大海捞针亦有道

——浅谈网络信息检索技术的现状与挑战

智能技术与系统国家重点实验室 智能检索组

刘奕群 2006年3月

网络信息检索技术的现状与挑战

清华大学
Tsinghua University

- 问题背景
- 网络信息检索系统的构成及运行原理
- 网络信息检索研究面临的挑战

网络信息检索技术的现状与挑战

清华大学
Tsinghua University

- 问题背景
- 网络信息检索系统的构成及运行原理
- 网络信息检索研究面临的挑战

问题背景

- 2005年搜索引擎市场的激烈竞争
  - Google市值的变化举世关注
  - Baidu上市造就数以百计的百万富翁
  - MSN推出新版搜索，MSRA建立搜索研究中心
  - Yahoo中国重组
  - 主要门户网站Sohu, Sina, Netease, 腾讯纷纷推出搜索引擎产品

问题背景

- How Much Info Project
  - How much information is produced in the world each year?
  - Produced by the School of Information Management and Systems at UC Berkeley
  - Financial support from Microsoft Research, Intel, Hewlett-Packard, and EMC.

问题背景

- What does How-Much-Info find?
  - Print, film, magnetic, and optical storage media produced about 5 exabytes of new information in 2002.
    - 1,000,000,000,000,000,000 bytes OR 10^{18} bytes
    - 5 Exabytes: All words ever spoken by human beings.
  - Equivalent of 250 megabytes per person for each man, woman, and child on earth.
  - 92 percent of the new information was stored on magnetic media, mostly in hard disks.

与Web的出现和发展直接相关

问题背景

- World Wide Web的出现与发展

38年

13年

4年

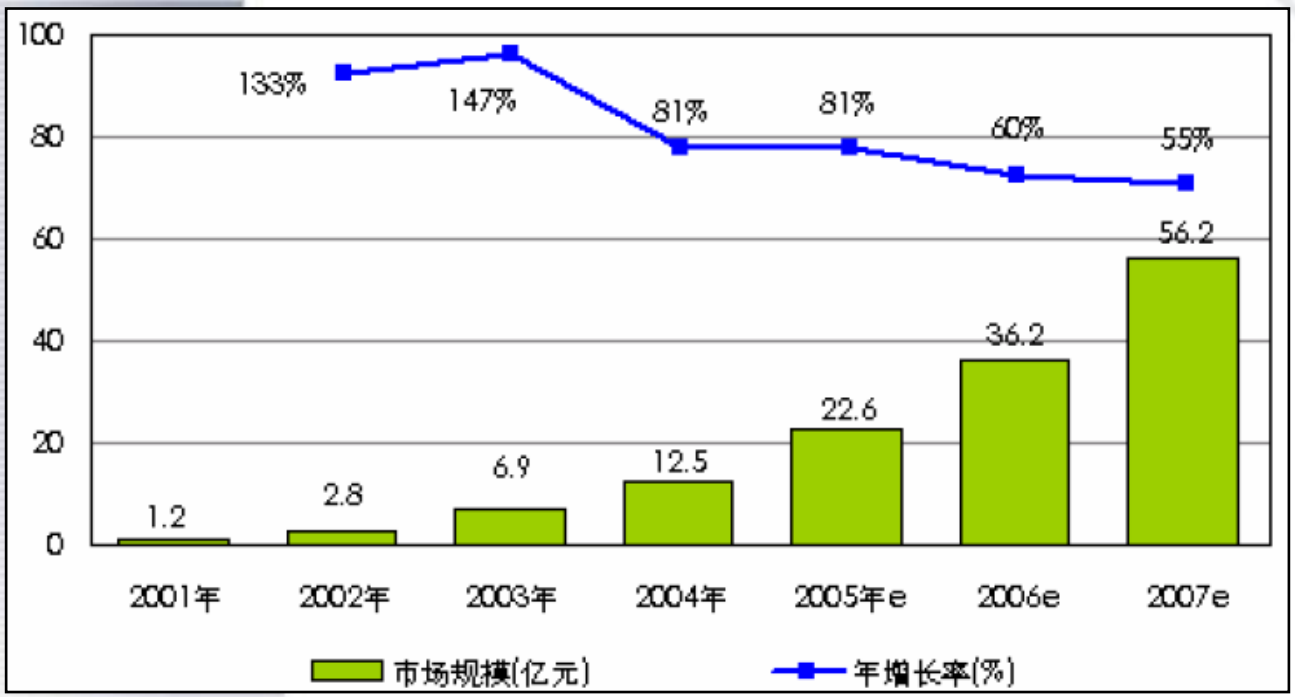
1994年个人浏览器诞生，到1998年用户超过5000万人

问题背景

- Web的发展带来了什么？
  - 信息数量的急剧膨胀
  - 知识的获取空前简单与繁荣
    - Information is no longer a scarce resource - attention is.
      （纽约时报，2005年10月16日）
    - 在信息化时代，知识实际上已经不是资源，智慧才是资源。（经管学院魏杰教授）
  - 从Web中有效的获取知识正在成为人们生活与工作的必须技能
    - 高科技企业员工1/3的时间用于查找资料
    - 由于无法找到有效信息而浪费的产值占企业收入1/5

问题背景

- 网络信息检索工具/搜索引擎的发展

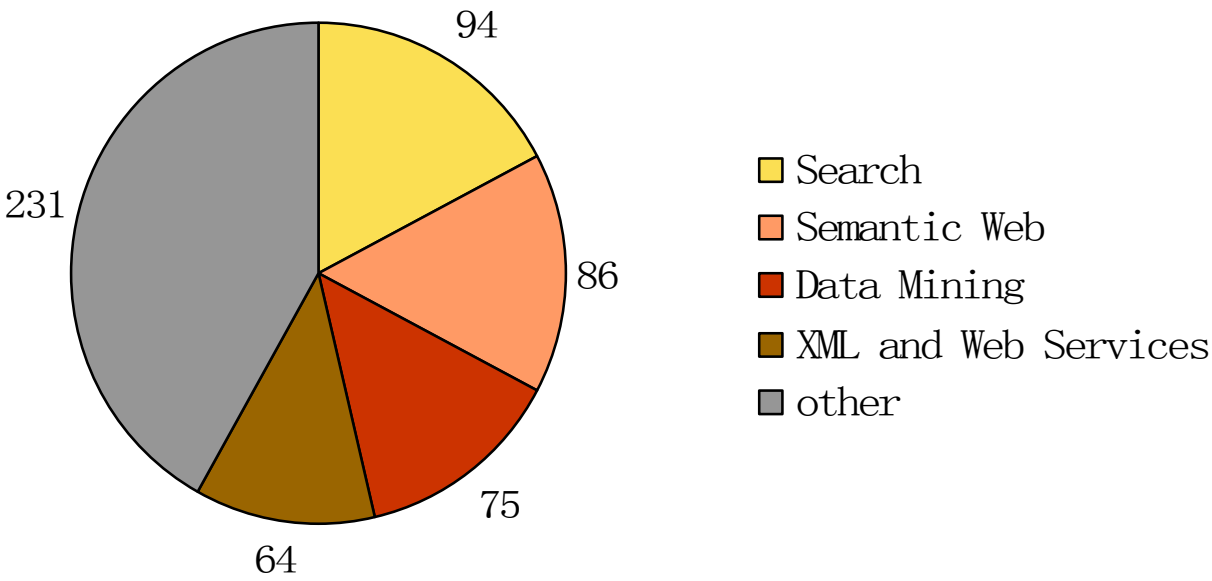


– 与Mp3的市场规模（50亿元）相当，但发展速度更快，利润值更大

问题背景

- 网络信息检索工具/搜索引擎的发展（续）

WWW 05’论文领域分布

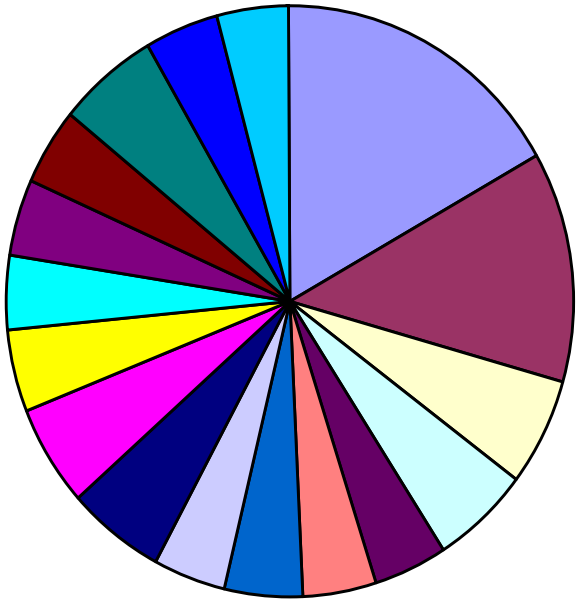


The 14th International World Wide Web Conference (WWW2005) JAPAN May 10 (Tue) - 14 (Sat), 2005

问题背景

- 网络信息检索工具/搜索引擎的发展（续）

SIGIR 05’论文领域分布



- Theory
- Web Search
- Relevance Feedback
- Distributed IR
- Filtering
- Categorization and Classification
- Evaluation
- Summarization
- Efficiency
- Categorization and Supervised Machine Learning
- Structured Data
- NLP
- Multimedia
- Question Answering
- User Studies

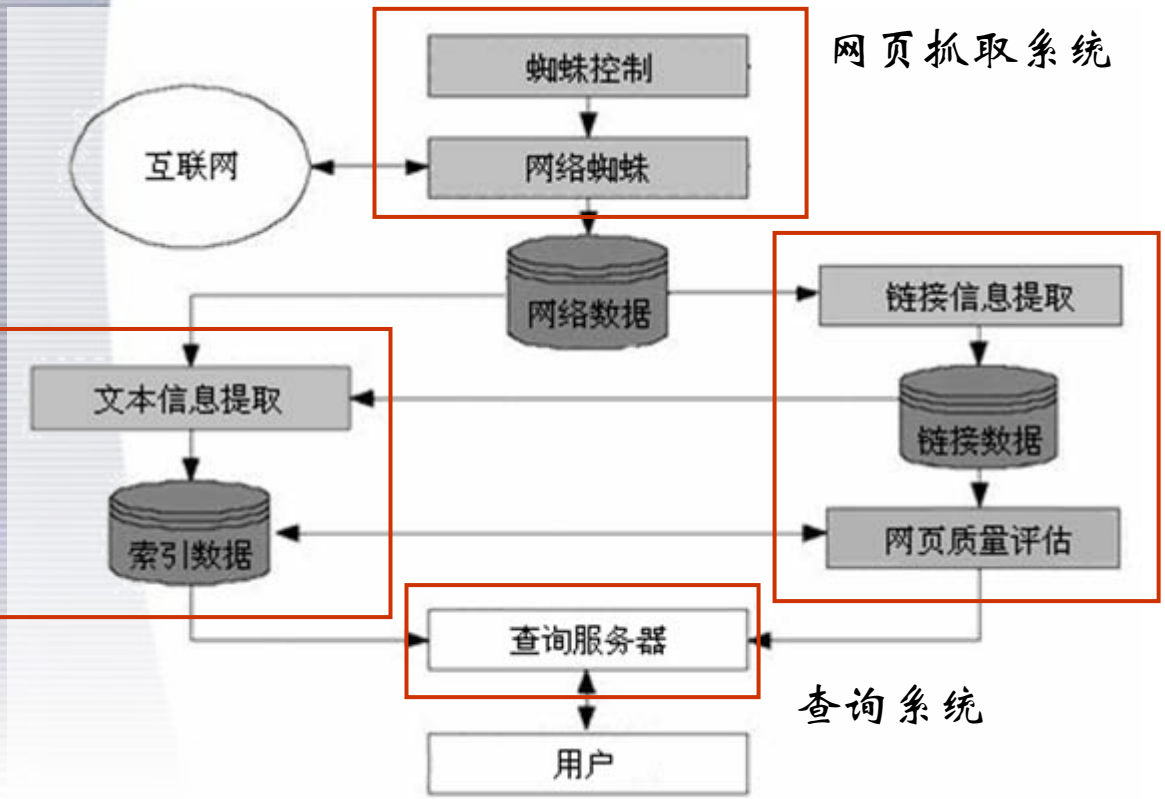
网络信息检索技术的现状与挑战

清华大学
Tsinghua University

- 问题背景
- 网络信息检索系统的构成及运行原理
- 网络信息检索研究面临的挑战

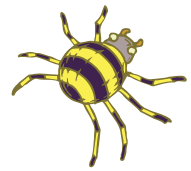
网络信息检索系统的构成及运行原理

- 网络信息检索系统的构成



网络信息检索系统的构成及运行原理

- 网页抓取系统
  - Crawler, Spider
  - 累积式抓取（cumulative crawling）
    - 用于索引的建立
    - 耗时长
    - 根据网页的重要性进行抓取
  - 增量式抓取（incremental crawling）
    - 用于索引的更新
    - 即时进行，绵绵不绝
    - 根据网页的时效性和重要性进行抓取



网络信息检索系统的构成及运行原理

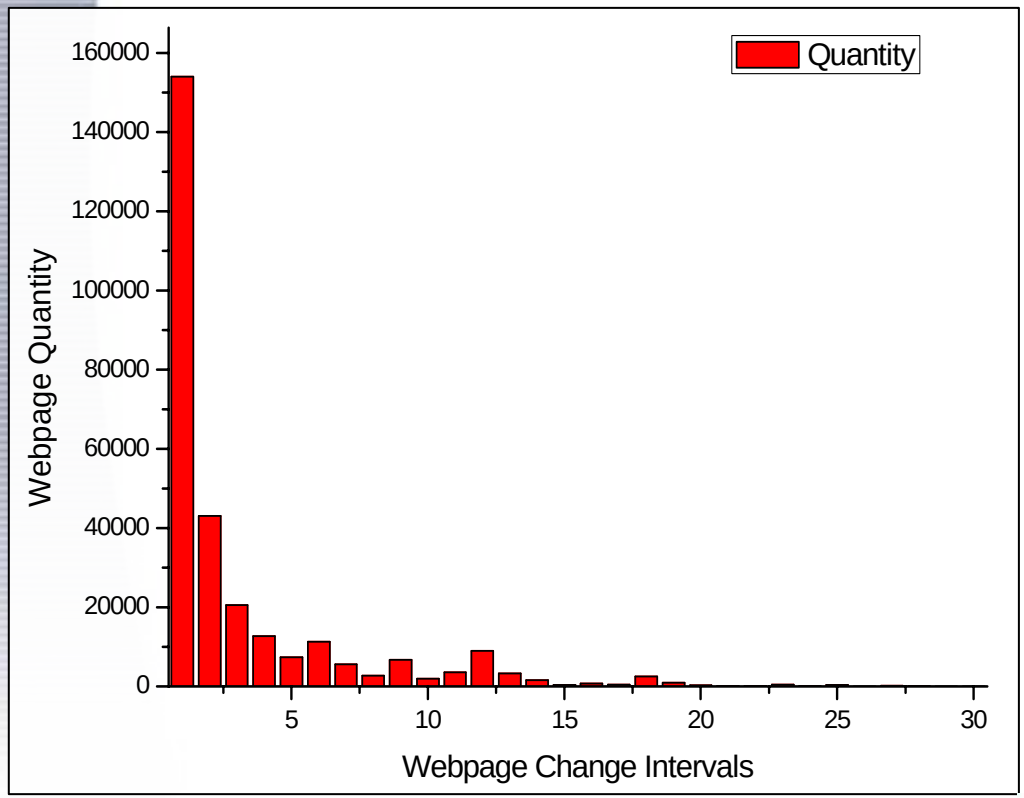
- 网页抓取系统
  - 累积式抓取
    - 以一定量的网页作为出发点，追随网页中的链接结构信息抓取新的页面，周而复始
    - 关键问题1：页面判重问题
      - URL层次的判重
      - 内容层次的判重
    - 关键问题2：页面质量问题
      - 如何利用有限的系统资源抓取尽量多的高质量页面
    - 关键问题3：黑洞
      - Blog, BBS, 论坛

网络信息检索系统的构成及运行原理

- 网页抓取系统
  - 增量式抓取
    - 根据网页寿命长短，进行更新以及新网页的发现和抓取。
    - 关键问题1：网页寿命
      - 变化越频繁的网页变化比例越小
      - 泊松分布的网页寿命模型
    - 关键问题2：更新策略
      - 保证时鲜度的更新
      - 保证重要网页的更新
    - 关键问题3：与累积式抓取类似的问题

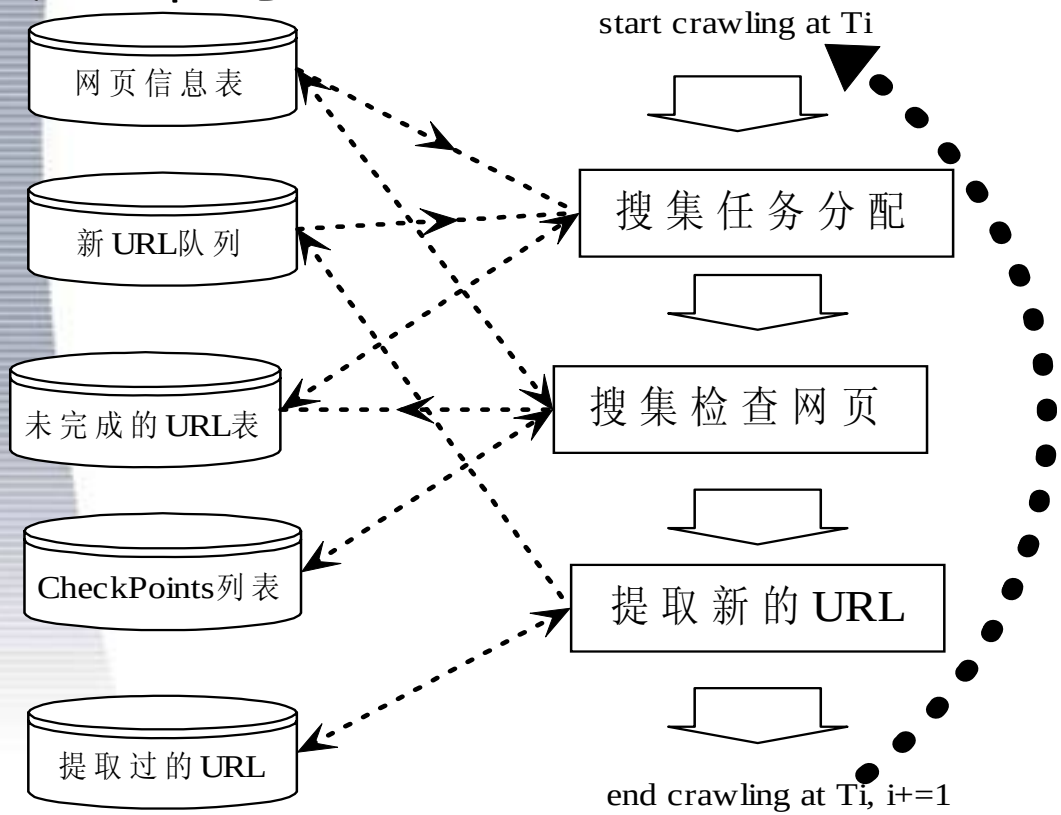
网络信息检索系统的构成及运行原理

- 网页抓取系统



网络信息检索系统的构成及运行原理

- 网页抓取系统
  - 大致的工作流程



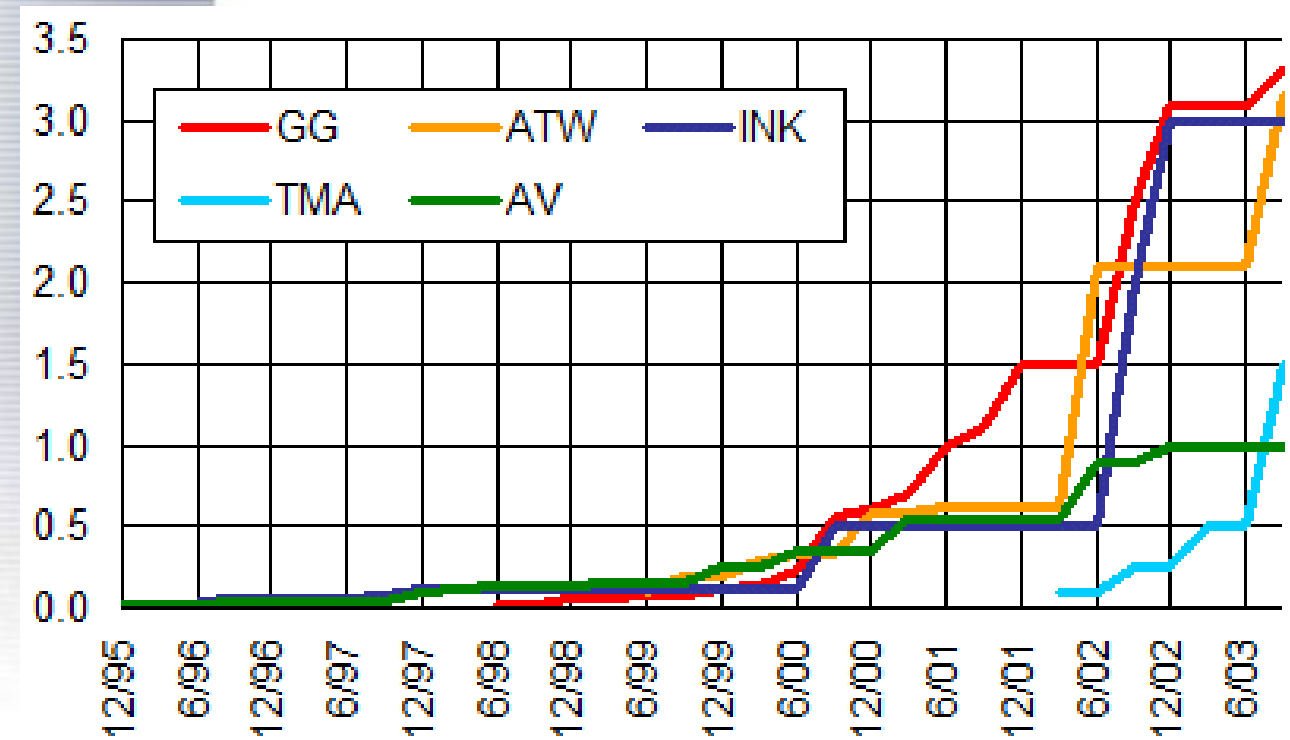
网络信息检索系统的构成及运行原理

- 索引系统
  - 文档、查询与词项
    - 文档（document）：检索的目标
    - 查询（query）：检索的要求
    - 词项（term）：构成文档与查询的基本单位
  - 索引系统的作用
    - 将抓取下来的网页信息予以保存
    - 倒排结构

Term 1	Doc 1, pos 1	Doc 1, pos 2	…	Doc m, pos n
…				
Term N	Doc 1, pos 1	Doc 1, pos 2	…	Doc p, pos q

网络信息检索系统的构成及运行原理

- 索引系统
  - 搜索引擎索引规模的竞争（Index Size War）



网络信息检索系统的构成及运行原理

- 索引系统

Search Engine	Reported Size	Page Depth
Google	**8.1 billion (Dec. 2004)**	101K
MSN	5.0 billion	150K
Yahoo	19.2 bilion (Aug. 2005)	500K
Ask Jeeves	2.5 billion	101K+
All the Web	**152 billion**	605K
All the Surface Web	**10 billion**	8K

2002.12

From Danny Sullivan, SearchEngineWatch web site

网络信息检索系统的构成及运行原理

- 索引系统
  - 任何搜索引擎都无法涵盖互联网上所有的信息

	搜索引擎WEB资源覆盖率(%)			
	Google	Yahoo!	MSN	Teoma
第一轮	76.30	69.28	62.03	57.58
第二轮	76.09	69.39	61.90	57.69
第三轮	76.27	69.37	61.87	57.70
第三轮	76.05	69.30	61.73	57.57
第五轮	76.11	69.26	61.96	57.56
平均	**76.16**	**69.32**	**61.90**	**57.62**

  - 9月27日，Google 宣布不再公布索引规模数据

“Absolute numbers are no longer useful ”

网络信息检索系统的构成及运行原理

- 索引系统需要解决的关键问题
  - 索引的有效性问题
    - 索引项的有效性（term, doc, pos）
    - 索引到的文档的有效性
  - 增量式系统的索引更新问题
  - 中文搜索引擎系统中的问题
    - “的”字的大麻烦
    - 分词粒度的考虑

网络信息检索系统的构成及运行原理

- 页面质量评估系统
  - 网络信息检索系统与传统检索系统最大的差别
  - 网页数据庞大的规模背后：
    - 中文网络环境中，完全镜像页面占25-40%（自2003年中国互联网络信息资源数量调查报告）
    - 垃圾页面、虚假页面、广告页面众多
    - ……
  - 如果无法保证索引到数据的质量，则只能是在低质量数据上浪费宝贵的存储和计算资源

网络信息检索系统的构成及运行原理

- 每天搜索引擎用户使用到多少被索引的页面
  - 李彦宏：搜索引擎里每天有400多万被检索的关键词。
  - 一般而言不重复的关键词会占总数的30%以内
  - 对于每个关键词，用户平均点击的页面数在2页以内
  - 假设：
    - 关键词返回的结果也没有交集
  - 则可以得到：
    - 用户使用到的被索引的页面数为2400万个左右。

网络信息检索系统的构成及运行原理

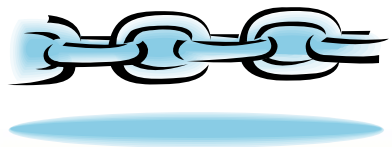
- 索引规模增长的终结？
  - 在百度的平均更新周期（1个月）内，用户共可能访问到的页面总数为7.2亿个，
  - 少于百度声称的索引量（8亿）
  - 更少于中文网页总数（20亿）
- 用户至多访问的页面数也少于搜索引擎的索引量，意味着搜索引擎索引到的页面中必然有一部分没有在任何一次检索中发挥作用。
- 去粗取精，去伪存真就是页面质量评估的目标

网络信息检索系统的构成及运行原理

- 宏观粒度的页面质量评估
  - 目的：找出对用户检索信息有用的页面
  - 当前的研究重点：Web链接结构分析
  - 如果存在超链接*L*从页面*P(source)*指向页面*P(destiny)*，则*P(source)*与*P(destiny)*之间满足：

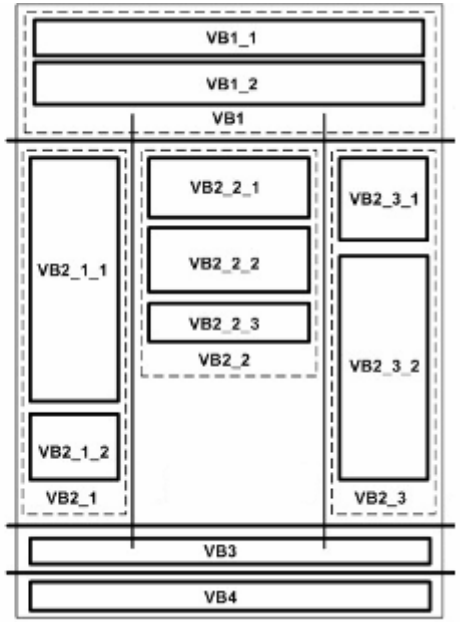
    假设1：（内容推荐假设）页面*P(source)*的作者推荐页面*P(destiny)*的内容，且利用*L*的链接文本对*P(destiny)*进行描述。

    假设2：（主题相关假设）被超链接连接的两个页面*P(source)*与*P(destiny)*比随机抽取的两个页面有更大的概率有内容相关性。
  - PageRank（Google），HITS（Kleinberg.）及众多的改进算法



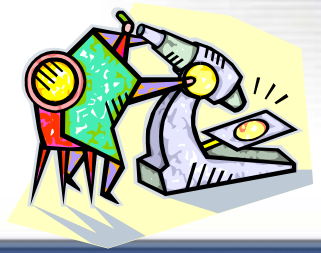
网络信息检索系统的构成及运行原理

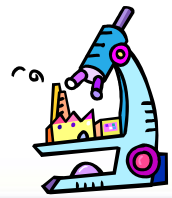
- 微观粒度的页面质量评估
  - 目的：找出对用户检索信息有用的页面的某个部分
  - 当前研究重点：去除特定垃圾信息；页面分块模型



网络信息检索系统的构成及运行原理

- 微观粒度的页面质量评估（续）
  - 去除特定垃圾信息（利用机器学习方法和一定量的训练）
    - 去除广告条（Davison et. al.）
    - 去除页面中的无关链接与垃圾链接（Kushmerick et. al.）
  - 页面分块模型
    - 依据语料统计信息计算页面块的信息量（Lin et. al.）
    - 基于模板频度检测构建站点模板（Yossef et. al. Yi et. al.）
    - 基于页面块的绝对位置和机器学习方法计算块的重要性（VIsion Based Page Segmentation, VIPS, MSRA）





网络信息检索系统的构成及运行原理

- 页面质量评估的研究现状
  - 微观粒度
    - 具有数据挖掘方面研究的积累（数据预处理、数据清理等）
    - 相对比较成熟完善
  - 宏观粒度
    - 搜索引擎竞价排名机制的引入，带来了大量的链接垃圾
    - 内容推荐和主题相关假设受到挑战
    - 过多关注页面自身的特性，忽略用户的实际需求
    - 只重视链接结构特征，忽略页面其他类型的查询无关特征

网络信息检索系统的构成及运行原理

- 页面质量评估：我们的做法
  - 有可能成为用户检索目标的页面才是高质量的
  - 使用页面成为用户检索目标的概率来评判质量
- 用户需要什么？
  - 反映在用户查询的目标页面中
  - 高质量页面：可能成为用户检索目标的Web页面
  - 矛盾：
    - 查询目标页面是与查询相关的
    - 页面质量评估是查询无关的过程必须使用查询无关特征
- 宏观上来讲，与查询相关的查询目标页面是否存在与查询无关的特征呢？

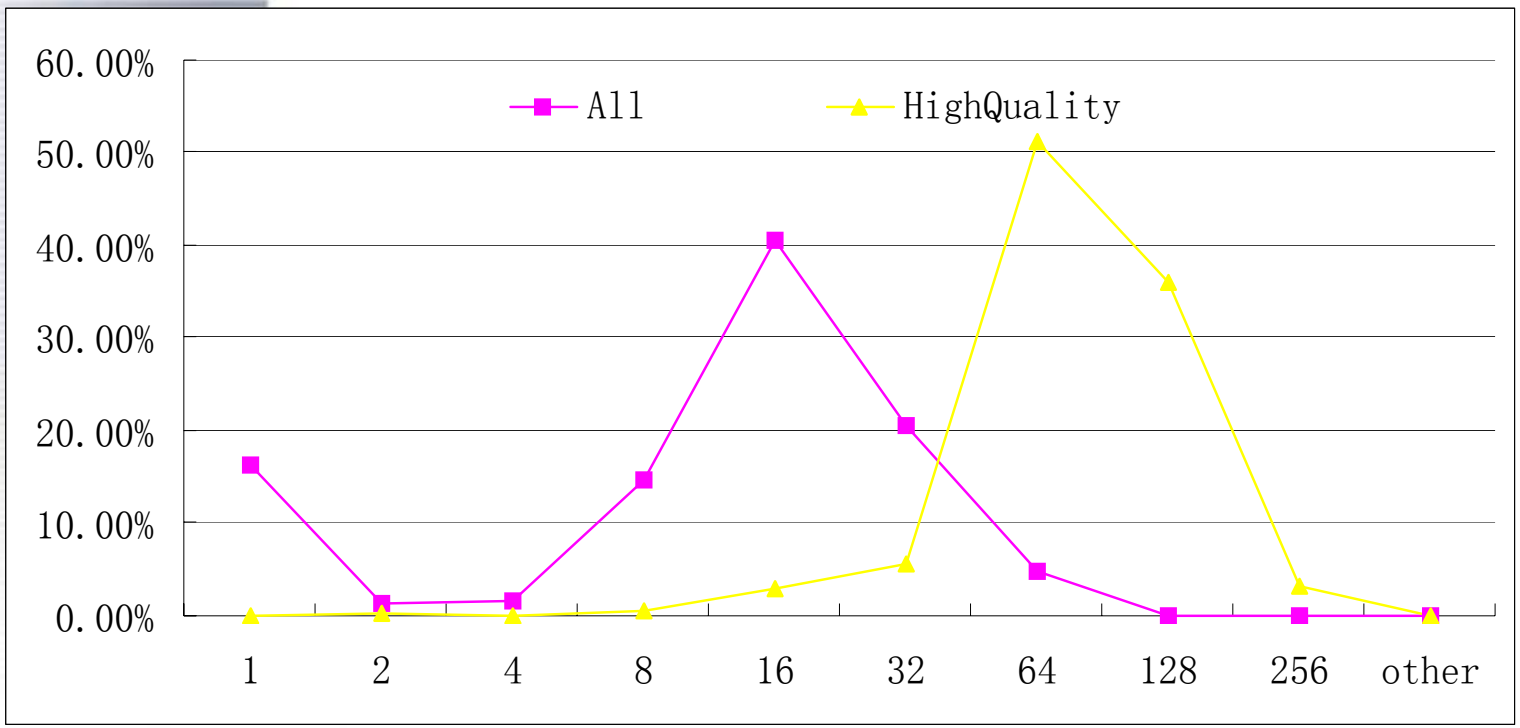
网络信息检索系统的构成及运行原理

- 页面质量评估：我们的做法
  - 根据在不同规模的网页数据集合，使用多个不同查询任务的查询目标页面的统计，查询目标页面确实存在查询无关特征
  - 在Sohu公司提供的超大规模训练数据集上进行的实验
    - 3700多万页面（是当前研究界所获得的最大规模真实网络页面数据）
    - 手工标注的上万个查询目标页面
    - 考察的特征包括PageRank, 入链接个数, 出链接个数, 链接文本长度, 页面镜像个数, 文档长度, 摘要长度, URL特征, Page Size等。

Sogou搜狗

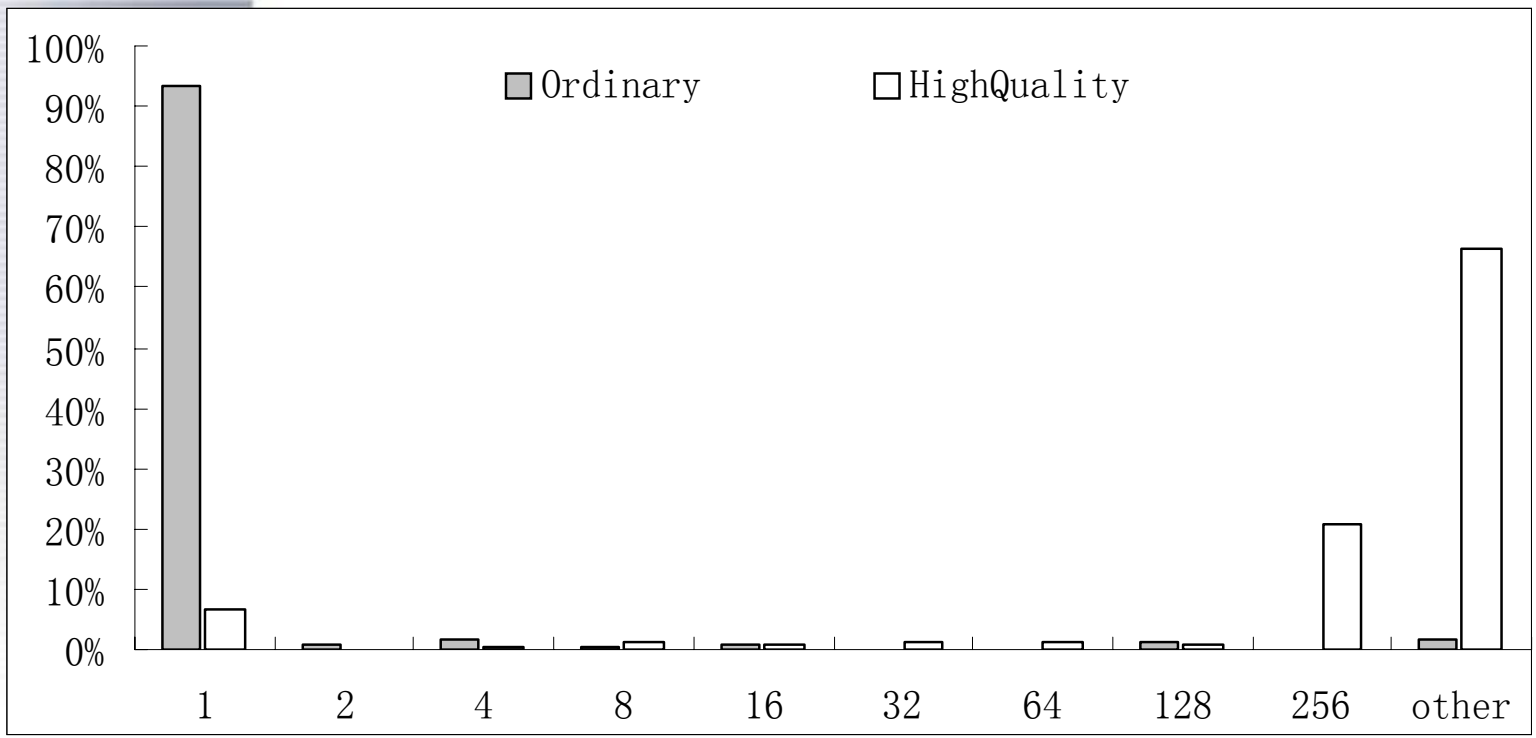
网络信息检索系统的构成及运行原理

- 页面质量评估：我们的做法
  - PageRank特征的分布



网络信息检索系统的构成及运行原理

- 页面质量评估：我们的做法
  - 利用贝叶斯方法计算出的查询目标概率



网络信息检索系统的构成及运行原理

- 页面质量评估：我们的做法
  - 测试集合：独立于训练集的近两万个高质量页面

	普通页面	查询目标页面训练集合	查询目标页面测试集合
算法判定出的低质量页面	95.04%	7.27%	7.63%
算法判定出的高质量页面	4.96%	92.73%	92.37%

  - 算法判定出的高质量页面仅占数据总量的5%，但能够满足超过92%以上的用户查询需求

网络信息检索系统的构成及运行原理

- 查询系统
  - 具有信息检索方面研究的积累，各个商业系统大同小异
  - 总的思路：内容相关度与页面质量评估相结合
  - 与传统信息检索系统中查询模块的不同
    - 页面质量评估的引入
    - 更加重视“精确匹配”的情况（与Web数据的庞大规模相关）
    - 半结构化信息的使用

网络信息检索系统的构成及运行原理

- 查询系统
  - 与传统信息检索系统不同的一些特殊技术
    - 查询提示技术

    - 查询修正技术

网络信息检索系统的构成及运行原理

- 查询系统
  - 与传统信息检索系统不同的一些特殊技术（续）
    - 简单的QA功能

网络信息检索技术的现状与挑战

清华大学
Tsinghua University

- 问题背景
- 网络信息检索系统的构成及运行原理
- 网络信息检索研究面临的挑战

网络信息检索研究面临的挑战

清华大学
Tsinghua University

- 2002年，Google公司的Monika R. Henzinger等人发表了“Challenges in Web Search Engines”一文，并在ICJAI03’进行了大会主题报告
- 文中提出了Spam, Content Quality, Quality Evaluation, Web Conventions, Duplicate Hosts, Vaguely-Structured Data等几个方面的挑战

网络信息检索研究面临的挑战

清华大学
Tsinghua University

- 在Henzinger提出这些挑战之后，搜索引擎技术也取得了迅猛的发展，有些挑战已经得到了一定的解决，例如Duplicate Hosts问题等，但又涌现出了更多的问题。
- 我们试图将Henzinger指出的挑战，连同新出现的挑战进行综合，重点从以下几个方面提出搜索引擎面临的问题：
  - SEO与SPAM的挑战
  - 多元数据融合的挑战
  - 评测方式的挑战

SEO与SPAM的挑战

清华大学
Tsinghua University

- 缘起
  - 注意力经济
    - 未来的因特网之争就是眼球之争 （Intel总裁葛洛夫）
    - CNNIC统计显示，搜索引擎是86.6%的用户得知新网站的主要途径。
    - 搜索引擎就是注意力的制高点，并在很大程度上决定了人们能够获取到哪些信息。
  - SEO：Search Engine Optimization
    - 排名等于商机

SEO与SPAM的挑战

- SEO的做法有哪些？
  - 用“合理”的方式改进网站在搜索引擎中的排名
    - 在域名中包含所要优化的关键字
    - 选择恰当个数与适当内容的关键词，把他们放置在适当的位置
    - 有序、合理安排文件目录结构，规范文件命名
    - 不要与质量低、而且还存在作弊的网站交换链接
    - 向所有能找到的相关网站目录提交你的网站
    - 围绕目标关键词在一些顶级站点的电子杂志或资源区里发表文章，并在其中引用你的链接。

SEO与SPAM的挑战

- SPAM的作法有哪些？
  - 早期
    - 堆砌关键词
    - 隐藏文字
    - Link Farm
  - 当前
    - 重定向作弊
    - 利用Blog，Wiki作弊
    - …

SEO与SPAM的挑战

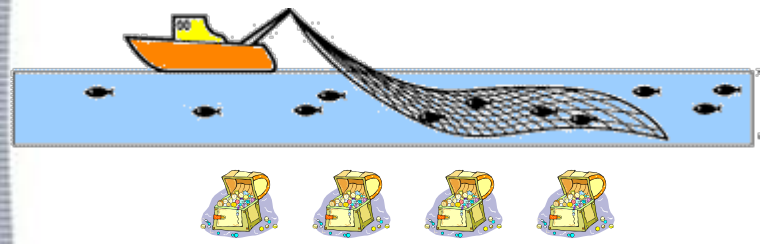
- 搜索引擎与Spam的战争
  - 事关用户能否高效的定位有用信息
    - 从这个角度讲，SEO对搜索引擎及其用户的影响不亚于Spam
  - 道高一尺，魔高一丈
    - 可能比病毒与反病毒软件的斗争更加激烈与长久
  - 必须由页面抓取系统、页面质量评估系统、查询系统多方面共同完成，而核心是达到查询无关的页面质量评估
    - 定位高质量，还是丢弃低质量？

多元数据融合的挑战

- 网络数据多元化的趋势愈演愈烈
  - 纯文本数据：文本，代码等
  - 半结构化数据：HTML, XML等
  - 结构化数据：表格，数据库数据（Deep Web）等
  - 二进制数据文件的多元信息表示
    - MP3，软件，图片
    - 检索功能多依赖于描述文字与二进制文件自身的特性相结合

多元数据融合的挑战

- 困难1：Deep Web数据的获取
  - Deep Web的含义



  - 数据规模：Surface Web规模的400-550倍
  - 访问量：是Surface Web站点访问量的1.5倍
  - 数据质量更高
  - 结构性更强

多元数据融合的挑战

- 困难2：如何将各种结构数据结合以向用户反馈结果
  - 应用假设：购买《纳尼亚传奇》
    - 各个购物站点的报价、版本信息（结构化），某些论坛的二手货相关帖子
    - 作者刘易斯的个人介绍、书籍的介绍（半结构化）
    - 与《纳尼亚传奇》相关的电影信息、其他书籍信息等（半结构化及结构化）
  - 如何组织这些信息，又应该是一个怎样的顺序反馈给用户？

多元数据融合的挑战

清华大学
Tsinghua University

- 困难3：对于二进制数据文件检索而言，如何将不同结构化的信息结合成对某个文件的统一描述
  - 应用假设：MP3检索“光良《约定》”
    - Mp3文件本身记录的元信息，艺术家，歌名乃至歌词等（结构化）
    - Mp3文件所在网页的信息（半结构化）
    - Mp3文件所在站点的信息（结构化）
  - 如何统一利用这些信息对这个Mp3文件给出合适的顺序？

评测方式的挑战

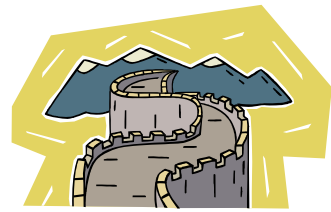
- 孰优孰劣，怎样评价？
  - 对于信息检索系统的评价，研究界有一套比较完整的理论与应用指标
    - 基于Precision, Recall的指标体系：前n位精度，前n位成功率，平均精度等
    - 这套体系需要基于一个“标准答案”集合，难于直接应用于网络信息检索系统/搜索引擎评价
      - TREC的评测，pooling方法标注答案
      - 标注与某个关键词相关的“词项”而不是页面（IBM Haifa）
    - 真的能够全面反映出搜索引擎的性能么？

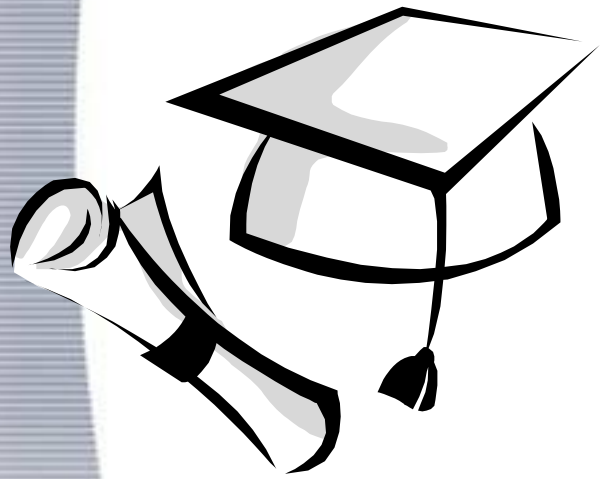
评测方式的挑战

- 搜索引擎评价中一些独特的考虑因素
  - 时效性
  - 死链率
  - 重复度
  - 垃圾比率
- 核心目的：发现用户是否对搜索结果满意
  - 依赖于对用户行为的考察与分析
  - 搜索引擎日志挖掘

评测方式的挑战

- 问题：对用户行为分析的再认识
  - 用户行为的歧义性
    - 用户点击了某个网页，就代表他满意这条检索结果么？
    - 用户进行若干次点击后，关闭检索结果页面，就代表他对检索结果很满意么？
    - 用户的点击行为，是否与他查询问题的类型相关呢？（“清华大学”，“长城”）





Thank you!

Questions or comments?