu/IR-book/>

提纲

- ① 上一讲回顾
- ② Web搜索系统
- ③ Web信息检索的特点
- ④ 互联网广告
- ⑤ 重复检测

提纲

- ① 上一讲回顾
- ② Web搜索系统
- ③ Web信息检索的特点
- ④ 互联网广告
- ⑤ 重复检测

奇异值分解

- 设矩阵 A 的维度为 $m \times n$ ，虽然 A 不是方阵，但是下面的矩阵却是方阵，且是对称方阵，维度分别为 $m \times m$ 、 $n \times n$ 。

$$AA^T \quad A^T A$$

- 分别对上面的方阵进行分解：

$$AA^T = U \Sigma U^T \quad A^T A = V \Sigma V^T$$

- 则 A 的奇异值分解为：

$$A = U \Sigma V^T$$

$$\begin{array}{c}
 \text{Blue box } A \\
 m \times n
 \end{array}
 =
 \begin{array}{c}
 \text{Green box } U \\
 m \times m
 \end{array}
 \times
 \begin{array}{c}
 \text{Blue box } \Sigma \\
 m \times n
 \end{array}
 \times
 \begin{array}{c}
 \text{Orange box } V^T \\
 n \times n
 \end{array}$$

SVD分解的例子 $C = U\Sigma V^T$: 所有的四个矩阵

C	d_1	d_2	d_3	d_4	d_5	d_6	
ship	1	0	1	0	0	0	
boat	0	1	0	0	0	0	
ocean	1	1	0	0	0	0	=
wood	1	0	0	1	1	0	
tree	0	0	0	1	0	1	
U	1	2	3	4	5		
ship	-0.44	-0.30	0.57	0.58	0.25		
boat	-0.13	-0.33	-0.59	0.00	0.73		
ocean	-0.48	-0.51	-0.37	0.00	-0.61		×
wood	-0.70	0.35	0.15	-0.58	0.16		
tree	-0.26	0.65	-0.41	0.58	-0.09		
Σ	1	2	3	4	5		
1	2.16	0.00	0.00	0.00	0.00		
2	0.00	1.59	0.00	0.00	0.00		
3	0.00	0.00	1.28	0.00	0.00		×
4	0.00	0.00	0.00	1.00	0.00		
5	0.00	0.00	0.00	0.00	0.39		
V^T	d_1	d_2	d_3	d_4	d_5	d_6	
1	-0.75	-0.28	-0.20	-0.45	-0.33	-0.12	
2	-0.29	-0.53	-0.19	0.63	0.22	0.41	
3	0.28	-0.75	0.45	-0.20	0.12	-0.33	
4	0.00	0.00	0.58	0.00	-0.58	0.58	
5	-0.53	0.29	0.63	0.19	0.41	-0.22	

将空间维度降为 2

U	1	2	3	4	5	
ship	-0.44	-0.30	0.00	0.00	0.00	
boat	-0.13	-0.33	0.00	0.00	0.00	
ocean	-0.48	-0.51	0.00	0.00	0.00	
wood	-0.70	0.35	0.00	0.00	0.00	
tree	-0.26	0.65	0.00	0.00	0.00	
Σ_2	1	2	3	4	5	
1	2.16	0.00	0.00	0.00	0.00	
2	0.00	1.59	0.00	0.00	0.00	
3	0.00	0.00	0.00	0.00	0.00	
4	0.00	0.00	0.00	0.00	0.00	
5	0.00	0.00	0.00	0.00	0.00	
V^T	d_1	d_2	d_3	d_4	d_5	d_6
1	-0.75	-0.28	-0.20	-0.45	-0.33	-0.12
2	-0.29	-0.53	-0.19	0.63	0.22	0.41
3	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00

实际上，我们只需将矩阵 Σ 中相应的维度置为0即可。此时，相当于矩阵 U 和 V^T 的相应维度被忽略，然后计算 $C_2 = U\Sigma_2 V^T$.

为什么新的低维空间更好？

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1
C_2	d_1	d_2	d_3	d_4	d_5	d_6
ship	0.85	0.52	0.28	0.13	0.21	-0.08
boat	0.36	0.36	0.16	-0.20	-0.02	-0.18
ocean	1.01	0.72	0.36	-0.04	0.16	-0.21
wood	0.97	0.12	0.20	1.03	0.62	0.41
tree	0.12	-0.39	-0.08	0.90	0.41	0.49

在原始空间中， d_2 和 d_3 的相似度为0；
 但是在新空间下， d_2 和 d_3 的相似度为：
 $0.52 * 0.28 + 0.36 * 0.16 + 0.72 * 0.36 + 0.12 * 0.20 + -0.39 * -0.08 \approx 0.52$

LSI实现

- 对词项-文档矩阵进行SVD分解
- 计算在新的低维空间下的文档表示
- 将查询映射到低维空间中 $\vec{q}_2^T = \Sigma_2^{-1} U_2^T \vec{q}$
- 上述公式来自： $C_2 = U_2 \Sigma_2 V_2^T \Rightarrow V_2^T = \Sigma_2^{-1} U_2^T C_2$
- 计算 q_2 和 V_2 中的所有文档表示的相似度
- 像以往一样按照相似度高低输出文档结果
- LSI可以看成是向量空间模型的降维方法

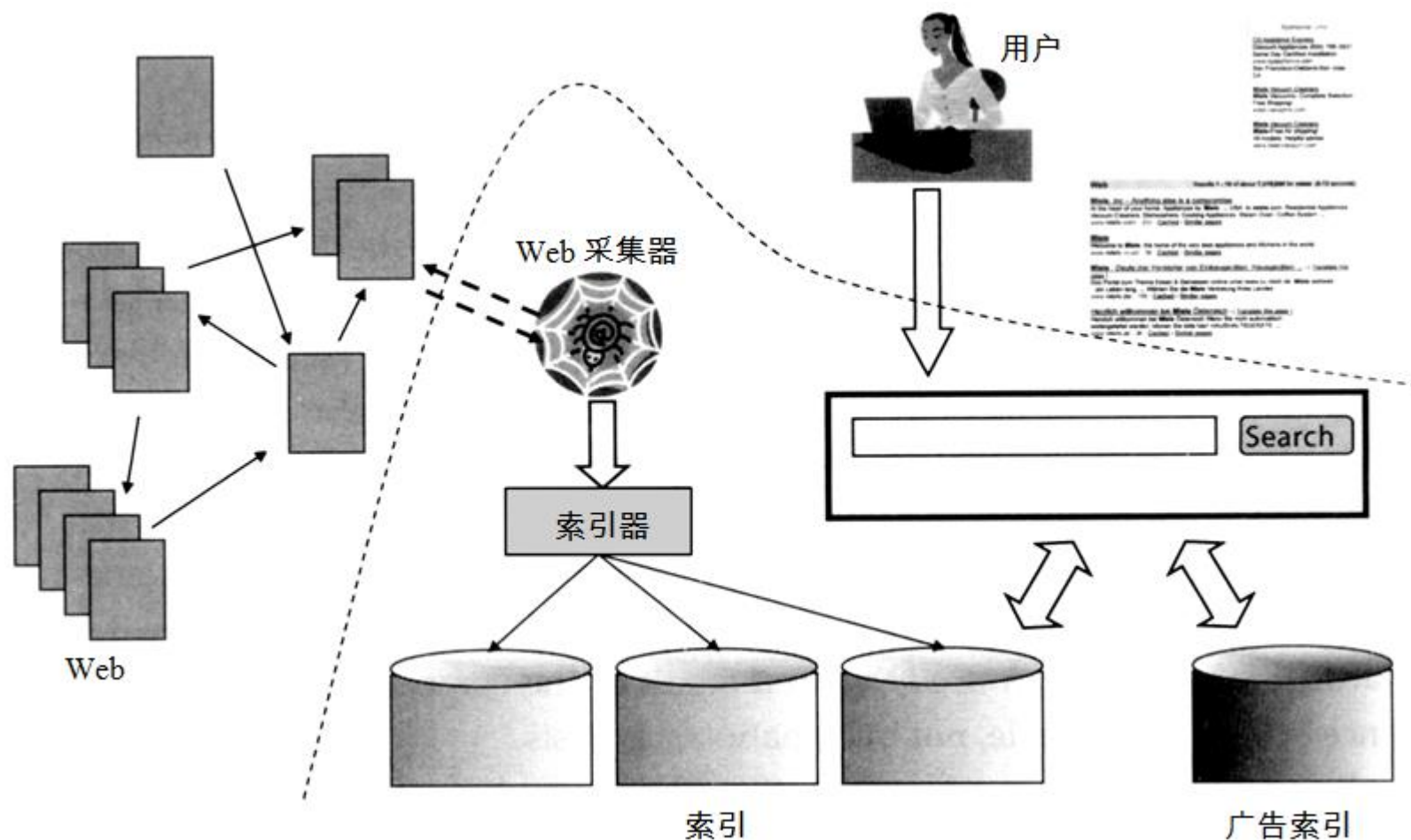
课程回顾

- ① 1~8章，构建搜索引擎的核心理论：倒排索引
- ② 第9章，相关反馈和查询扩展，增加返回结果的相关性
- ③ 第11~12章，基于概率论的信息检索模型
 - a) 传统概率检索模型
 - b) 语言模型
- ④ 第13~18章，机器学习和数值方法
 - a) 13~15章，文档分类
 - b) 16~18章，文档聚类
- ⑤ **Web搜索**，信息检索的具体应用
 - a) Web搜索面临的基本挑战和普遍采用的技术
 - b) Web数据采集技术
 - c) Web搜索中的链接分析技术

提纲

- ① 上一讲回顾
- ② Web搜索系统
- ③ Web信息检索的特点
- ④ 互联网广告
- ⑤ 重复检测

Web搜索系统组成



搜索引擎简史回顾

1986年，
Internet
正式形成

Archie

1990年，加拿大蒙特利尔McGill大学学生Alan Emtage发明是对FTP文件名搜索，现代搜索引擎的祖先

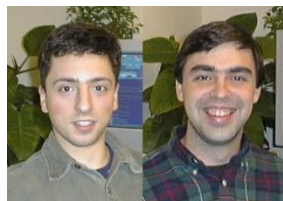


Lycos

1994年7月，CMU机器翻译中心的Michael Mauldin创建，第一个现代意义上的WEB搜索引擎

Google

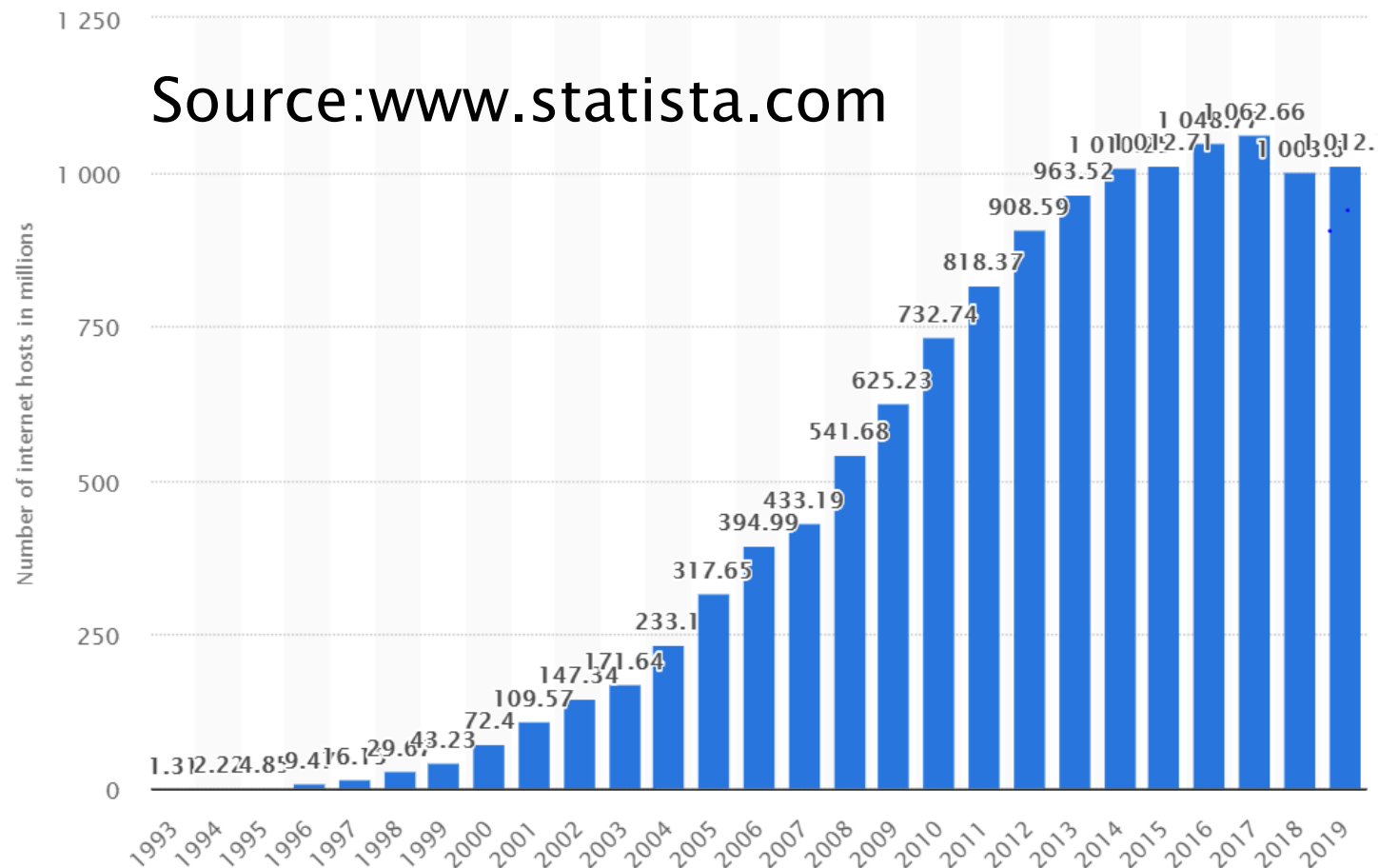
1998年9月，斯坦福大学博士生Sergey Brin与Larry Page创建，目前是世界最受欢迎的搜索引擎。



Baidu

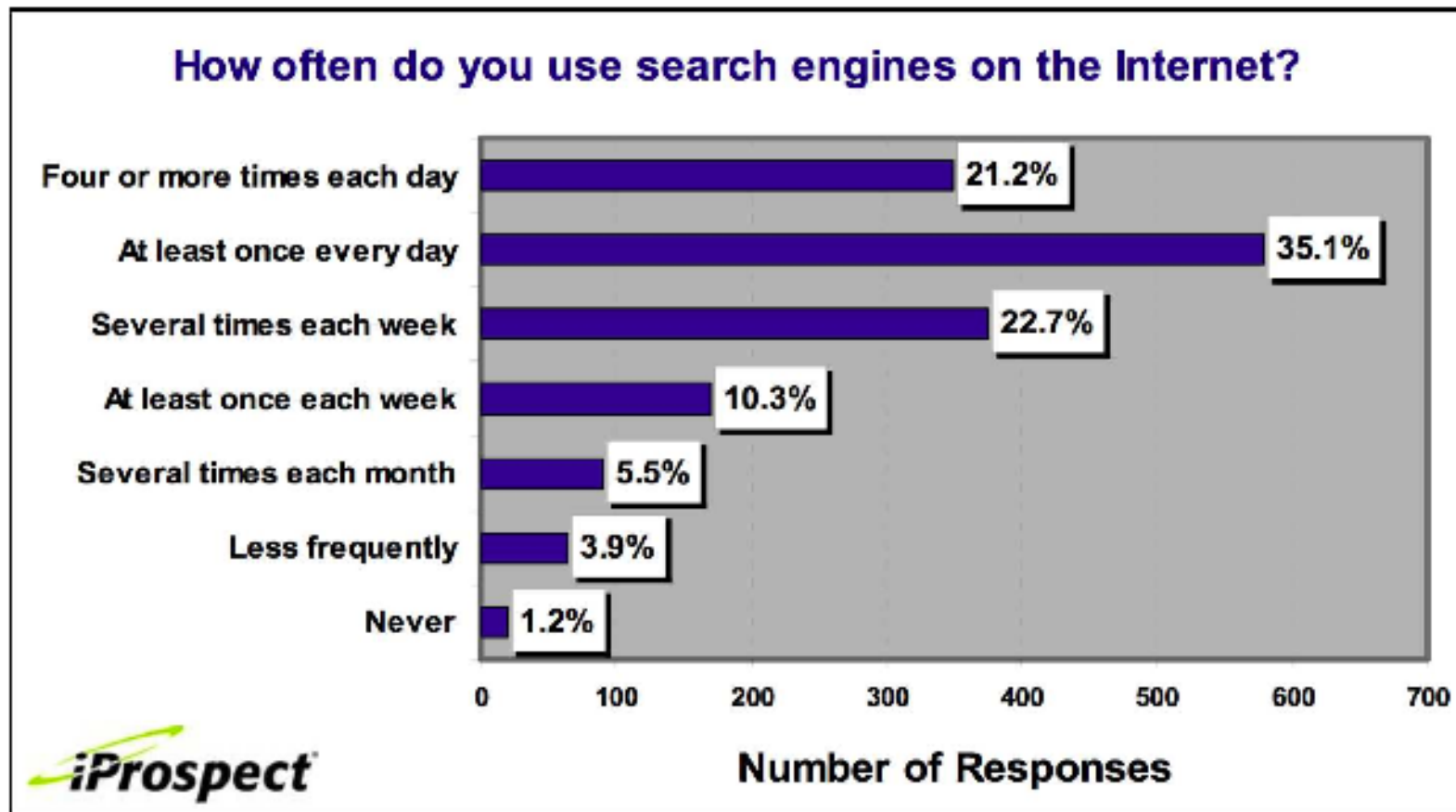
2001年，李彦宏与徐勇在北京中关村创立了百度公司，10月22日发布百度搜索引擎，是目前最受欢迎的中文搜索引擎之一

Web的增长速度



- DNS主机规模已达10亿级

搜索是Web上使用最多的应用之一



没有搜索引擎，Web甚至无法运转

- 没有搜索，很难找到所需的内容
- 没有搜索，在Web上创建内容也就缺了动机
 - 如果没人看为什么要发布内容？
 - 如果没有任何回报为什么要发布内容？
- Web运转必须要有人买单
 - 服务器、Web基础设施、内容创建过程等需要费用支持
 - 这些费用的大部分都是通过搜索广告支付
 - 可以说，搜索为Web买单

兴趣聚合(Interest aggregation)

- Web的特点：具有相同兴趣的人，即使所处地理位置分散，也可以通过Web找到对方
 - 喜欢极限运动的人们
 - 研究computer vision的学者 (开源项目和社区)
- 搜索引擎是实现兴趣聚合的关键事物

Web IR vs. 一般的IR

- 在Web上，搜索不仅仅是一个优异的功能
 - 搜索是Web的关键组成部分: ...
 - ... 筹资、内容创建、兴趣聚合等等 → 参考搜索广告
- Web是一个充满噪声数据且组织失调的集合体 → 大量的重复需要检测
- 用户可以无控制和无限制地发布内容 → 大量作弊内容需要检测
- Web规模非常大 → 需要知道Web的规模大小

提纲

① 上一讲回顾

② Web搜索系统

③ Web信息检索的特点 

- 查询请求
- 上下文
- 用户
- 文档
- 链接
- 大小

④ 互联网广告

⑤ 重复检测

提纲

① 上一讲回顾

② Web搜索系统

③ Web信息检索的特点 

- 查询请求
- 上下文
- 用户
- 文档
- 链接
- 大小

④ 互联网广告

⑤ 重复检测

查询请求

2020年12月5日，百度、搜狗、360的实时热搜榜

百度热搜榜

排名	热点	搜索指数
1	重庆永川煤矿事故已致18人遇难	4827303
2	美国将对更多中国公民实施签证限制	4658348
3	特朗普在佐治亚州提起选举诉讼	4495305
4	美国叫停5个中方资助的交流项目	4337970
5	31省份新增本土确诊2例 在内蒙古	4186141
6	明日将有火箭残骸坠落云南	3898239
7	湖南干部猥亵女企业家获刑三年	3761800
8	丁真得到一匹名叫青龙的赛马	3630137

实时热点

更多>>

1	新冠疫苗重磅消息	9542 ↑
2	火箭残骸将坠云南	9038 ↑
3	史上第二热年份	8664 ↑
4	人造太阳首次放电	8119 ↓
5	西藏首例克隆藏猪	7631 ↓
6	重庆永川煤矿事故	7309 ↑
7	内蒙古现三日当头	6804 ↓
8	澳回应假肢当酒杯	6260 ↑
9	美医院黑暗震撼照	5788 ↓
10	南大贫困生炫富	5527 ↓

360热搜榜

排名	热点	搜索指数
1	索嫖案犯人下周出狱	97971
2	央视曝苹果安全漏洞	97922
3	警察目视女孩溺亡	97854
4	大爷狂偷政府机关	97850
5	重庆煤矿事故致18死	97730
6	司机拒超载遭围殴	97702
7	谈莉娜被记者问哭	97541
8	内蒙古出现三日当头	97511

查询请求的分布

- 查询请求符合幂次分布 (power law distribution)
- Zipf 定律：一个单词出现的次数与它在频率表中的排名成反比
- 类似的：高频查询请求的数量少，低频查询请求的数量巨大
- 低频查询的例子：搜索人姓、地名、书名等。

查询请求的类型 / 用户需求分类

- **信息类**查询：奇异值分解、PageRank算法
- **导航类**查询：“中国工商银行”、“中科院信息工程研究所”的主页
- **事务类**查询：用户进行事务处理前的先导型查询，比如
 - 购买车票：“北京到上海”
 - 预定旅店：“斐济 香格里拉酒店”
 - 软件下载：“Foxmail”
- 难点：搜索引擎如何知道用户输入一个特定查询请求的**真实需求或目的**？

提纲

① 上一讲回顾

② Web搜索系统

③ Web信息检索的特点 

- 查询请求
- 上下文
- 用户
- 文档
- 链接
- 大小

④ 互联网广告

⑤ 重复检测

用户意图

- 如何获取用户意图？
- 上下文无关的方式：
 - 拼写校正
 - 预先计算用户查询的类型（见下页）
- 更好的方式：基于上下文预测用户意图：
 - 地理信息上下文（见后页）
 - 当前会话的上下文（例如，用户之前的查询请求）
 - 用户个人信息包含的上下文（Yahoo/MSN/Google已参考用户个人信息）

对查询请求进行分类

算数计算: $3/7*2/5+4/5*2/7$

单位转换: 1 kg 磅

货币转换: 1美元 人民币

快递号: 8167 2278 6764

航班动态: CA1549

区号: 010

地图: 国科大雁栖湖校区

股价: msft

专辑/ 电影等: 流浪地球

3/7*2/5+4/5*2/7

Q 网页 资讯 视频 图片 知道 文库 贴吧 地图

百度为您找到相关结果约100,000,000个

搜索工具

3/7*2/5+4/5*2/7			
0.4			
()	%	C
7	8	9	÷
4	5	6	x
1	2	3	-
0	.	=	+

1美元 人民币

Q 网页 资讯 视频 图片 知道 文库 贴吧

百度为您找到相关结果约100,000,000个


货币兑换

1美元=6.5301人民币
1人民币≈0.1531美元

数据仅供参考, 交易时以银行柜台成交价为准 更新时间:2020-12-05 16:31
[更多汇率信息>>](#)

北京到上海

Q 网页 资讯 视频 图片 知道 文库 贴吧 地图

百度为您找到相关结果约100,000,000个

搜索工具

搜索结果涉及价格仅作参考, 请以商家官网为准

[北京到上海机票](#) [特价机票查询预订](#) [携程旅行网](#)

12-06	12-07	12-08	
18:20 大兴国际	→	20:55 浦东T1	¥554起 3.4折 联航KN5977
08:20 大兴国际	→	10:45 浦东T2	¥299起 1.9折 吉祥HO1254
07:25 大兴国际	→	09:35 虹桥T2	¥387起 2.4折 联航KN5737

地理信息上下文: Geo-search

- **3个相关的位置**
 - Web服务器位置 (nytimes.com → New York)
 - Web页面内容中的位置 (nytimes.com关于Hong Kong的报道)
 - 用户位置(北京、上海)
- **如何确定用户位置?**
 - IP地址
 - 用户提供的信息 (例如, 用户注册信息)
 - 移动电话
- **地理位置标记:** 分析文本, 确定地理实体的坐标
 - 例如: East Palo Alto CA → Latitude: 37.47 N, Longitude: 122.14 W
 - 这是一个重要的NLP问题

如何利用上下文改变查询结果？

- 结果限定：不考虑不合适的结果
 - 对于使用google.fr...的用户
 - 返回来自.fr...域名的结果
- 调整打分：使用粗略的通用排名算法，再根据个人背景重新排名
- 情境化/个性化（Contextualization/ personalization）搜索是Web搜索的一个研究领域，具有很大的改进空间

提纲

① 上一讲回顾

② Web搜索系统

③ Web信息检索的特点 

- 查询请求
- 上下文
- 用户
- 文档
- 链接
- 大小

④ 互联网广告

⑤ 重复检测

互联网用户规模



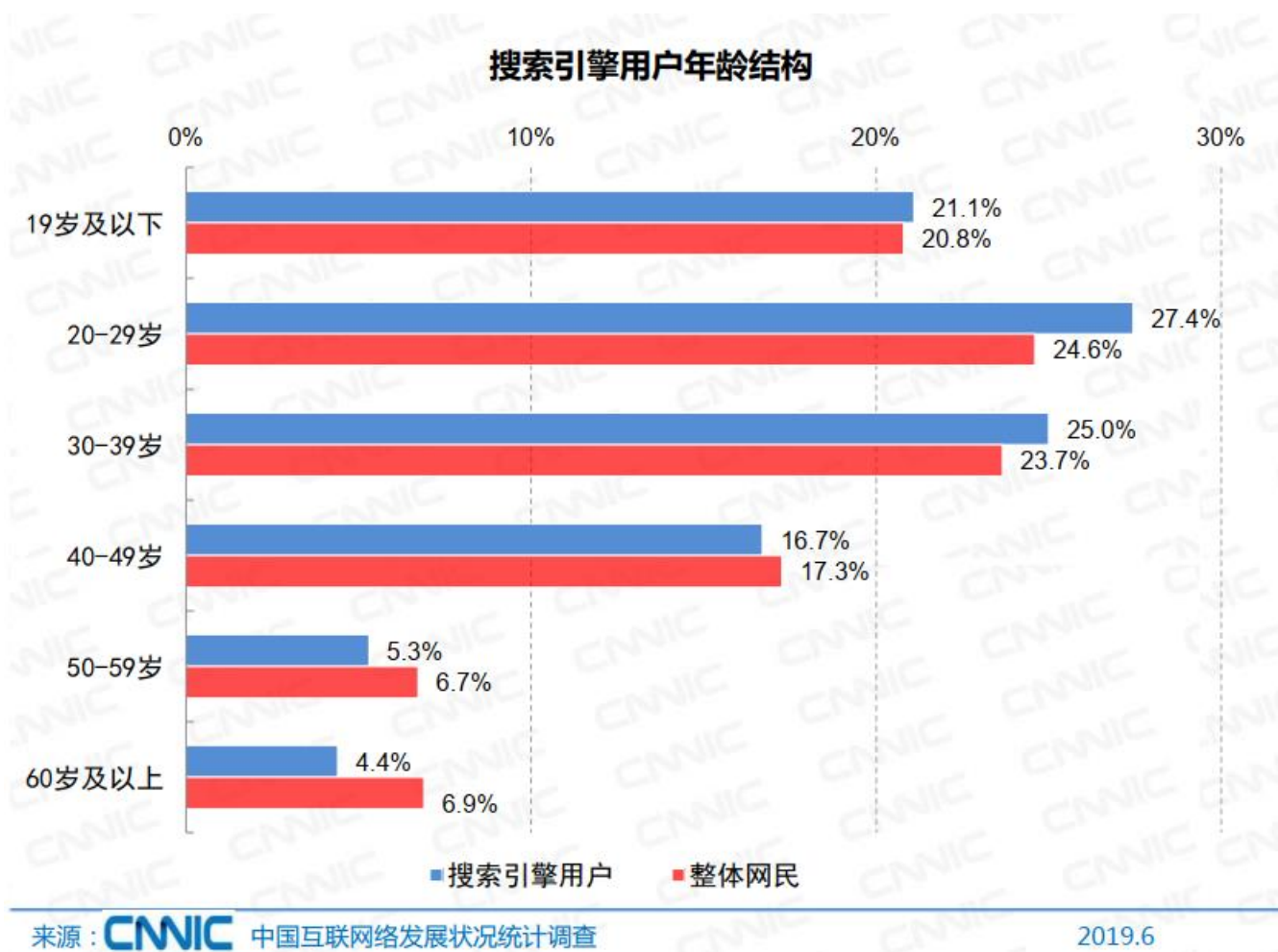
截至2019年互联网用户为43.9亿, 搜索引擎用户规模大于30亿

国内搜索引擎用户规模及使用率



截至2020年6月，我国网民规模为9.4亿，搜索引擎用户达7.66亿

搜索引擎用户年龄结构



Web搜索的用户特点

- 查询请求短（平均长度小于3个词）
- 基本不使用操作符
- 不想花时间拼装查询请求
- 仅看最前面的部分结果
- 需要简洁的UI，不希望开始页面嵌入图形
- 用户需求、用户期望、经历、知识水平参差不齐
 - 工业界/学术界、英语/爱沙尼亚语、老人/年轻人、富人/穷人
 - 在文化和社会层级上都存在差异
- 一个接口，满足巨大的、千差万别的需求

用户如何评价搜索引擎？

- 典型的信息检索系统相关性指标（F），同样适用于Web搜索
- 其它重要的因素：可信度，消冗，可读性，快速加载，没有弹窗
- 在Web搜索中，精确率比召回率更重要
 - Top 1的精确率、top 10的精确率、前2-3页的精确率
 - 但部分查询需要关注召回率

需要高召回率的Web查询需求

- 某个想法是否已注册专利
- 搜索潜在财务顾问的信息
- 搜索潜在员工的信息
- 搜索某个日期的信息

提纲

① 上一讲回顾

② Web搜索系统

③ Web信息检索的特点 

- 查询请求
- 上下文
- 用户
- 文档
- 链接
- 大小

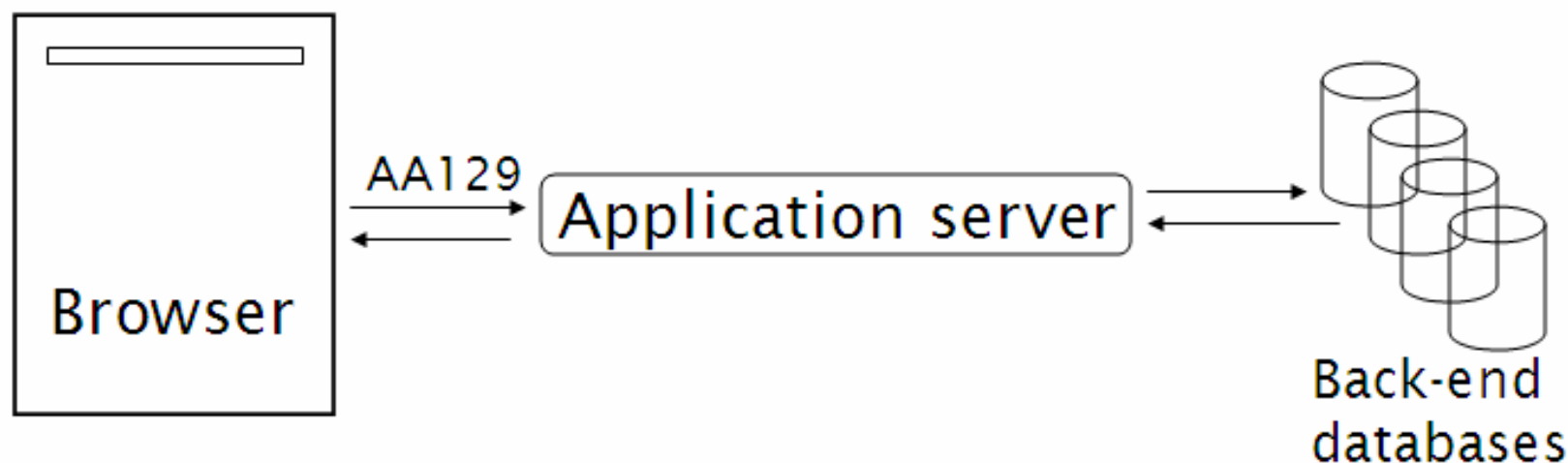
④ 互联网广告

⑤ 重复检测

Web文档：与其它IR数据集的区别

- 去中心化的内容创建：没有设计、没有协同
 - “信息发布民主化”
 - 结果： Web文档极端的异构性
- 无结构（text、html）、半结构（html、xml）、结构化/关系型(databases)
- 动态生成内容

动态生成内容 (1)



- 动态页面是在用户发出请求时重新生成的——通常来自于底层的数据库
- 例如：查询航班AA129的当前状态

动态生成内容 (2)

- 大多数动态生成网页都被Web爬虫忽略
 - 要索引所有动态网页代价太大
- 事实上，很多“静态”网页也是实时生成的
 - 如，asp、php网页中的标题、日期、广告等

多语种

- 网页文档包含大量不同语言
- 用户查询请求包含大量不同语言
- 简单处理：搜索日文，不返回英语结果
- 然而：查询请求和文档中经常会出现多个语种
- 许多人可以理解，但不能用一种语言查询
- 需要进行翻译
- Google示例：“Beaujolais Nouveau -wine”

重复文档

- 大量重复文档：30%~40%重复率
- 早期Web搜索中，在检索结果中通常会包含大量重复结果
- 现在搜索引擎有效的消除了重复文档
- 这对提高用户满意度非常重要

可信度

- 很多文档集，很容易评估文档的可信度
 - 路透社通讯社文集
 - 20世纪80年代的塔斯社（苏联电报局）新闻专线文集
 - 个人过去三年的Outlook电子邮件
- Web文档是不同的：在许多情况下，我们不知道如何评估信息是否真实
- 恶作剧/谣言比比皆是

提纲

① 上一讲回顾

② Web搜索系统

③ Web信息检索的特点 

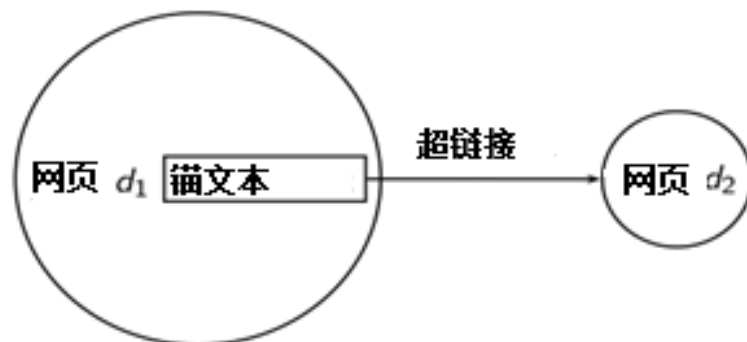
- 查询请求
- 上下文
- 用户
- 文档
- 链接
- 大小

⑤ 互联网广告

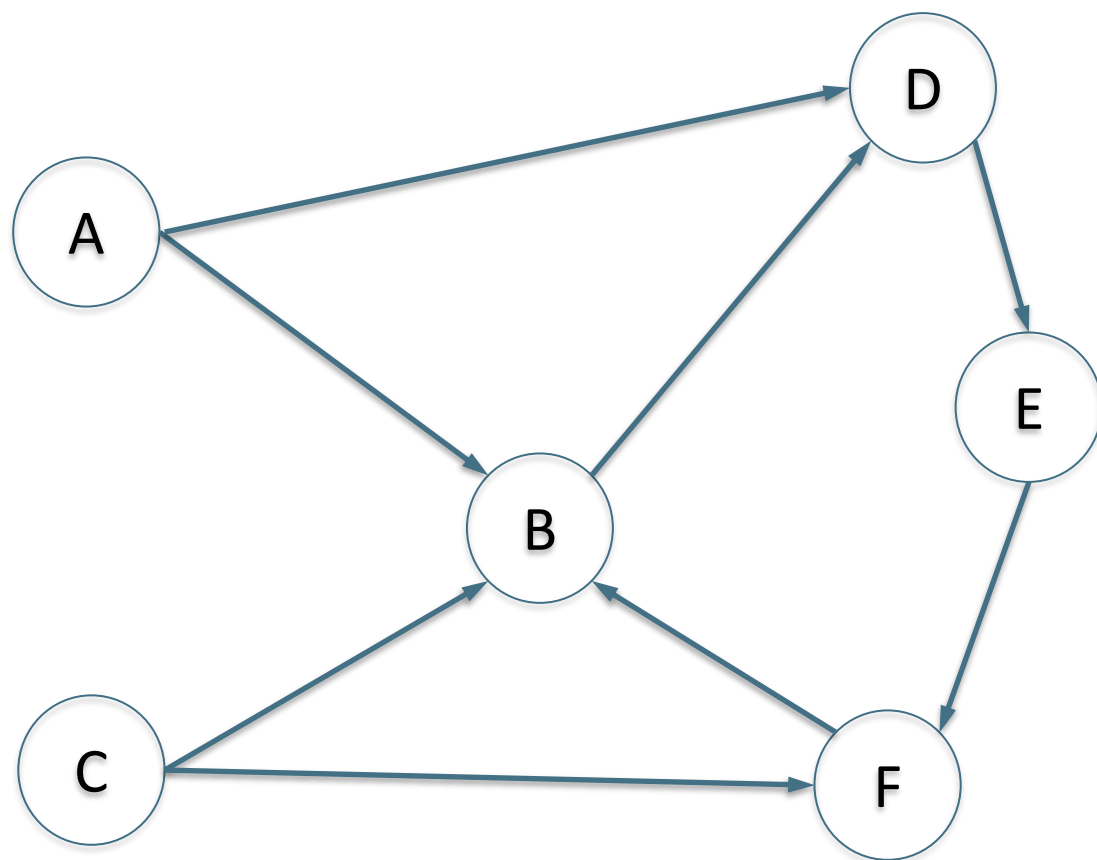
⑥ 重复检测

在一个超链接组成的集合中查询

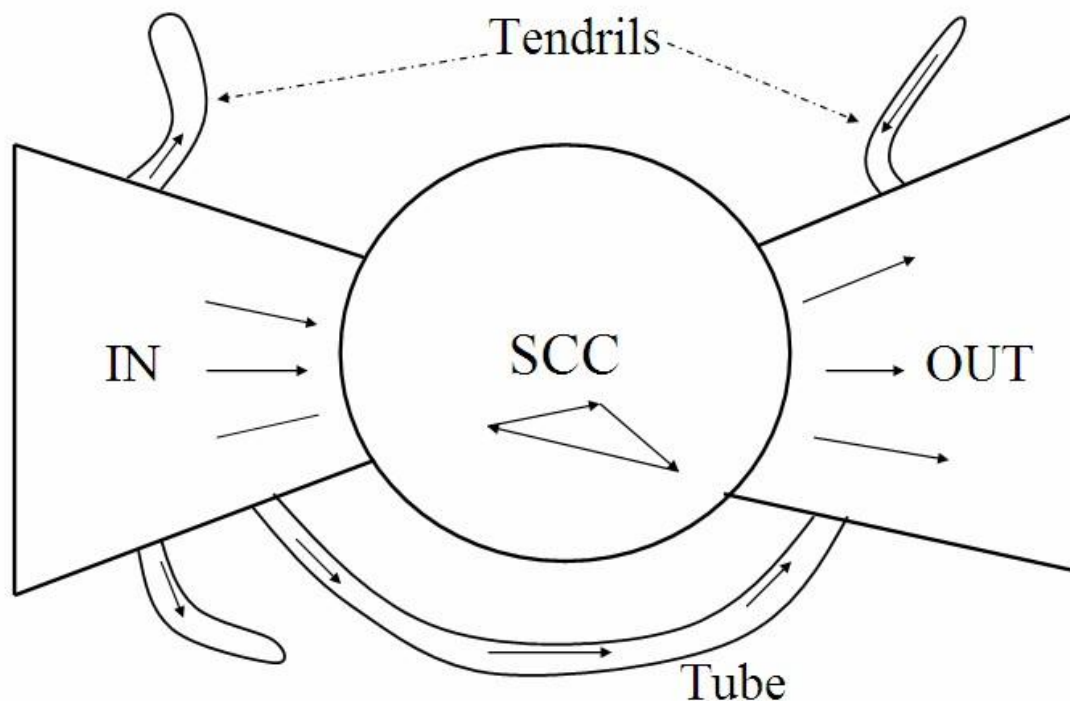
- Web搜索的结果通常都包含导航信息
- 即，Web搜索结果中通常都有超链接
- 这与大多数IR数据集不同



一个小型Web图的例子



Web图的蝴蝶结形结构



Strongly connected component (**SCC**) in the center

Lots of pages that get linked to, but don't link (**OUT**)

Lots of pages that link to other pages, but don't get linked to (**IN**)

Tendrils(卷须), **tubes**(管道), **islands**(孤岛)

提纲

① 上一讲回顾

② Web搜索系统

③ Web信息检索的特点 

- 查询请求
- 上下文
- 用户
- 文档
- 链接
- 大小

④ 互联网广告

⑤ 重复检测

Web的大小：问题

- **Web大小如何定义？**
 - 是Web服务器数量？
 - 还是网页数量？
 - 还是存储了多少TB的数据？
- **有的服务器未与Web联通**
 - 例如：个人笔记本运行Web服务
 - 它们是否是Web的一部分？
- **动态网页数量无穷大**
 - 任意两个数的和，在搜索引擎中都是一个独立的动态页面（例如：“ $2+4$ ”）
 - 动态网站生成无数的动态网页

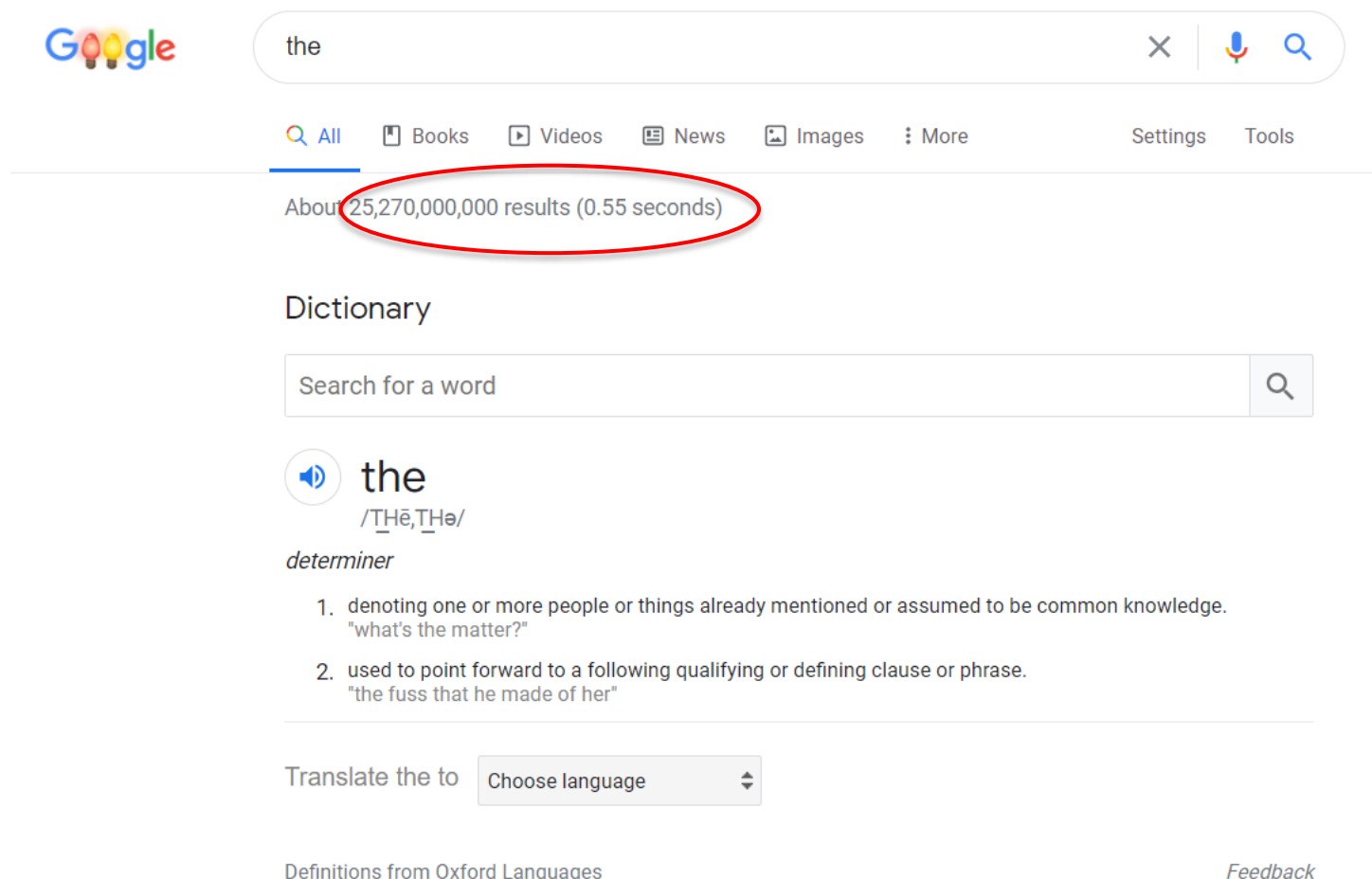
“搜索引擎索引的网页有多少”：问题

- 如果仅对前4K字节建立索引，能否认为这个网页包含在索引中？
- 如果仅对指向一个网页的锚文本建立索引，能否认为这个网页包含在索引中？
 - 仍有数十亿的页面仅通过锚文本建立索引

确定Web规模下限的简单方法

- 使用不同语言，对常用词进行OR组合查询
- 基于该方法估计：
 - 2007. 07. 07, Web大小 \geq 21, 450, 000, 000
 - 2008. 07. 03, Web大小 \geq 25, 350, 000, 000
- 缺点：Google搜索结果中的网页数量不是精确值

确定Web规模下限的简单方法



The screenshot shows a Google search interface. The search bar contains the word "the". Below the search bar, the results are displayed as "About 25,270,000,000 results (0.55 seconds)". This result count is circled in red. Below the search bar, there are links for "All", "Books", "Videos", "News", "Images", and "More". To the right of these links are "Settings" and "Tools". Below the search bar, there is a "Dictionary" section with a search input field and a magnifying glass icon. Below the dictionary section, the word "the" is displayed with its phonetic transcription "/THē, THə/" and the word "determiner". Below this, there are two numbered definitions: 1. denoting one or more people or things already mentioned or assumed to be common knowledge. "what's the matter?" 2. used to point forward to a following qualifying or defining clause or phrase. "the fuss that he made of her". Below the definitions, there is a "Translate the to" section with a "Choose language" dropdown menu. At the bottom of the page, there is a footer that reads "Definitions from Oxford Languages" and a "Feedback" link.

Google


the

[All](#)
[Books](#)
[Videos](#)
[News](#)
[Images](#)
[More](#)
[Settings](#)
[Tools](#)

About 25,270,000,000 results (0.55 seconds)

Dictionary

Search for a word


the
 /THē, THə/

determiner

1. denoting one or more people or things already mentioned or assumed to be common knowledge.
"what's the matter?"
2. used to point forward to a following qualifying or defining clause or phrase.
"the fuss that he made of her"

Translate the to

Definitions from Oxford Languages [Feedback](#)

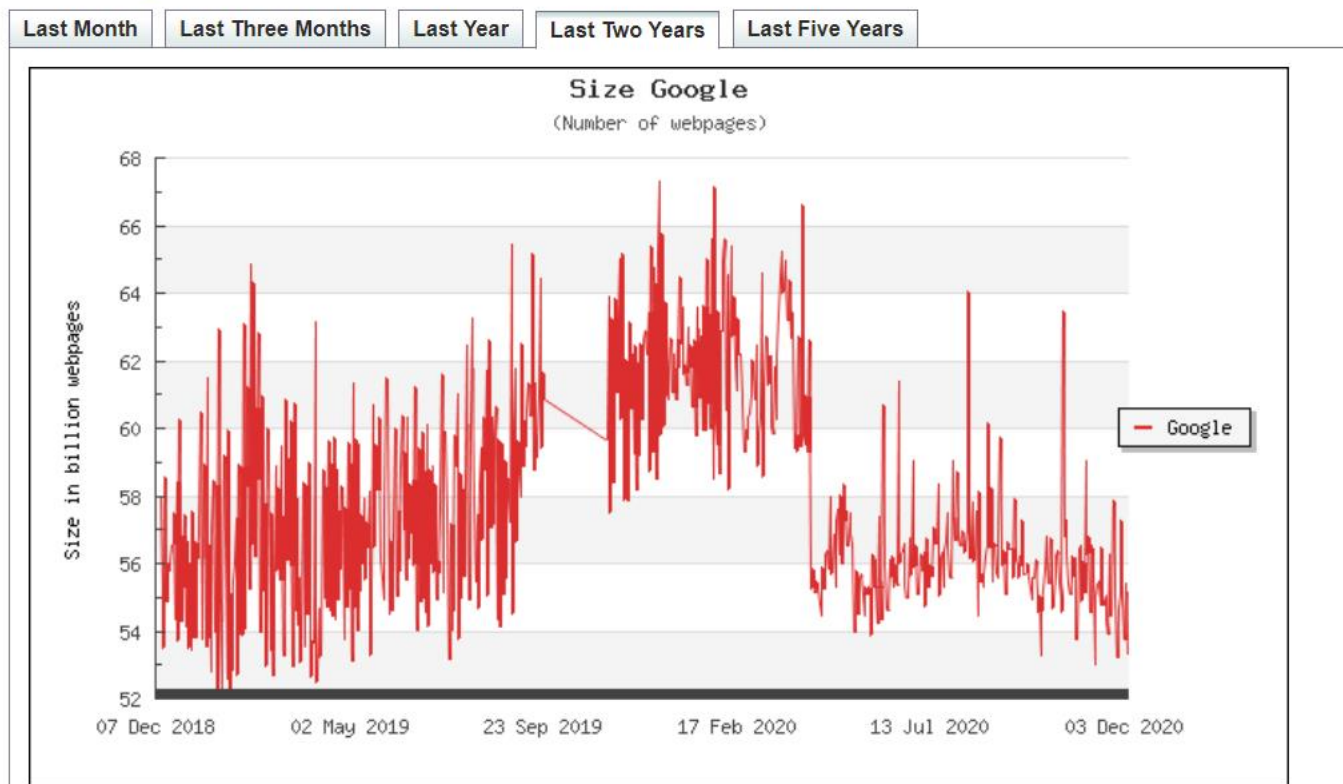
确定Web规模下限的简单方法

- The size of the World Wide Web (The Internet)

<https://www.worldwidewebsite.com/>, 荷兰蒂尔堡大学



The size of the World Wide Web:
Estimated size of Google's index



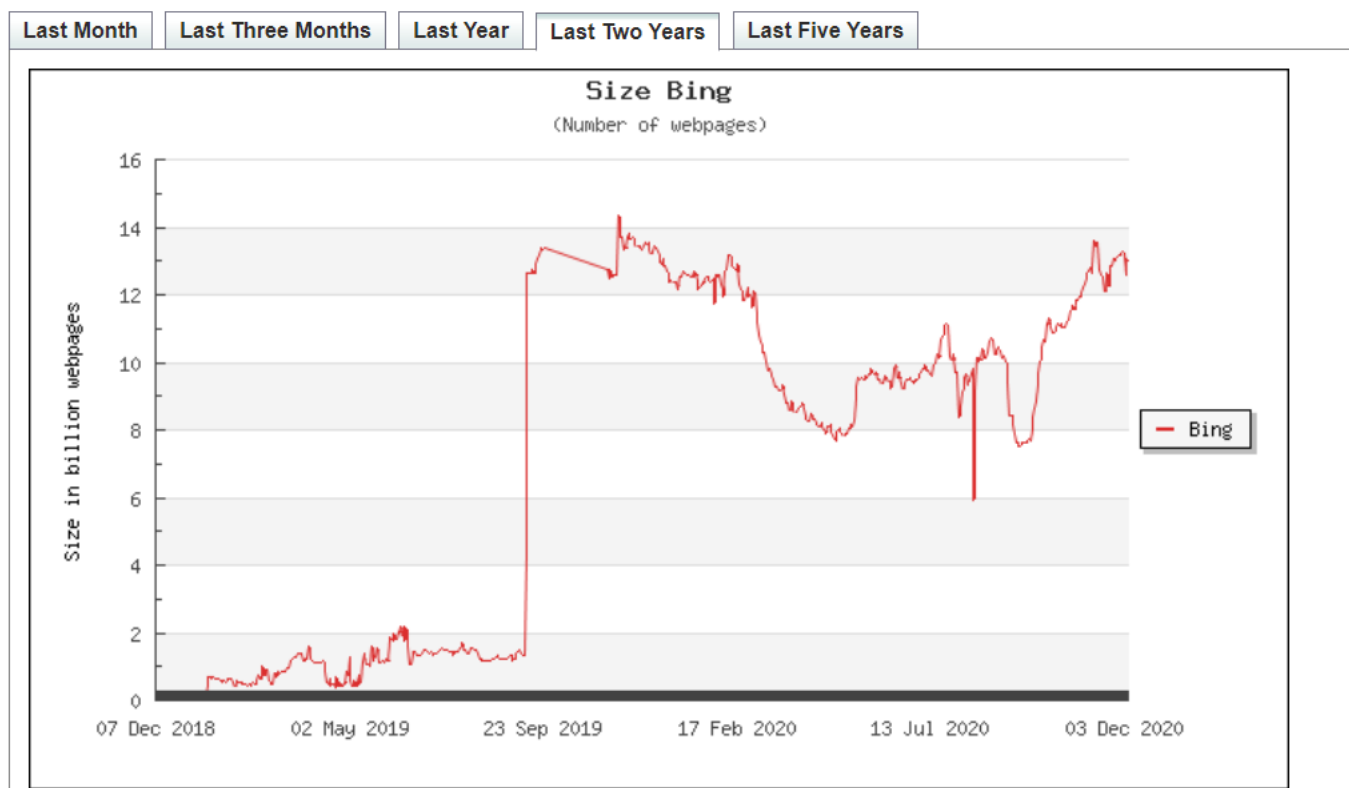
确定Web规模下限的简单方法

- The size of the World Wide Web (The Internet)

<https://www.worldwidewebsite.com/>, 荷兰蒂尔堡大学



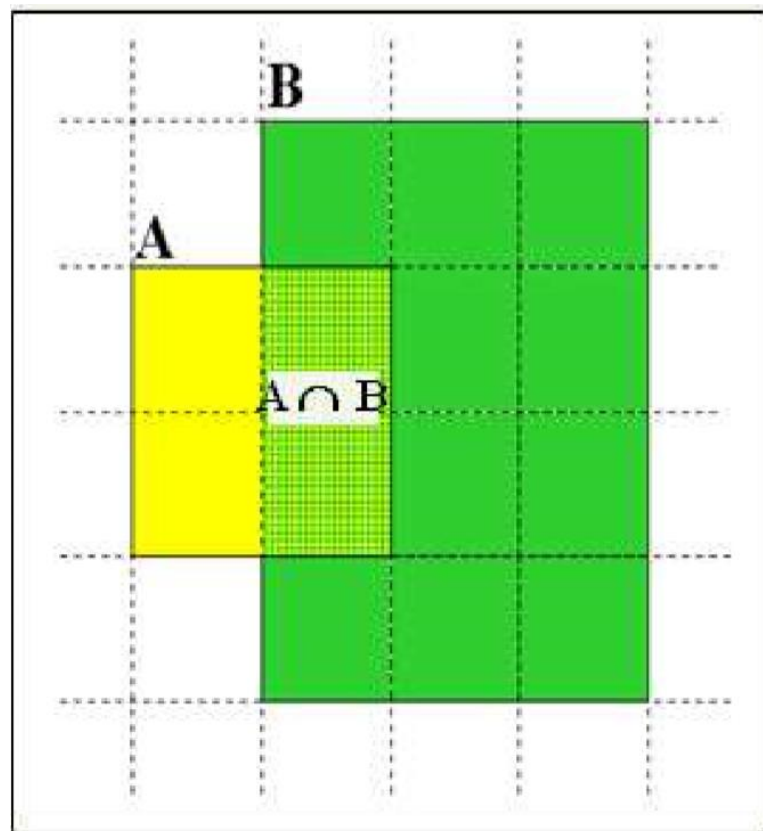
The size of the World Wide Web:
Estimated size of Bing index



问题变种: 估算索引的相对大小

- 不同搜索引擎的索引有很大差异
- 不同搜索引擎有不同的爬取偏好
 - 最大URL深度
 - 最大数量
 - 最大主机数
 - 反作弊规则
 - 优先级规则
- 对相同URL的索引内容不同
 - 锚文本、整个网页、meta标签中的keywords、前N个词

基于重叠区域的相对大小估算[Bharat & Broder, 98]



- (1) 从A中随机抽样一个URL集,
- (2) 检查在B中是否包含这些URL
- (3) 反之亦然

$$A \cap B = (1/2) * \text{Size } A$$

$$A \cap B = (1/6) * \text{Size } B$$

$$(1/2) * \text{Size } A = (1/6) * \text{Size } B$$

$$\therefore \text{Size } A / \text{Size } B =$$

$$(1/6) / (1/2) = 1/3$$

每次估算包括两个步骤: (i) 抽样 (ii) 检查

抽样URL

- 理想的策略：生成随机的URL集合
- 问题：难以找到随机的URL集合(抽样的分布应体现出“用户兴趣”)
- 方法1：随机游走/IP地址
 - 理论上：可以得到Web的真正近似大小（而不仅仅是索引的相对大小）
- 方法2：生成一个包含在给定搜索引擎内的随机URL
 - 可满足准确估算相对大小的需求

基于随机查询生成随机URL方法

- 思想：基于Web词典生成查询请求
- 词典可从Web爬取的数据中生成
- 使用合取查询 w_1 AND w_2
 - 例如：“歌手” AND “指标”
- 从源引擎中得到100个URL结果集
- 从结果集中随机选择一个URL
- 该方法由Bharat and Broder提出(1998)

检查页面是否包含在索引中

- 如果引擎支持URL检索，使用页面的URL进行查询
- 或者：构造一个能大概率找到页面 d 的查询
 - 下载网页，抽取词语
 - 使用8个低频词语，进行合取查询
 - 查询被称为 d 的强查询
 - 运行查询
 - 检查 d 是否在结果集中
- 存在的问题
 - 内容近似重复的网页影响结果
 - 引擎查询超时

随机查询方法的优劣

- 在统计上，使用了较好的抽样技术
- 随机查询引起的偏差
 - 查询偏差：倾向于选择词典语言中内容丰富的页面
 - 排序偏差：使用连接查询和全部获取
 - 检查偏差：忽略动态、低质网页
 - 文档或查询限制偏差：引擎可能无法处理8个单词的联合查询
 - 恶意偏差：引擎破坏
 - 操作问题：超时、故障、引擎不一致、索引修改

常用抽样方法

- 随机查询法（ Lawrence & Giles 97 ）
- 随机搜索法（ Lawr98, Lawr99 ）
- 随机IP地址法（ Lawrence & Giles 99, ONei97, Lawr99 ）
- 随机游走法（Henzinger2000）

提纲

- ① 上一讲回顾
- ② Web搜索系统
- ③ Web信息检索的特点
- ④ 互联网广告
- ⑤ 重复检测

传统广告(1)

- 品牌广告(Brand Advertising)



TIFFANY & CO.
Tiffany T

传统广告(2)

■ 直接营销(Direct marketing)

网上订餐 专享优惠

天天半价

• 每款特价半价的食品原品只售1份
• 内含珍珠奶茶，请小心吸食，儿童请在家长监护下食用。
• 活动时间：2013年6月3日-2013年7月21日

周一	周二
海鲜至尊比萨 单品价58元 半价29元 大 单品价98元 半价45元	红豆珍珠冻奶茶 单品价10元 半价5元
夏威夷风光芝士比萨 单品价62元 半价30.5元	摩卡咖啡冰沙 单品价14元 半价7元
奥尔良风味烤肉芝士比萨 单品价73元 半价35.5元	特浓冻奶茶 单品价9元 半价4.5元
秘制飘香烤肉芝士比萨 单品价61元 半价30.5元	红豆珍珠冻奶茶 单品价10元 半价5元
巧克力熔岩珍珠奶茶 单品价10元 半价5元	特浓冻奶茶 单品价9元 半价4.5元
黑椒小牛排比萨 单品价58元 半价29元 大 单品价98元 半价45元	雪梨烤鸭米线 单品价24元 半价12元
蜜汁小牛排比萨 单品价58元 半价29元 大 单品价98元 半价45元	秘制江肉饭 单品价29元 半价14.5元

周五
巧克力熔岩珍珠奶茶 单品价10元 半价5元

周六
特浓冻奶茶 单品价9元 半价4.5元

周日
雪梨烤鸭米线 单品价24元 半价12元

温馨提示：
产品和包装以实物为准，产品均加强检测，请放心食用。
为保证产品质量，我们在接到您的订单后点餐，餐厅有送餐范围限制，超出送餐范围无法送餐或需另加运费，敬请谅解。
订餐时间：10:00-22:00
不送外带，每单收取一次运费，谢绝小单。
产品、价格以及外送范围如有调整恕不另行通知，具体以订餐时实际确认为准。
菜单中个别菜品因城市差异会有所不同，具体请查询www.必胜宅急送.com，敬请谅解！
促销活动不适用于会员储值卡。
促销活动不适用于会员储值卡。
促销活动不适用于会员储值卡。
如遇不可抗力因素，必胜宅急送将暂时不提供相关服务，敬请谅解！

Note 9 Pro

液冷游戏芯 | 一亿像素夜景相机 | 120Hz 变速高刷屏

¥1599起

[立即购买](#)

购机得2倍米金 | 享碎屏保、延保5折优惠



传统广告的不足

- 广告投放场地或媒介相对有限：报纸、电视、杂志、橱窗、公汽、电梯等
- 广告场地的费用一般不菲：CCTV 标王
- 很难进行个性化
- 投放效果取决于广告商的智慧
- 投放效果很难度量



数据：公开资料，《知识经济》统计整理，2011年、2014年和2015年数据未公开

- 互联网的出现改变了这一切.....

互联网广告的优点

- 无限机会
- 无限创意
- 完全可以个性化处理
- 每次点击花费的代价很低
- 定量度量程度高

互联网广告的主要形式(1)

■ 图片广告

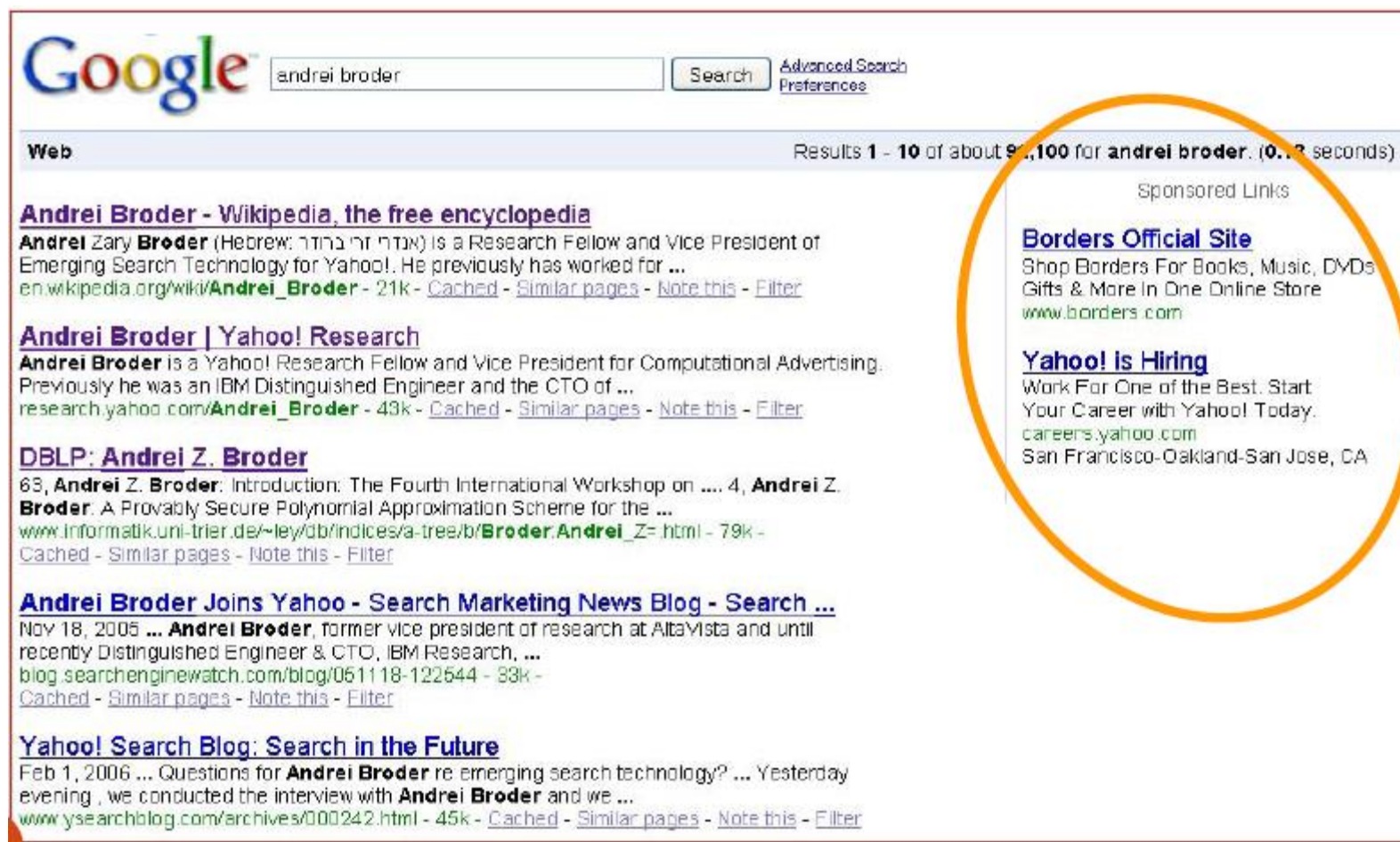
The image shows a screenshot of a CSDN blog page. At the top is the CSDN header with navigation links like '博客', '学院', '下载', etc., and a search bar. The main content area displays a blog post titled 'GNN图神经网络详述-01' by user 'Chris_34'. The post's metadata includes the date '2020-05-27 05:19:25', 618 views, and 1 collection. Below the title are tags for '机器学习', 'GNN', 'NLP', and '图神经网络'. The post text begins with a note about the terminology of 'Graph' and 'Network'. At the bottom of the page are interaction buttons for '点赞', '评论', '分享', '收藏', and '手机看'. A large blue speech bubble with the text '图片广告' (Image Ad) points to a blue PayPal advertisement on the left side of the page. The ad features the PayPal logo and text: '一个账户, 收款全球', '0 费用开户, 享卖家保障', '赢逾 2 亿用户', and a button '启用PayPal收款'.

互联网广告的主要形式(2)

- 文本广告
 - 搜索关键词触发的广告(Sponsored Search driven Ad, paid Ad), 也称搜索广告。Google Adswords
 - 网页内容触发的广告(Web Page driven Ad), 也称上下文广告(Contextual Ad)。Google Adverb

互联网文本广告的主要类型(1)

■ 搜索广告



The screenshot shows a Google search results page for the query "andrei broder". The search bar at the top contains the text "andrei broder" and a "Search" button. To the right of the search bar are links for "Advanced Search" and "Preferences". Below the search bar, the results are categorized under "Web". The main results list includes:

- Andrei Broder - Wikipedia, the free encyclopedia**
Andrei Zary Broder (Hebrew: אנדרזר ברודר) is a Research Fellow and Vice President of Emerging Search Technology for Yahoo!. He previously has worked for ...
en.wikipedia.org/wiki/Andrei_Broder - 21k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)
- Andrei Broder | Yahoo! Research**
Andrei Broder is a Yahoo! Research Fellow and Vice President for Computational Advertising. Previously he was an IBM Distinguished Engineer and the CTO of ...
research.yahoo.com/Andrei_Broder - 43k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)
- DBLP: Andrei Z. Broder**
63, Andrei Z. Broder: Introduction: The Fourth International Workshop on ... 4, Andrei Z. Broder: A Provably Secure Polynomial Approximation Scheme for the ...
www.informatik.uni-trier.de/~ley/db/indices/a-tree/b/Broder.Andrei_Z.html - 79k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)
- Andrei Broder Joins Yahoo - Search Marketing News Blog - Search ...**
Nov 18, 2005 ... Andrei Broder, former vice president of research at AltaVista and until recently Distinguished Engineer & CTO, IBM Research, ...
blog.searchenginewatch.com/blog/061118-122644 - 33k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)
- Yahoo! Search Blog: Search in the Future**
Feb 1, 2006 ... Questions for Andrei Broder re: emerging search technology? ... Yesterday evening, we conducted the interview with Andrei Broder and we ...
www.ysearchblog.com/archives/000242.html - 45k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)

On the right side of the page, there is a section titled "Sponsored Links" which is circled in orange. It contains two advertisements:

- Borders Official Site**
Shop Borders For Books, Music, DVDs, Gifts & More In One Online Store
www.borders.com
- Yahoo! is Hiring**
Work For One of the Best. Start Your Career with Yahoo! Today.
careers.yahoo.com
San Francisco-Oakland-San Jose, CA

互联网广告的主要类型(2)

- 网页广告(内容匹配广告, Content Match)


YAHOO! NEWS Search

HOME U.S. BUSINESS WORLD ENTERTAINMENT SPORTS TECH POLITICS ELECTIONS

Entertainment Video Celebrity TV Movies Music Reviews Fashion Books Arts

Beastie Boys recording "political" album REUTERS

By John Benson — Fri Oct 24, 4:13 am ET Buzz Up Send Share Print



Reuters — Michael 'Mike D' Diamond (L) and Adam 'MCA' Yauch of the Beastie Boys perform on the main stage during ...

CLEVELAND (Billboard) — A Get Out and Vote tour, the 2007 instrumental album "T ...

"We're actually in the middle Adam "Ad-Rock" Horowitz te sometime next year. It's a lo And it's political, depending toilet talk and fart jokes are yeah, very."

Any chance of new material 08" tour? "I don't think so," E you play the new songs that songs, it always seems like

sponsored links

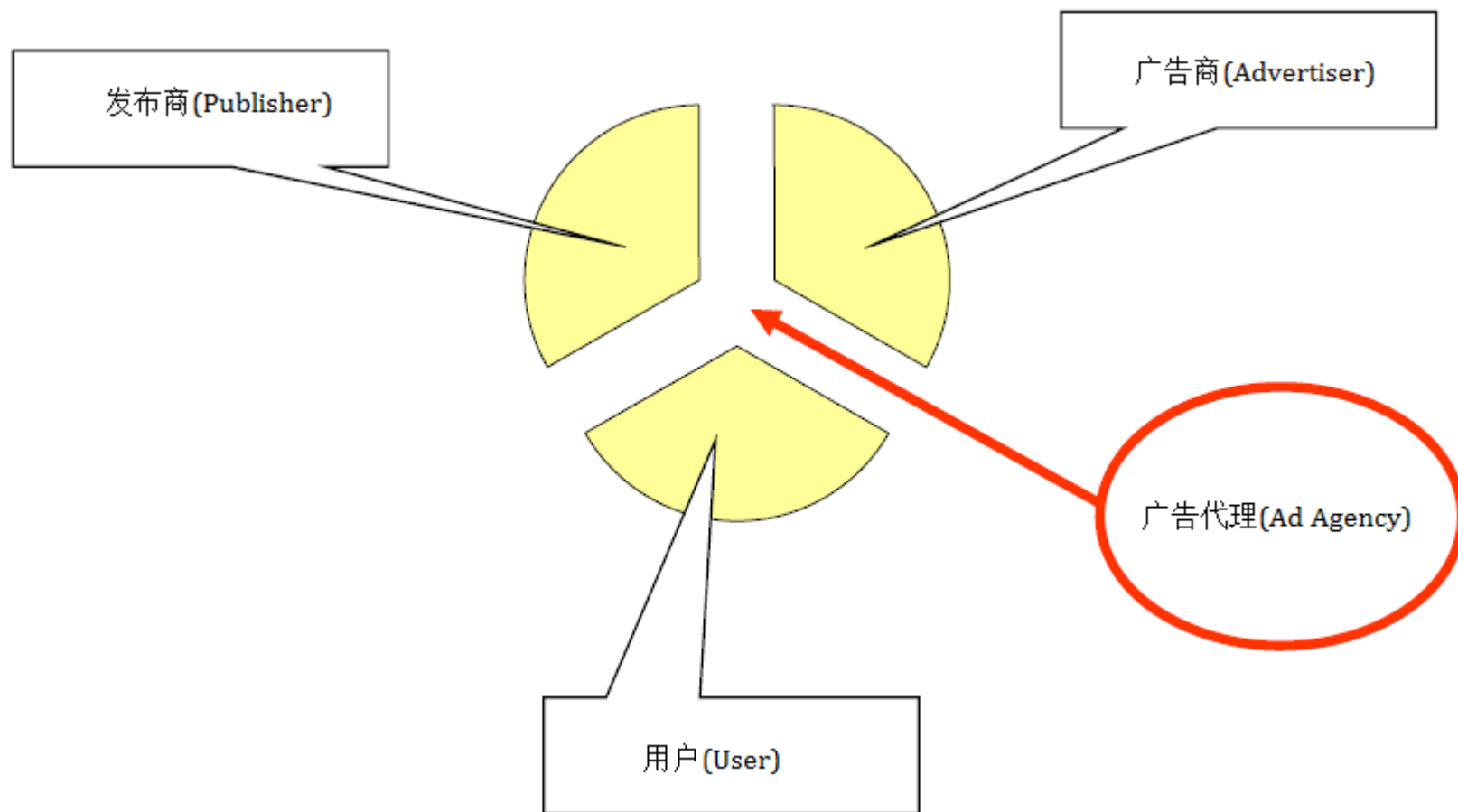
Beastie Boys Vs. Jay-Z - New York
Beastie Boys battle Jigga for NY supremacy.
Check it out on fuse.
fuse.tv/all-american-face-off

Beastie Boys
Browse a huge selection now. Find exactly what you want today.
www.ebay.com

Cell Phone Ringtones
Download Your Favorite Ringtone In Seconds.
www.RingtoneOcean.com

Video: D.L. Hughley on the election CNN

互联网广告中的利益三方



第一代搜索广告: Goto (1996) 竞价排名

www.goto.com/d/search/?sessionid=AQ42T4AAH0R5QFIEF3QPUQ?type=home&tr=1&Keywords=Wilmington+
Wilmington real estate. **Wilmington Real Estate**

Access 75% of all users now!
Premium Listings reach 75% of all
Internet users. [Sign up](#) for Premium
Listings today!

1. [Wilmington Real Estate - Buddy Blake](#)
Wilmington's information and real estate guide. This is your on
anything to do with Wilmington.
www.buddyblake.com (Cost to advertiser: **10.28**)
2. [Coldwell Banker Sea Coast Realty](#)
Wilmington's number one real estate company.
www.cbseacoast.com (Cost to advertiser: 10.37)
3. [Wilmington, NC Real Estate Becky Bullard](#)
Everything you need to know about buying or selling a home c
on my Web site!
www.iwwc.net (Cost to advertiser: 10.35)

第一代搜索广告: Goto (1996)



- Buddy Blake(房地产公司) 为此搜索投出最高价 (\$0.38)
- 只要某个人点击了该链接，Buddy Blake就要付\$0.38的费用给 Goto 公司
- 搜索结果按照投标价格的顺序排序 – Goto可以获得最大的收益
- 不区分广告还是文档，仅仅是一个结果列表！
- 广告预售，坦诚公开，没有相关度排序
- ...但是Goto并不假装存在相关度

第二代搜索广告: Google (2000/2001)

- 严格区分搜索结果和搜索广告

Google

andrei broder Search Advanced Search Preferences

Web Results 1 - 10 of about 91,100 for **andrei broder**. (0.12 seconds)

Andrei Broder - Wikipedia, the free encyclopedia
Andrei Zary Broder (Hebrew: אַנְדְּרֵי זָרִי בְּרֹדֶר) is a Research Fellow and Vice President of Emerging Search Technology for Yahoo!. He previously has worked for ...
en.wikipedia.org/wiki/Andrei_Broder - 21k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)

Andrei Broder | Yahoo! Research
Andrei Broder is a Yahoo! Research Fellow and Vice President for Computational Advertising. Previously he was an IBM Distinguished Engineer and the CTO of ...
research.yahoo.com/Andrei_Broder - 43k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)

DBLP: Andrei Z. Broder
 63, **Andrei Z. Broder**: Introduction: The Fourth International Workshop on ... 4, **Andrei Z. Broder**: A Provably Secure Polynomial Approximation Scheme for the ...
www.informatik.uni-trier.de/~ley/db/indices/a-tree/b/Broder.Andrei_Z_.html - 79k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)

Andrei Broder Joins Yahoo - Search Marketing News Blog - Search ...
 Nov 18, 2005 ... **Andrei Broder**, former vice president of research at AltaVista and until recently Distinguished Engineer & CTO, IBM Research, ...
blog.searchenginewatch.com/blog/061118-122544 - 83k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)

Yahoo! Search Blog: Search in the Future
 Feb 1, 2006 ... Questions for **Andrei Broder** re emerging search technology? ... Yesterday evening, we conducted the interview with **Andrei Broder** and we ...
www.ysearchblog.com/archives/000202.html - 45k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)

Sponsored Links

Borders Official Site
 Shop Borders For Books, Music, DVDs, Gifts & More In One Online Store
www.borders.com

Yahoo! Is Hiring
 Work For One of the Best. Start Your Career with Yahoo! Today.
careers.yahoo.com
 San Francisco-Oakland-San Jose, CA

两个列表结果: Web 网页 (左图) 及广告 (右图)

Web Images Maps News Shopping Gmail more Sign in

Google discount broker Search Advanced Search Preferences

Web Results 1 - 10 of about 807,000 for discount broker [definition]. (0.12 seconds)

Discount Broker Reviews
Information on online discount brokers emphasizing rates, charges, and customer comments and complaints.
www.broker-reviews.us/ - 94k - Cached - Similar pages

Discount Broker Rankings (2008 Broker Survey) at SmartMoney.com
Discount Brokers. Rank/ Brokerage/ Minimum to Open Account, Comments, Standard Commission*, Reduced Commission, Account Fee Per Year (How to Avoid), Avg. ...
www.smartmoney.com/brokers/index.cfm?story=2004-discount-table - 121k - Cached - Similar pages

Stock Brokers | Discount Brokers | Online Brokers
Most Recommended. Top 5 Brokers headlines. 10. Don't Pay Your Broker for Free Funds May 15 at 3:39 PM. 5. Don't Discount the Discounters Apr 18 at 2:41 PM ...
www.fool.com/investing/brokers/index.aspx - 44k - Cached - Similar pages

Discount Broker
Discount Broker - Definition of Discount Broker on Investopedia - A stockbroker who carries out buy and sell orders at a reduced commission compared to a ...
www.investopedia.com/terms/d/discountbroker.asp - 31k - Cached - Similar pages

Discount Brokerage and Online Trading for Smart Stock Market ...
Online stock broker SogoTrade offers the best in discount brokerage investing. Get stock market quotes from this Internet stock trading company.
www.sogotrade.com/ - 38k - Cached - Similar pages

15 questions to ask discount brokers - MSN Money
Jan 11, 2004 ... If you're not big on hand-holding when it comes to investing, a discount broker can be an economical way to go. Just be sure to ask these ...
moneycentral.msn.com/content/Investing/StartInvesting/P68171.asp - 34k - Cached - Similar pages

Sponsored Links

Rated #1 Online Broker
No Minimums. No Inactivity Fee.
Transfer to FirstTrade for Free!
www.firsttrade.com/

Discount Broker
Commission free trades for 30 days.
No maintenance fees. Sign up now.
TDAMERITRADE.com

TradeKing - Online Broker
\$4.95 per Trade, Market or Limit
SmartMoney Top Discount Broker 2001
www.TradeKing.com

Scottrade Brokerage
\$7 Trades, No Share Limit. In-Depth Research. Start Trading Online Now!
www.Scottrade.com

Stock trades \$1.95-\$3
100 free trades, up to \$100 back for transfer costs, \$500 minimum
www.sogotrade.com

\$3.95 Online Stock Trades
Market/Limit Orders, No Share Limit and No Inactivity Fees
www.Marsco.com

INGDIRECT | ShareBuilder
Discount Broker with 100% Free Trades

SogoTrade 出现在搜索结果中

SogoTrade 出现在广告中

搜索引擎是不是把广告商的结果放在非广告商的结果之前?

所有的主流搜索引擎都否认这一点

单个列表结果：广告在前，Web网页在后

Baidu 百度 北京动物园  [百度一下](#)

[网页](#) [资讯](#) [视频](#) [图片](#) [知道](#) [文库](#) [贴吧](#) [采购](#) [地图](#) [更多»](#)

百度为您找到相关结果约12,600,000个 [搜索工具](#)

搜索结果涉及价格仅作参考，请以商家官网为准

【携程】玩转上海野生动物园 门票120元起! 携程旅行 抢低价门票!



名称: 上海野生动物园 现价: **120元起**
 游玩类型: 动植物园
 介绍: 用携程app订上海野生动物园门票, 有特价, 方便实用, 快速入园! 携程旅行, 让出行说走就走, 在网上规划行程, 让旅游度假更...

上海野生动物园门票成人票	120元
上海野生动物园门票+水域探秘普通船票+投喂车票成人票	180元
上海野生动物园门票家庭票	200元

[携程旅游信息技术](#) 2020-06-01 [评价](#) [广告](#)

北京动物园 [官方](#)

门票: 每张15元(4月1日起至10月31日止) 门票半价票: 每张7.5元(4月1日起至10月31日止) 联票: 每张19元(4月1日起至10月31日止) 联票半价票: 每张9.5元...
www.bjzoo.com

北京动物园 [百度百科](#)



北京动物园位于北京市西城区西直门外大街, 东邻北京展览馆和莫斯科餐厅, 占地面积约86公顷, 水面8.6公顷。始建于清光绪三十二年(1906年), 是中国开放最早、饲养展出动物种类...
[历史沿革](#) [动物馆舍](#) [建筑古迹](#) [文化活动](#) [公园荣誉](#) [更多»](#)
<https://baike.baidu.com/>

广告是否会影响编辑的内容？

- 在报纸、电视上存在类似问题
- 报纸一般不会刊登针对其主要广告商的严厉指责性质的文章
- 在报纸和TV上，广告和编辑内容之间的界限往往变得很模糊
- 现在还不清楚搜索引擎广告是否和上面一样

广告在右部如何排序？

Web Images Maps News Shopping Gmail more

Sign in



discount broker

Search

[Advanced Search](#)
[Preferences](#)

Web

Results 1 - 10 of about 807,000 for **discount broker** [definition]. (0.12 seconds)

Discount Broker Reviews

Information on online **discount brokers** emphasizing rates, charges, and customer comments and complaints.

www.broker-reviews.us/ - 94k - [Cached](#) - [Similar pages](#)

Discount Broker Rankings (2008 Broker Survey) at SmartMoney.com

Discount Brokers. Rank/ **Brokerage**/ Minimum to Open Account, Comments, Standard Commis- sion*, Reduced Commission, Account Fee Per Year (How to Avoid), Avg. ...

www.smartmoney.com/brokers/index.cfm?story=2004-discount-table - 121k -

[Cached](#) - [Similar pages](#)

Stock Brokers | Discount Brokers | Online Brokers

Most Recommended. Top 5 **Brokers** headlines. 10. Don't Pay Your **Broker** for Free Funds May 15 at 3:39 PM. 5. Don't **Discount** the Discounters Apr 18 at 2:41 PM ...

www.fool.com/investing/brokers/index.aspx - 44k - [Cached](#) - [Similar pages](#)

Discount Broker

Discount Broker - Definition of **Discount Broker** on Investopedia - A stockbroker who carries out buy and sell orders at a reduced commission compared to a ...

www.investopedia.com/terms/d/discountbroker.asp - 31k - [Cached](#) - [Similar pages](#)

Discount Brokerage and Online Trading for Smart Stock Market ...

Online stock **broker** SogoTrade offers the best in **discount brokerage** investing. Get stock market quotes from this internet stock trading company.

www.sogotrade.com/ - 39k - [Cached](#) - [Similar pages](#)

15 questions to ask discount brokers - MSN Money

Jan 11, 2004 ... If you're not big on hand-holding when it comes to investing, a **discount broker** can be an economical way to go. Just be sure to ask these ...

moneycentral.msn.com/content/Investing/StartInvesting/P66171.asp - 34k -

[Cached](#) - [Similar pages](#)

Sponsored Links

Rated #1 Online Broker

No Minimums. No Inactivity Fee
Transfer to Firsttrade for Free!

www.firsttrade.com

Discount Broker

Commission free trades for 30 days.
No maintenance fees. Sign up now.

TDAMERITRADE.com

TradeKing - Online Broker

\$4.95 per Trade, Market or Limit
SmartMoney Top **Discount Broker** 2007

www.TradeKing.com

Scottrade Brokerage

\$7 Trades, No Share Limit. In-Depth
Research. Start Trading Online Now!

www.Scottrade.com

Stock trades \$1.50 - \$3

100 free trades, up to \$100 back
for transfer costs, \$500 minimum

www.sogotrade.com

\$3.95 Online Stock Trades

Market/Limit Orders, No Share Limit
and No Inactivity Fees

www.Marsco.com

INGDIRECT | ShareBuilder

如何对广告排序？

- 广告商对关键词竞标 – 拍卖方式
- 拍卖系统公开，任何人都可以参与关键词竞标
- 广告商仅在用户点击广告时才真正付费
- 拍卖机制如何确定某条广告的排序以及该广告的支付价格？
- 基本思路是次高价拍卖(second price auction)原则
- 搜索引擎中最重要的研究领域之一---计算广告学
 - 对每条广告压榨出一分钱也就意味着为搜索引擎公司带来上亿的额外收益

如何对广告排序？

- 简单的方法: 按照类似Goto的方式, 即按照投标价格排序
 - 不好的方法: 可能会被滥用
 - 例如: query [does my husband cheat?] → 有关离婚律师的广告
 - 我们并不想得到相关性上无关的广告
- 替代方法: 按照**投标价格和相关性排序**
- 相关度度量的关键指标: 点击率(clickthrough rate)
 - $\text{clickthrough rate} = \text{CTR} = \text{clicks per impressions}$
- 结果: 无关的广告将得到很低的排名
 - 即使在短期时间内降低了搜索引擎的收益
 - 希望: 如果用户能通过系统获得有价值的信息, 那么系统的总体接受程度和整体收益将最大化
- 其他排序因子: 位置、一天内的时间、着陆页(landing page)的质量和装载速度
 - 最主要的排序因子: 查询

Google AdsWords 的例子

Google次高竞标价格拍卖机制

advertiser	bid	CTR	ad rank	rank	paid
A	\$4.00	0.01	0.04	4	(minimum)
B	\$3.00	0.03	0.09	2	\$2.68
C	\$2.00	0.06	0.12	1	\$1.51
D	\$1.00	0.08	0.08	3	\$0.51

- **bid**: 每个广告商为每次点击给出的最大投标价格
- **CTR**: 点击率，即一旦被显示后被点击的比率。CTR是一种相关性度量指标。
- **ad rank**: $\text{bid} \times \text{CTR}$: 这种做法可以在 (i) 广告商愿意支付的价钱 (ii) 广告的相关度高低 之间进行平衡。
- **rank**: 拍卖中的排名
- **paid**: 广告商的次高竞标价格

Google次高竞标价格拍卖机制

advertiser	bid	CTR	ad rank	rank	paid
A	\$4.00	0.01	0.04	4	(minimum)
B	\$3.00	0.03	0.09	2	\$2.68
C	\$2.00	0.06	0.12	1	\$1.51
D	\$1.00	0.08	0.08	3	\$0.51

- 次高竞标价格拍卖：广告商支付其维持在拍卖中排名所必须的价钱(加上一分钱)(用它的下一名计算其支付价格)

$$\text{price}_1 \times \text{CTR}_1 = \text{bid}_2 \times \text{CTR}_2 \text{ (使得排名 } \text{rank}_1 = \text{rank}_2 \text{)}$$

$$\text{price}_1 = \text{bid}_2 \times \text{CTR}_2 / \text{CTR}_1$$

$$p_1 = \text{bid}_2 \times \text{CTR}_2 / \text{CTR}_1 = 3.00 \times 0.03 / 0.06 = 1.50$$

课堂练习：计算 p_2 和 p_3

Google次高竞标价格拍卖机制

advertiser	bid	CTR	ad rank	rank	paid
A	\$4.00	0.01	0.04	4	(minimum)
B	\$3.00	0.03	0.09	2	\$2.68
C	\$2.00	0.06	0.12	1	\$1.51
D	\$1.00	0.08	0.08	3	\$0.51

- 次高竞标价格拍卖：广告商支付其维持在拍卖中排名所必须的价钱(加上一分钱)(用它的下一名计算其支付价格)

$$\text{price}_1 \times \text{CTR}_1 = \text{bid}_2 \times \text{CTR}_2 \text{ (使得排名 } \text{rank}_1 = \text{rank}_2 \text{)}$$

$$\text{price}_1 = \text{bid}_2 \times \text{CTR}_2 / \text{CTR}_1$$

$$p_1 = \text{bid}_2 \times \text{CTR}_2 / \text{CTR}_1 = 3.00 \times 0.03 / 0.06 = 1.50$$

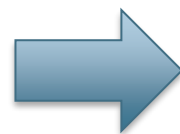
$$p_2 = \text{bid}_3 \times \text{CTR}_3 / \text{CTR}_2 = 1.00 \times 0.08 / 0.03 = 2.67$$

$$p_3 = \text{bid}_4 \times \text{CTR}_4 / \text{CTR}_3 = 4.00 \times 0.01 / 0.08 = 0.50$$

具有高投标价格的关键词

参考<http://www.cwire.org/highest-paying-search-terms/>

\$69.1	mesothelioma treatment options
\$65.9	personal injury lawyer michigan
\$62.6	student loans consolidation
\$61.4	car accident attorney los angeles
\$59.4	online car insurance quotes
\$59.4	arizona dui lawyer
\$46.4	asbestos cancer
\$40.1	home equity line of credit
\$39.8	life insurance quotes
\$39.2	refinancing
\$38.7	equity line of credit
\$38.0	lasik eye surgery new york city
\$37.0	2nd mortgage
\$35.9	free car insurance quote



- 医疗
- 律师
- 保险
- 融资
- 信贷

搜索广告：三赢？

- 每次用户点击广告，搜索引擎公司将会获得收益
- 用户只会点击其感兴趣的广告
 - 搜索引擎会对误导性和不相关的广告进行惩罚
 - 于是，用户在点击广告后往往会感到满意
- 广告商通过广告能够在物有所值的情况下找到新的客户

课堂练习

- 为什么和TV、报纸和电台相比，Web搜索对广告商更有吸引力？
- 广告商会为所有一切买单，那么它们会受到欺骗吗？
- 这对用户来说究竟是好消息还是坏消息？
- 当然，不论如何做，这都会危害搜索引擎

并非三赢：关键词套现(Keyword arbitrage)

- 比如从Google买一个关键词
- 然后将流量导向一个第三方页面，该页面对应机构付的钱将比你付给Google的多得多
 - 比如，重定向到一个满是广告的页面
- 该页面对于搜索用户来说基本没意义
- 广告作弊者一直在钻营想招
- 搜索引擎必须要花时间对付他们

并非三赢：商标侵权

- 例子: geico (美国政府雇员保险公司, 是美国第四大私人客户汽车保险公司)
- 2005年的部分时间内: 搜索词项 “geico” 在Google上可以买到
- Geico 在美国控告Google侵权
- Louis Vuitton(LV) 在欧洲控告Google侵权
- 参考 http://google.com/tm_complaint.html
- 如果采用商标做关键词, 那么用户可能被误导到一个页面, 该页面实际和用户期望购买的牌产品无关

提纲

- ① 上一讲回顾
- ② Web搜索系统
- ③ Web信息检索的特点
- ④ 互联网广告
- ⑤ 重复检测

重复检测

- Web上充斥重复内容
- 相对其它文档集合，Web上的重复内容更多
- **完全重复**(Exact duplicate)
 - 易剔除，比如采用哈希/指纹的方法
- **近似重复**(Near-duplicate)
 - Web上存在大量近似重复
 - 很难剔除
- 对用户而言，如果搜索结果中存在不少几乎相同的页面，那么体验非常不好
- 边缘相关度(Marginal relevance)为0：如果一篇高度相关的文档出现在另一篇高度近似的文档之后，那么该文档变得不相关(冗余)
- 必须要去除这些近似重复


近似重复例子

嫦娥五号探测器对接组合体成功分离


2020-12-06 13:12

新京报
BJNEWS.COM.CN

新京报快讯 据探月与航天工程中心官方微博消息，12月6日12时35分，嫦娥五号轨道返回组合体与上升器成功分离，进入环月等待阶段，将择机返回地球。



轨道组合体与上升器分离前模拟图（图片来源：探月与航天工程中心官方微博）



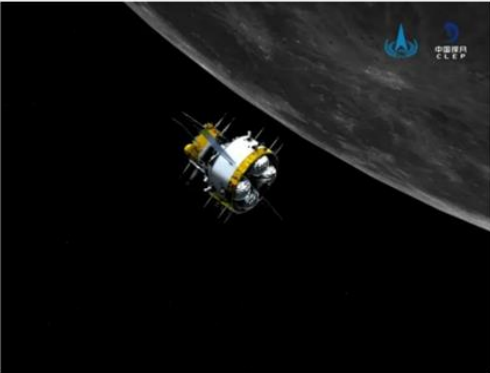
轨道组合体与上升器分离后模拟图（图片来源：探月与航天工程中心官方微博）

嫦娥五号探测器对接组合体成功分离


2020-12-06 13:10

中国军网

北京时间12月6日12时35分，嫦娥五号轨道器和返回器组合体与上升器成功分离，进入环月等待阶段，准备择机返回地球。



▲ 轨道组合体与上升器分离前模拟图



▲ 轨道组合体与上升器分离后模拟图

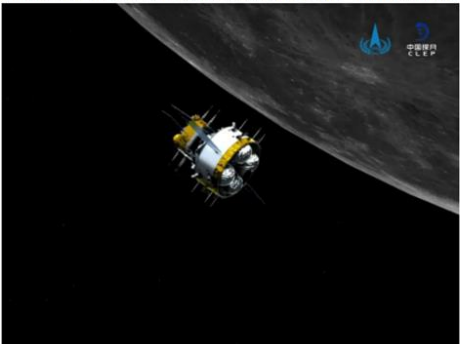
嫦娥五号探测器对接组合体成功分离，将择机返回地球

澎湃
THE PAPER

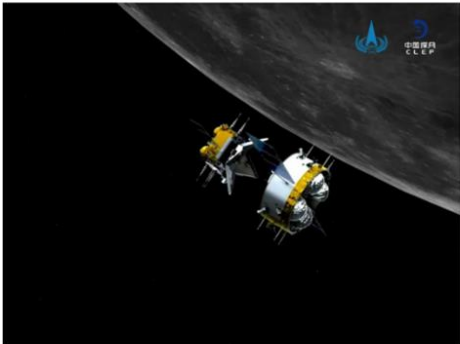
首页 > 澎湃防务

崔霞、王世玉、李厦/央视新闻客户端
2020-12-06 13:12

央视新闻客户端12月6日消息，记者从国家航天局获悉，北京时间12月6日12时35分，嫦娥五号轨道器和返回器组合体与上升器成功分离，进入环月等待阶段，准备择机返回地球。



轨道组合体与上升器分离前模拟图



轨道组合体与上升器分离后模拟图

重复检测

嫦娥五号探测器对接组合体成功分离

百度为您找到相关结果约7,360,000个 搜索工具

嫦娥五号探测器对接组合体成功分离
 北京时间12月6日12时35分,嫦娥五号轨道器和返回器组合体与上升器成功分离,进入环月等待阶段,准备择机返回地球。[详细>](#)
 央视新闻 2小时前

视频

 02:33 **嫦娥五号探测器完成在轨样品转移** 好看视频

 01:05 **视频来了!嫦娥五号探测器组合体成功分离** 好看视频

 00:48 **嫦娥五号探测器组合体成功分离** 好看视频

 00:47 **现场视频:嫦娥五号探测器组合体成功分离** 好看视频

资讯

 **权威发布 | 嫦娥五号探测器对接组合体成功分离**
 解放军报北京12月6日电(记者安普忠、贺逸舒)记者从国家航天局获悉,今天12时35分,嫦娥五号轨道器和返回器组合体与上升器成功分离...
 腾讯新闻 16分钟前

 **嫦娥五号探测器对接组合体成功分离**
 【嫦娥五号探测器对接组合体成功分离】记者从国家航天局获悉,12月6日12时35分,嫦娥五号轨道器返回器组合体与上升器成功分离...
 环球时报 2小时前

 **嫦娥五号探测器对接组合体成功分离 进入环月准备阶段**
 记者从国家航天局获悉,北京时间12月6日12时35分,嫦娥五号轨道器和返回器组合体与上升器成功分离,进入环月等待阶段,准备择机返回...
 央视新闻 2小时前

其他人还在搜

[嫦娥五号探测器的意义](#) [嫦娥五号探测器结构](#) [嫦娥五号交会对接](#)
[嫦娥五号月球探测器直播](#) [嫦娥五号探测器高清图](#) [嫦娥五号探测器图片](#)

[祝贺!嫦娥五号探测器对接组合体成功分离_手机网易网](#)
 36分钟前 【祝贺!嫦娥五号探测器对接组合体成功分离】记者从国家航天局获悉,北京时间12月6日12时35分,嫦娥五号轨道器和返回器组合体与上升器成功分离,进入环月...
 网易新闻 百度快照

课堂思考题

如何去掉Web上的近似重复页面呢？

近似重复的检测

- 采用Jaccard距离、编辑距离等指标计算页面之间的相似度
- 需要说明的是，我们希望检测那些“语法”(syntactic)上而不是“语义”(semantic)上相似的页面
 - 页面的语义相似度(即内容语义之间的相似度)非常难以计算
- 也就是说，我们并不考虑那些内容意义上相似但是表达方式不同的近似重复
- 引入一个相似度阈值 θ 来判定“两个页面之间是否近似重复”
 - 比如，如果两篇文档的相似度 $>$ 阈值 $\theta = 80\%$ ，那么认为两篇文档近似重复

将每篇文档表示成一个shingle集合

- 每个 shingle 是一个基于词语的n-gram
- 使用shingle来计算文档之间的语法相似度
- 比如，对于 $n = 3$ ，那么文档 “a rose is a rose is a rose” 就可以表示成shingle的集合：
 - { a-rose-is, rose-is-a, is-a-rose }
- 我们可以通过指纹(fingerprinting)算法将shingle映射到 $1..2^m$ (例如 $m = 64$)之间
- 接下来我们用 s_k 来代表某个shingle映射到 $1..2^m$ 之间的一个数
- 两个文档的相似度定义为它们的shingle集合的Jaccard距离

Jaccard距离计算回顾

- 一个常用的计算两个集合重合度的方法
- 令 A 和 B 分别表示两个集合
- Jaccard距离:

$$\text{JACCARD}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$$(A \neq \emptyset \text{ or } B \neq \emptyset)$$

- $\text{JACCARD}(A, A) = 1$
- $\text{JACCARD}(A, B) = 0$ if $A \cap B = \emptyset$
- 并不要求 A 和 B 的大小一样
- Jaccard距离取值在 $[0, 1]$ 之间

课堂练习：Jaccard距离计算

- 3篇文档：

d_1 : “Jack London traveled to Oakland”

d_2 : “Jack London traveled to the city of Oakland”

d_3 : “Jack traveled from Oakland to London”

- 基于2-gram的shingle表示，计算文档 d_1 和 d_2 之间的Jaccard距离 $J(d_1, d_2)$ ，计算文档 d_1 和 d_3 之间的Jaccard距离 $J(d_1, d_3)$

Jaccard距离计算的例子

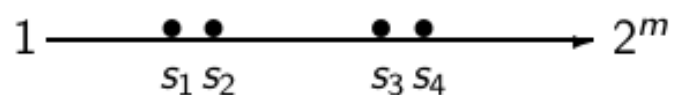
- 3篇文档：
 - d_1 : “Jack London traveled to Oakland”
 - d_2 : “Jack London traveled to the city of Oakland”
 - d_3 : “Jack traveled from Oakland to London”
- 基于2-gram的shingle表示，可以计算文档之间的Jaccard距离如下：
 - $J(d_1, d_2) = 3/8 = 0.375$
 - $J(d_1, d_3) = 0$
- 注意：Jaccard距离对差异十分敏感

将文档表示成梗概(sketch)

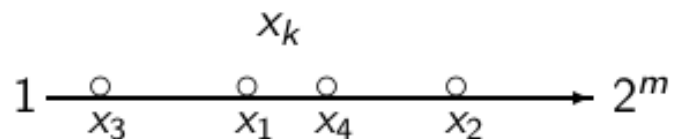
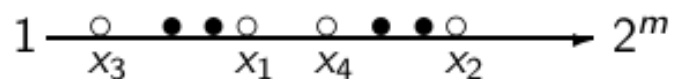
- 每篇文档的shingle的个数非常多
- 为提高效率，接下来我们使用文档的梗概来表示文档，它由文档的shingle集合中精巧挑选出的子集构成
- 比如，梗概中shingle的数目为 $n = 200 \dots$
- \dots 通过一系列置换 $\pi_1 \dots \pi_{200}$ 来定义
- 每个置换 π_i 都是 $1..2^m$ 上的随机置换
- 文档 d 的梗概定义为：
 $\langle \min_{s \in d} \pi_1(s), \min_{s \in d} \pi_2(s), \dots, \min_{s \in d} \pi_{200}(s) \rangle$
(一个200维的数字向量)

置换和最小值：例子

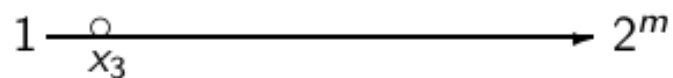
文档 1: $\{s_k\}$



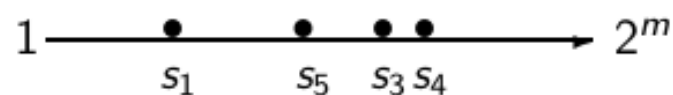
$$x_k = \pi(s_k)$$



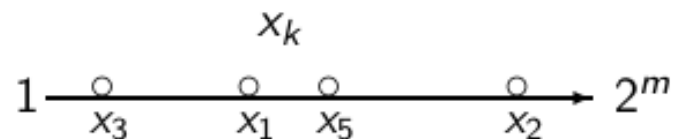
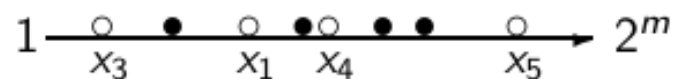
$$\min_{s_k} \pi(s_k)$$



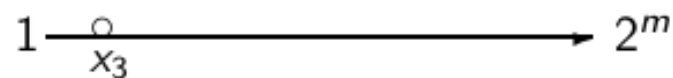
文档2: $\{s_k\}$



$$x_k = \pi(s_k)$$



$$\min_{s_k} \pi(s_k)$$



使用 $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$ 作为文档 d_1 和 d_2 是否近似重复的测试条件？该例子中置换 π 表明: $d_1 \approx d_2$

计算梗概之间的Jaccard距离 (1)

- 现在每篇文档都变成了一个 $n=200$ 维的数字向量
- 该向量比高维空间下的shingle容易处理得多
- 如何计算Jaccard距离?

计算梗概之间的Jaccard距离 (2)

- 如何计算?
- 令 U 为 d_1 和 d_2 的 shingle 并集, I 为它们的交集
- 对于 U 而言就存在 $|U|!$ 个置换集合
- 对于 $s' \in I$, 有多少置换 π 会使得

$$\operatorname{argmin}_{s \in d_1} \pi(s) = s' = \operatorname{argmin}_{s \in d_2} \pi(s)?$$
- 答案是: $(|U| - 1)!$
- 对于 I 的每个 s , 存在着 $(|U| - 1)!$ 个不同的置换集合 \Rightarrow
 于是总共有 $|I|(|U| - 1)!$ 个置换能够保证

$$\operatorname{argmin}_{s \in d_1} \pi(s) = \operatorname{argmin}_{s \in d_2} \pi(s) \text{ 为真}$$
- 因此, 使得 $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$ 为真的置换比例为:

$$\frac{|I|(|U| - 1)!}{|U|!} = \frac{|I|}{|U|} = J(d_1, d_2)$$

Jaccard距离估计

- 因此，成功的置换比例就等于Jaccard距离
 - 置换 π 成功当且仅当 $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$
- 随机选取一个置换，当置换成功时输出1，否则输出0，该过程是一个贝努利试验过程
- 成功概率的估计：在 n 次贝努利试验中成功比率 ($n = 200$)
- 我们使用的梗概是基于置换的随机选择
- 因此，为了计算Jaccard距离，统计 $\langle d_1, d_2 \rangle$ 上的成功置换个数 k ，然后除以 $n = 200$.
- $k/n = k/200$ 就是 $J(d_1, d_2)$ 的估计值

实现

- 使用哈希函数来实现高效的置换:
$$h_i : \{1..2^m\} \rightarrow \{1..2^m\}$$
- 以任意顺序扫描两个集合并集中的所有shingle s_k
- 对每个哈希函数 h_i 及文档 d_1, d_2, \dots : 在某个固定存储位置中保留当前的最小值
- 如果 $h_i(s_k)$ 小于当前的最小值, 那么对固定存储位置上的值进行更新

例子

	d_1	d_2
s_1	1	0
s_2	0	1
s_3	1	1
s_4	1	0
s_5	0	1

$$h(x) = x \bmod 5$$

$$g(x) = (2x + 1) \bmod 5$$

$$\min(h(d_1)) = 1 \neq 0 =$$

$$\min(h(d_2)) \quad \min(g(d_1)) =$$

$$2 \neq 0 = \min(g(d_2))$$

$$\hat{J}(d_1, d_2) = \frac{0+0}{2} = 0$$

	d_1 slot		d_2 slot	
h	∞		∞	
g	∞		∞	
$h(1) = 1$	1	1	–	∞
$g(1) = 3$	3	3	–	∞
$h(2) = 2$	–	1	2	2
$g(2) = 0$	–	3	0	0
$h(3) = 3$	3	1	3	2
$g(3) = 2$	2	2	2	0
$h(4) = 4$	4	1	–	2
$g(4) = 4$	4	2	–	0
$h(5) = 0$	–	1	0	0
$g(5) = 1$	–	2	1	0

最终的梗概

课堂练习

	d_1	d_2	d_3
s_1	0	1	1
s_2	1	0	1
s_3	0	1	0
s_4	1	0	0

$$h(x) = (5x + 5) \bmod 4$$

$$g(x) = (3x + 1) \bmod 4$$

Estimate $\hat{J}(d_1, d_2)$, $\hat{J}(d_1, d_3)$, $\hat{J}(d_2, d_3)$

解答 (1)

	d_1	d_2	d_3
s_1	0	1	1
s_2	1	0	1
s_3	0	1	0
s_4	1	0	0

$$h(x) = (5x + 5) \bmod 4$$

$$g(x) = (3x + 1) \bmod 4$$

	d_1 slot	d_2 slot	d_3 slot
	∞	∞	∞
	∞	∞	∞
$h(1) = 2$	— ∞	2 2	2 2
$g(1) = 0$	— ∞	0 0	0 0
$h(2) = 3$	3 3	— 2	3 2
$g(2) = 3$	3 3	— 0	3 0
$h(3) = 0$	— 3	0 0	— 2
$g(3) = 2$	— 3	2 0	— 0
$h(4) = 1$	1 1	— 0	— 2
$g(4) = 1$	1 1	— 0	— 0

最终的梗概

解答 (2)

$$\hat{J}(d_1, d_2) = \frac{0 + 0}{2} = 0$$

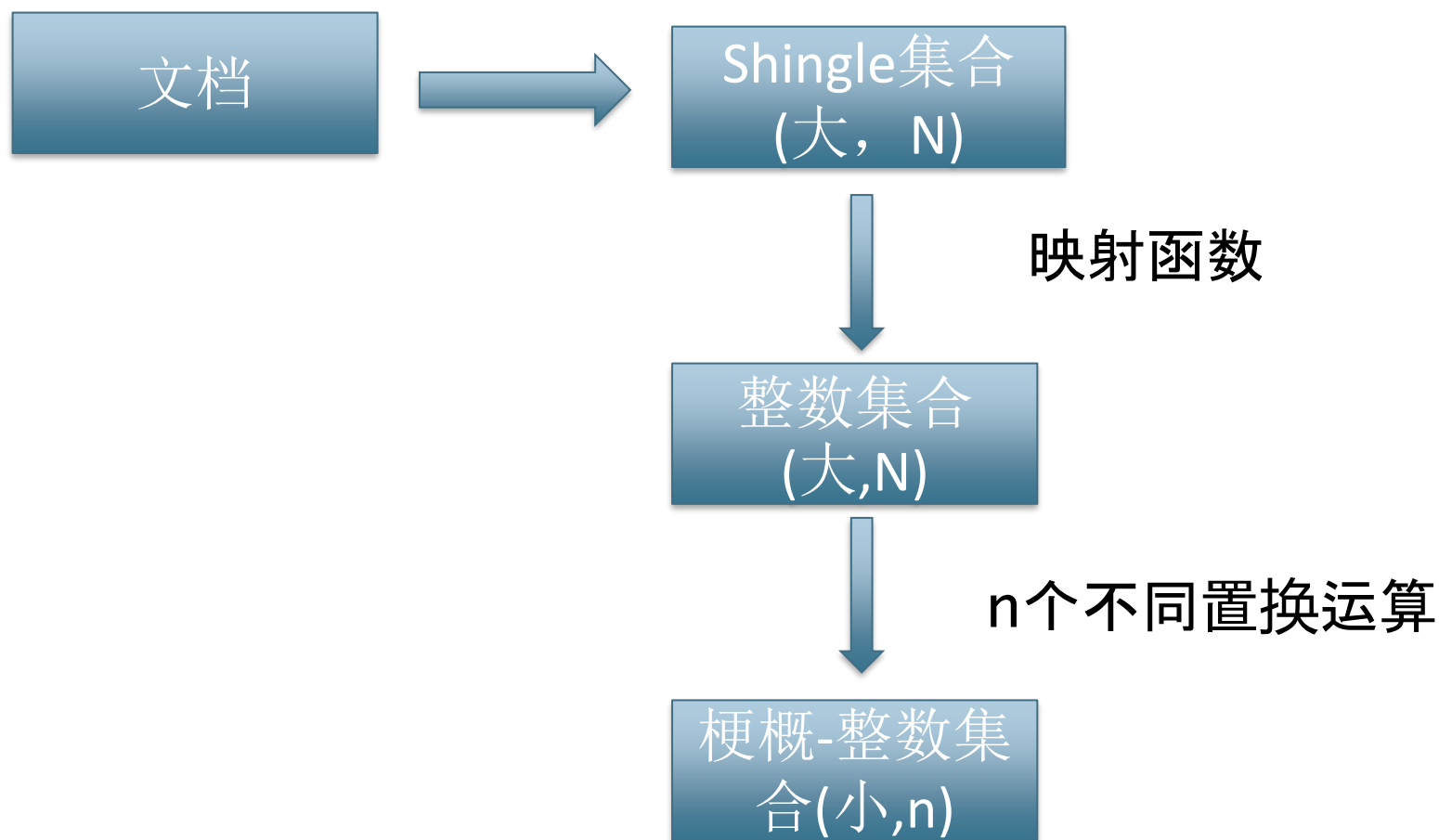
$$\hat{J}(d_1, d_3) = \frac{0 + 0}{2} = 0$$

$$\hat{J}(d_2, d_3) = \frac{0 + 1}{2} = 1/2$$

Shingling技术概要

- 输入： N 篇文档
- 选择用于生成shingle的n-gram的大小，例如 $n = 5$
- 选择200个随机置换，每个置换可以通过哈希函数表示
- 计算 N 个梗概值: 得到一个 $200 \times N$ 的矩阵(参考前面的例子)，其中每一行对应一个置换，每一列对应一个文档
- 计算 $\frac{N \cdot (N-1)}{2}$ 个两两文档之间的相似度
- 将所有两两之间相似度大于 θ 的文档构成一个传递闭包
- 对每一个传递闭包只索引一篇文档

文档的表示过程



高效的近似重复检测

- 现在我们已经得到了一个非常高效的算法来估计两篇文档的Jaccard距离
- 但是，如果Web网页数目为 N ，那么仍然需要估计 $O(N^2)$ 个相似度
- 仍然无法处理
- 一种解决办法: 局部敏感哈希(locality sensitive hashing , 简称LSH, 也常译成位置敏感哈希)
- 另一种解决办法: 排序 (Henzinger 2006)

本讲小结

- Web搜索系统分析
- Web信息检索的特点
- 互联网广告
- 重复检测

参考资料

- 《信息检索导论》第 19 章
- <http://ifnlp.org/ir>
- Stanford 计算广告学课程,
<http://www.stanford.edu/class/msande239/>
- van den Bosch A, Bogers T, de Kunder M. Estimating search engine index size variability: a 9-year longitudinal study. *Scientometrics*. 2016;107:839–856.
doi:10.1007/s11192-016-1863-z
- [The size of the World Wide Web \(The Internet\)](https://www.worldwidewebsize.com/)
<https://www.worldwidewebsize.com/>

课后练习

- 无！