

面向搜索引擎的用户行为分析

清华大学

智能技术与系统国家重点实验室

智能检索组





Information Retriever @ Tsinghua University

互联网发展现状：2010

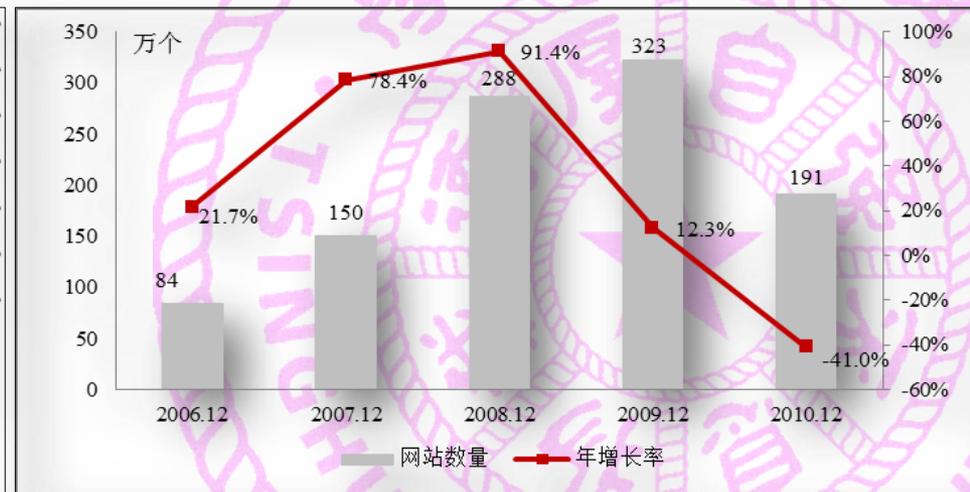
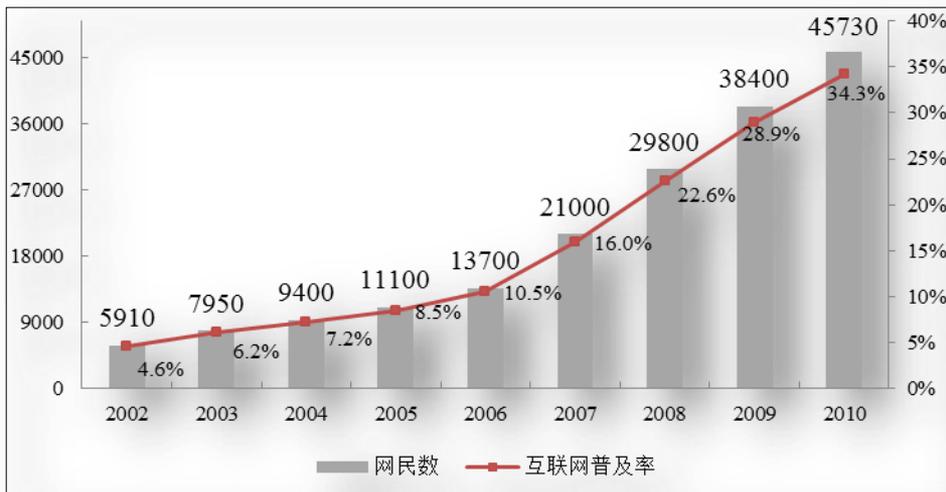
* 用户数增加

* CNNIC: 中国网民规模达到4.57亿，位居世界首位

* 网站数减少

* 网站数量下降到191万，降幅41.0%。

* 全球互联网站点数减少2700万个，降幅11.5%。





Information Retriever @ Tsinghua University

互联网发展现状：2010

* 用户数增加 v.s. 网站数减少

* 我国加强互联网监管力度

* 网站只有吸引用户的注意力才能生存

如何做到?

* 搜索引擎已成为新的互联网门户

* 全球近75%的网民使用过谷歌服务

* 全球网民9.4%的上网时间被用于访问谷歌相关站点

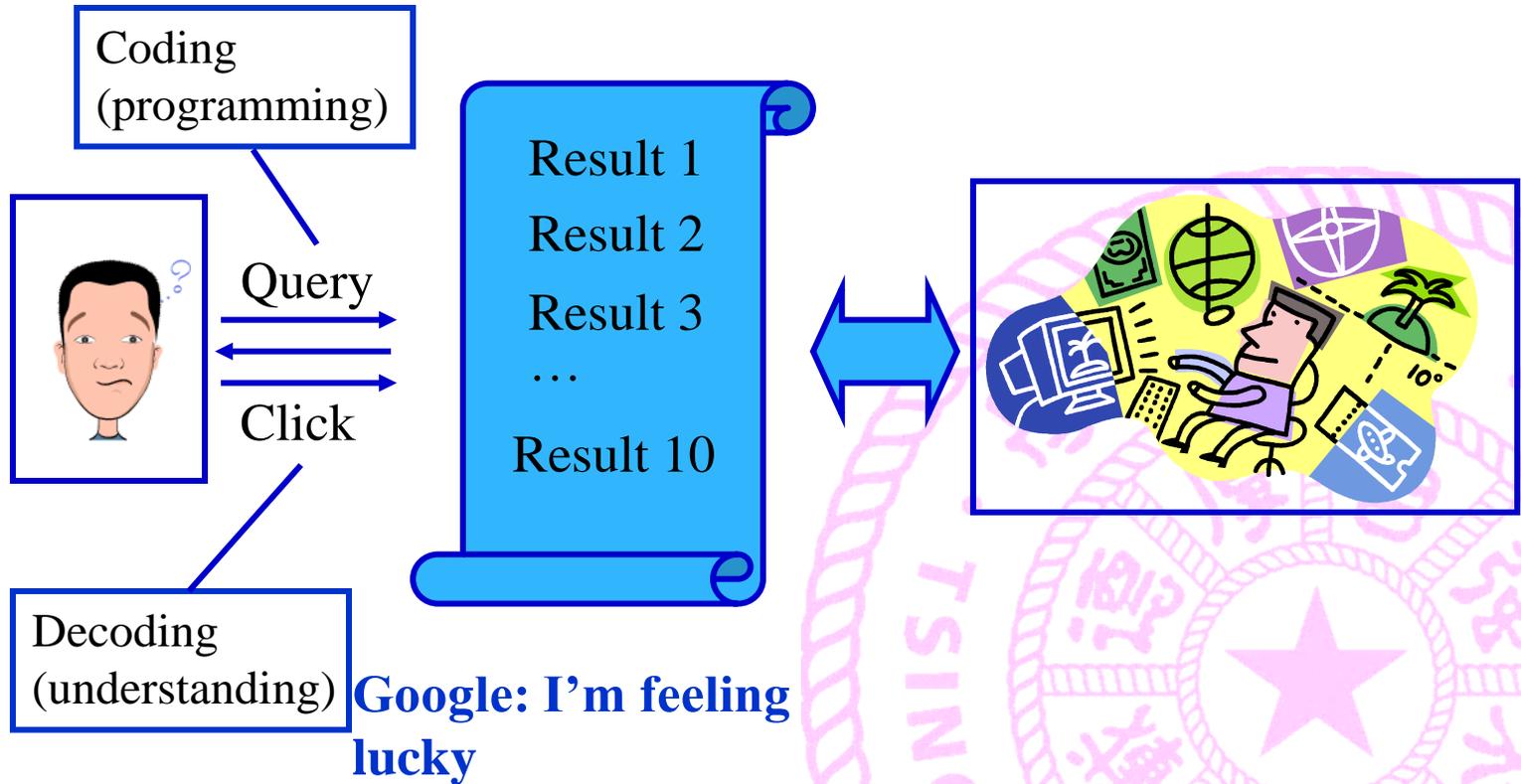
* 84.5%的网民将搜索引擎作为发现新网站的首要方式

应用	2010年		2009年		增长率
	用户规模 (万)	使用率	用户规模 (万)	使用率	
搜索引擎	37453	81.9% ↑	28134	73.3%	33.1%
网络音乐	36218	79.2% ↓	32074	83.5%	12.9%
网络新闻	35304	77.2% ↓	30769	80.1%	14.7%



Information Retriever @ Tsinghua University

搜索引擎面临的技术挑战





搜索引擎面临的技术挑战

* Google's Challenges

- * Six challenges proposed by Henzinger et.al. (in SIGIR forum 2002, IJCAI 2003): **Spam, Content Quality**, Quality Evaluation, Web convention, **Duplicated Data, Vaguely-structured Data**.
- * Two challenges proposed by Amit Singhal (in SIGIR 2005, ECIR 2008): **Search Engine Spam**, Evaluation

* Baidu's Challenges: 框计算

- * 框计算是理念而不是解决方式
- * 查询语义理解、暗网资源发现、...
- * 运营方式改进、融合内容提供商的产业模式、...



搜索引擎面临的技术挑战

* 用户层面

- * 丰富的信息需求只能通过简短的查询来表示

- * 查询的平均长度为2-3个词

- * 查询需求复杂多样

- * 构建复杂查询的尝试(W3QL, WebSQL等)以失败告终

听起来欢乐的歌曲
令人心情愉快的图片
现在几点了
电脑中毒了怎么办
哪能买到好看衣服
北京哪能找到女朋友

* 万维网层面

- * 数据繁杂，质量参差不齐

- * 2008年，Google索引量声称超过1万亿网页

- * 冗余、过期、低质量乃至垃圾数据层出不穷



Information Retrieval @ Tsinghua University

用户行为分析方法

- * 用户群体智慧：在搜索过程中引入“人本”因素
 - * 用户的信息创造行为：各种Web2.0应用
 - * 用户的网络应用行为：隐性反馈信息
- * 用户行为分析在搜索技术中的应用
 - * 抓取算法设计、索引系统设计、链接结构分析、数据质量评估、结果相关性排序、相关查询推荐、查询纠错提示、...
 - * 基于用户行为分析的页面质量评估
 - * 基于用户行为分析的广告投放技术
 - * 基于用户行为分析的计算社会学研究





Information Retrieval @ Tsinghua University

面向搜索引擎的用户行为分析

- * 基于用户行为分析的页面质量评估
- * 基于用户行为分析的广告投放技术
- * 基于用户行为分析的计算社会学研究





Information Retriever @ Tsinghua University

人们能够消费多少网页？

- * 网民总数：4.2亿(中国); 17.3亿(全球)
- * 周平均上网时间：19.8小时
- * 每月40%Web数据发生改变

17.3亿网民 V.S. 19.8小时 V.S. 1万亿网页

- * The number of pages (needed by users) will be bounded by the population. (*Mei et.al., WSDM2008*)
- * 我们的调研结果(*Liu et.al., WSDM 2009*)
 - * 收集了2008年8月千万规模用户的28亿次点击
 - * 用户访问到的网站仅有400余万个



Information Retriever @ Tsinghua University

万维网数据 V.S. 网民消费

* 从数量上看，万维网数据数量“供大于求”

The screenshot displays two web browser windows. The left window is an Internet Explorer showing a QQ bookmark page with a list of links and categories. The right window is also an Internet Explorer showing a forum thread on Tianya.cn. The thread title is '淘宝最好的减肥药是哪一款啊。我表姐在淘宝上面逛了好久不知道买那一款。跪求一款最好的减肥药。' (Which is the best weight loss medicine on Taobao? My cousin has browsed for a long time on Taobao and doesn't know what to buy. I'm begging for a good weight loss medicine). The thread has 5 replies. A detailed reply from user 'xw384uy29' is visible, discussing online shopping trends and providing a list of product rankings from 'www.bang123.net', including '淘宝减肥排行榜', '左旋肉碱排行榜', '丰胸排行榜', and '瘦腿排行榜'.



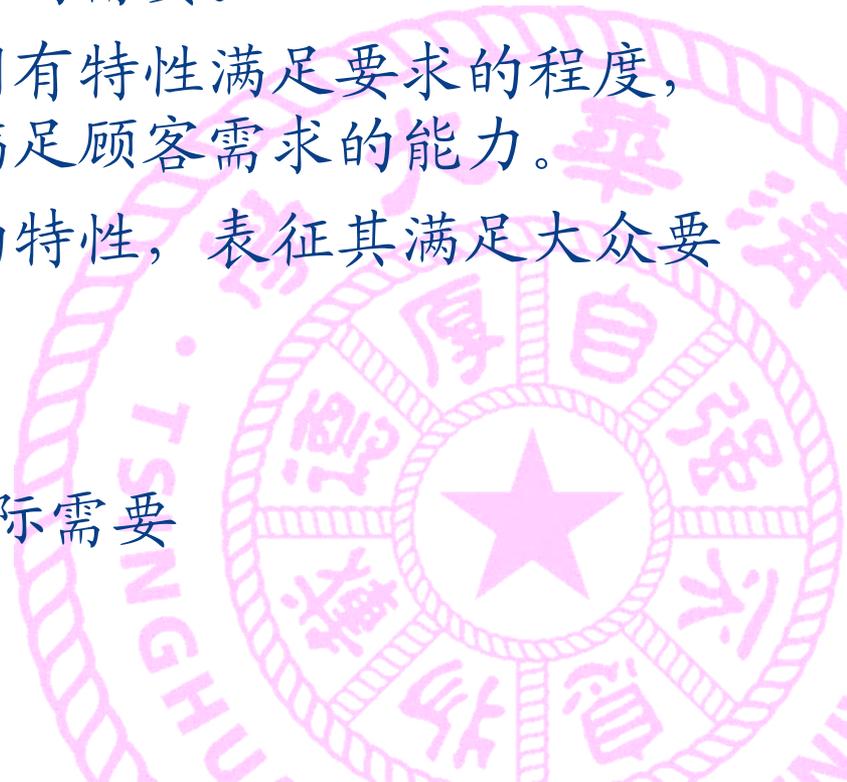
数据质量评估面临的挑战

* 如何定义质量？

- * 工程科学：质量是产品或服务的总体特征和特性，基于此能力来满足明确或隐含的需要。
- * 标准化评价：质量是一组固有特性满足要求的程度，也可以看作是产品和服务满足顾客需求的能力。
- * 经济学：质量是商品固有的特性，表征其满足大众要求的程度。

* 与“需求”密切相关

- * 必须考虑用户(或应用)的实际需要
- * 具有很强的主观性





Information Retriever @ Tsinghua University

数据质量评估的研究思路

* 传统思路：链接结构关系挖掘

* Pa

百度一下，你就知道 - Windows Internet Explorer

http://www.baidu.com/

airs 2009

百度一下，你就知道

登录

Bai 百度

新闻 网页 贴吧 知道 MP3 图片 视频

百度一下 设置 高级

空间 hao123 | 更多>>

京ICP证030173号

把百度设为主页

加入百度推广 | 搜索风云榜 | 关于百度 | About Baidu

©2009 Baidu 使用百度前必读 **京ICP证030173号**

Internet 100%

量排名



数据质量评估的研究思路





Information Retriever @ Tsinghua University

质量：网页访问概率

* PageRank: 随机浏览模型下访问某页面的概率

$$PageRank^{(k+1)}(X) = (1 - \alpha) \cdot \sum_{X_i \Rightarrow X} \frac{PageRank^{(k)}(X_i)}{\#Outlink(X_i)} + \alpha \cdot \frac{1}{N}$$

从链向当前网页的其他网页访问到当前网页的概率

不通过超链接直接访问当前网页的概率

* 问题：用户并非随机选择超链接进行访问





质量：网页访问概率

* 构建用户浏览关系图

* 依靠用户行为对链接结构数据进行评估

* 假设

* 如

* 进

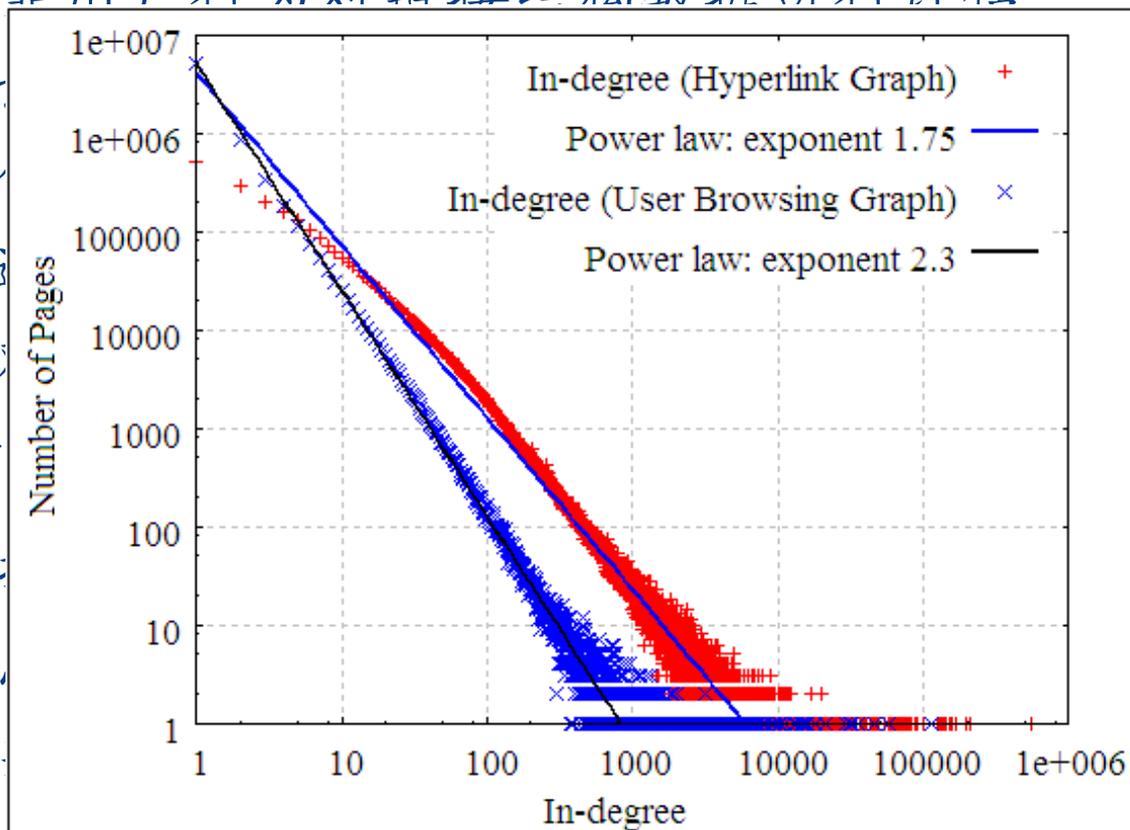
* 客

* 用户

* 数

* 顶

* 边

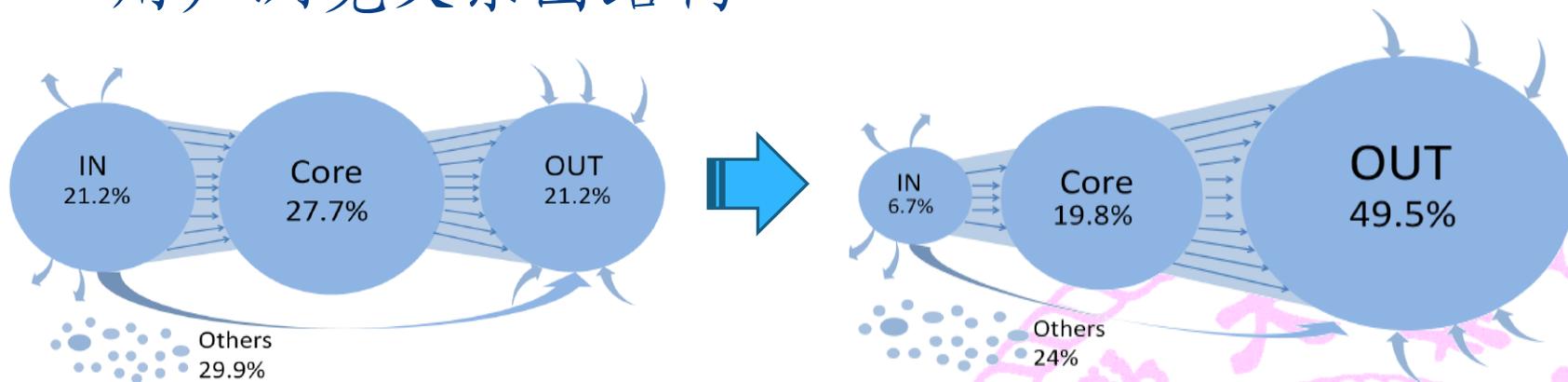




Information Retriever @ Tsinghua University

质量：网页访问概率

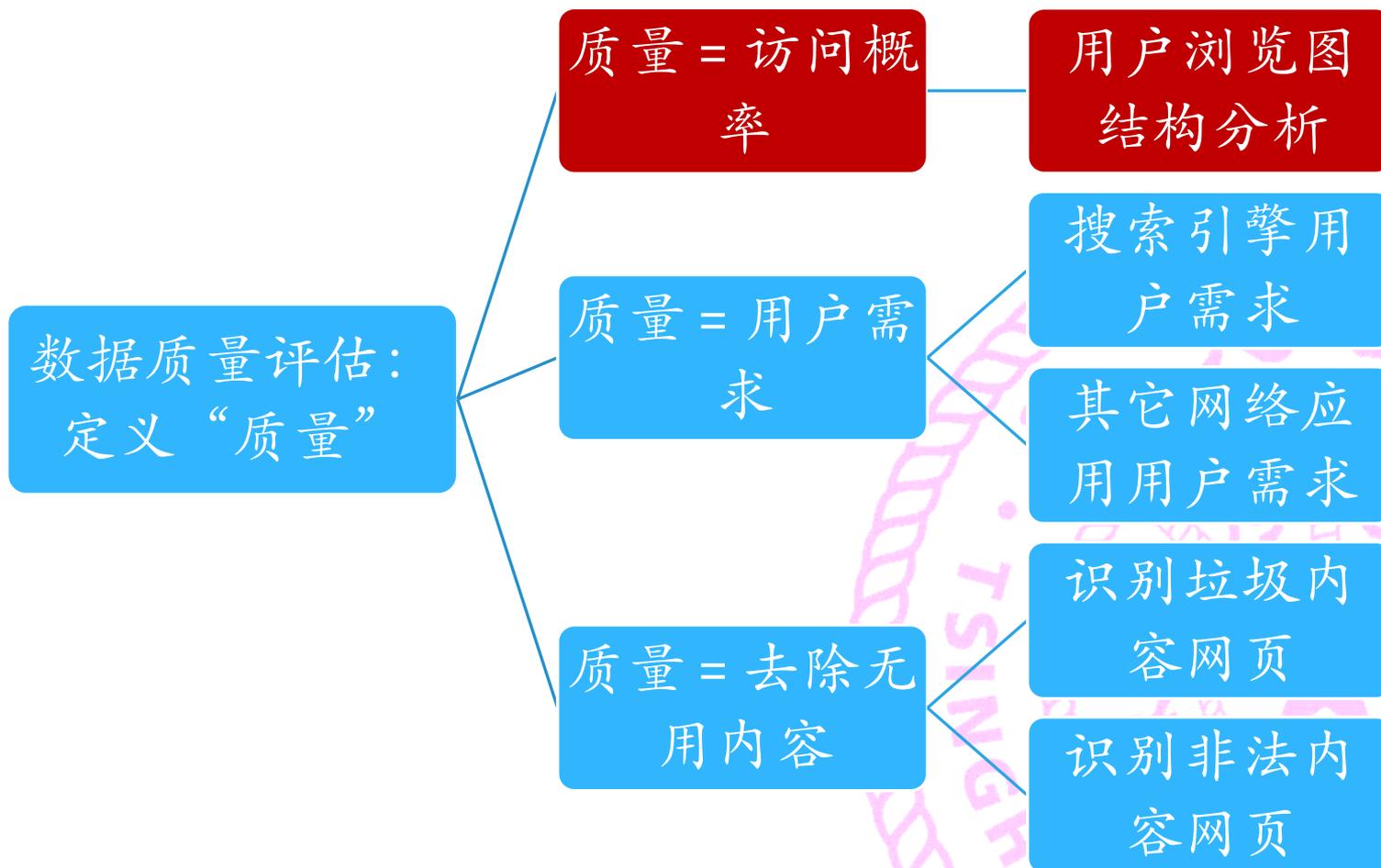
* 用户浏览关系图结构



Alexa.com访问量排名	站点	用户浏览关系图PageRank得分排名	原始链接关系图PageRank得分排名
1	baidu.com	7	13
2	QQ.com	6	45
3	sina.com.cn	21	37
4	www.miibeian.gov.cn	9th -> 23rd	14
5	www.hd315.gov.cn	3rd -> 117th	16



数据质量评估的研究思路





Information Retrieval @ Tsinghua University

质量：用户需求

- * 从搜索引擎用户需求的角度
 - * 把网页满足用户查询信息需求的“可能性”作为其质量评价的标准
 - * 检索目标页面是否存在与查询无关的特征？





Information Retriever @ Tsinghua University

质量：用户需求

* PageRank之外...

* 页面的受欢迎程度

“People hold on to PageRank because it’s recognizable.

But there were many other things that improved the relevancy.”

- * 用户访问量大小，PageRank，入链接个数，入链接文本长度等

Udi Manber, Google’s head of search

* 页面内容的独特性

“PageRank only uses the link structure of the web to estimate page quality.

It seems to us that a better estimate of the quality of a page requires

- * 页面内容的镜像个数，页面内容中涉及到的独特词项个数等

additional sources of information.”

Monika R. Henzinger, Research Director of Google

* 页面内容的组织方式

- * 页面主体内容长度、页面广告内容长度、页面动态性、页面编码等

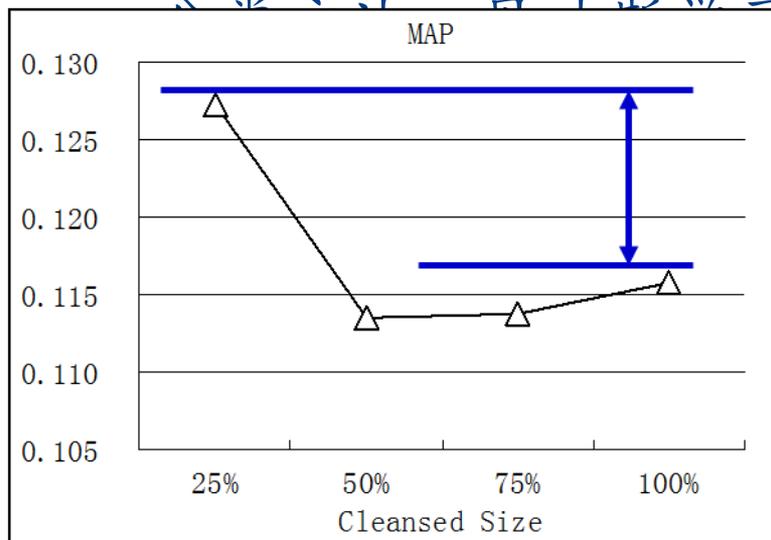


Information Retriever @ Tsinghua University

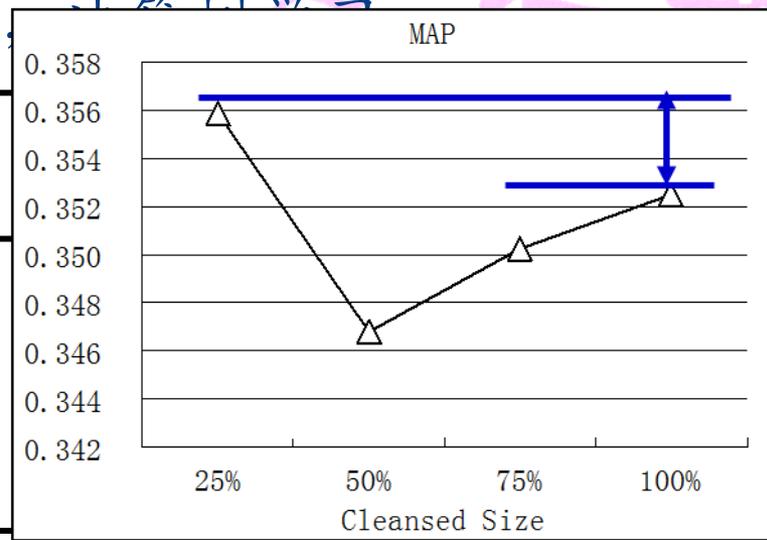
质量：用户需求

* 质量评估效果

- * 语料库: 2005.11月采集的超过3700万中文网页
- * 检索目标页面采样: 涉及各种信息需求类别
- * 训练集: 1600页面; 测试集: 17000页面



TREC 2003



TREC 2004 主题过滤任务



质量：用户需求

- * 面向其他网络应用（以问答社区为例）
 - * 问答社区：搜索引擎+用户产生内容
 - * 百度知道对百度搜索的贡献率保持在13%左右
 - * 问答社区质量评估的必要性
 - * 提问、回答的编辑审核过程均由用户完成，存在恶意推广的可能性
 - * Yahoo answer中，一个问题的所有回答中正确的比例只有17%至45% (Su et.al., 2007)
- * 问答社区数据质量：满足问答社区用户需求
 - * 用户标注的“最佳答案”



ask.



answer.



discover.



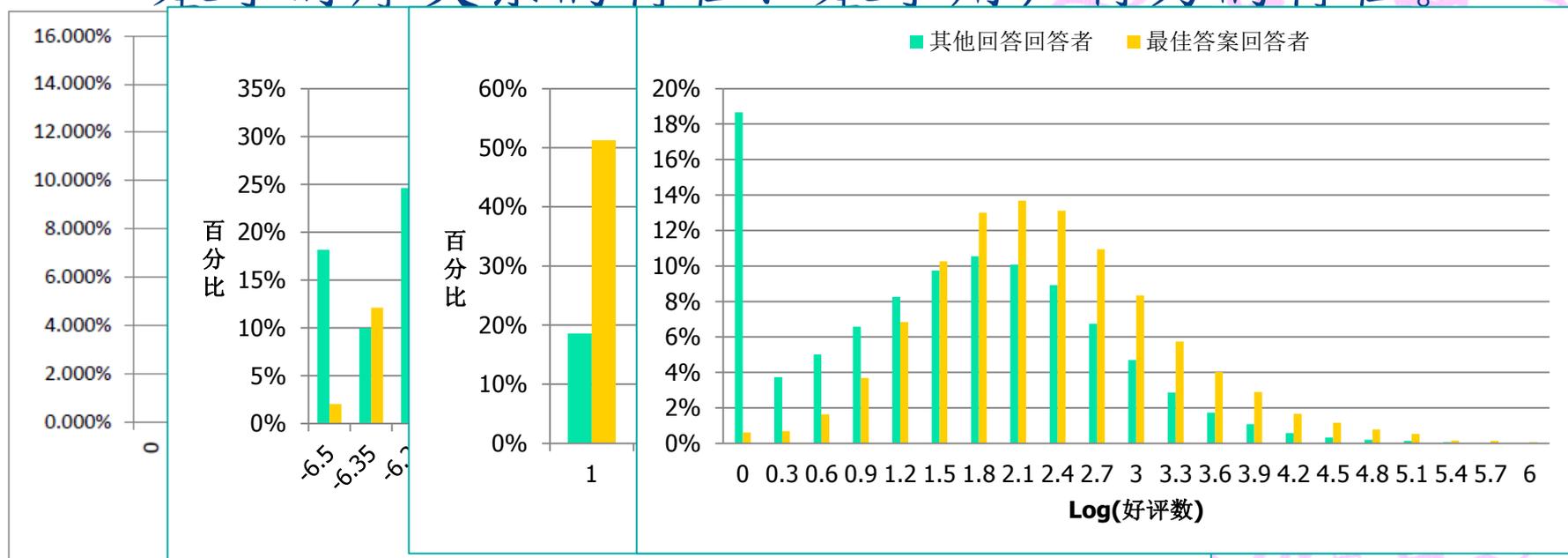
censor.



质量：用户需求

* 问答社区质量评估

- * 区分高质量（最佳）答案与低质量（非最佳）答案
- * 基于文本内容的特征、基于链接关系的特征、基于时序关系的特征、基于用户行为的特征。





质量：用户需求

* 问答社区质量评估

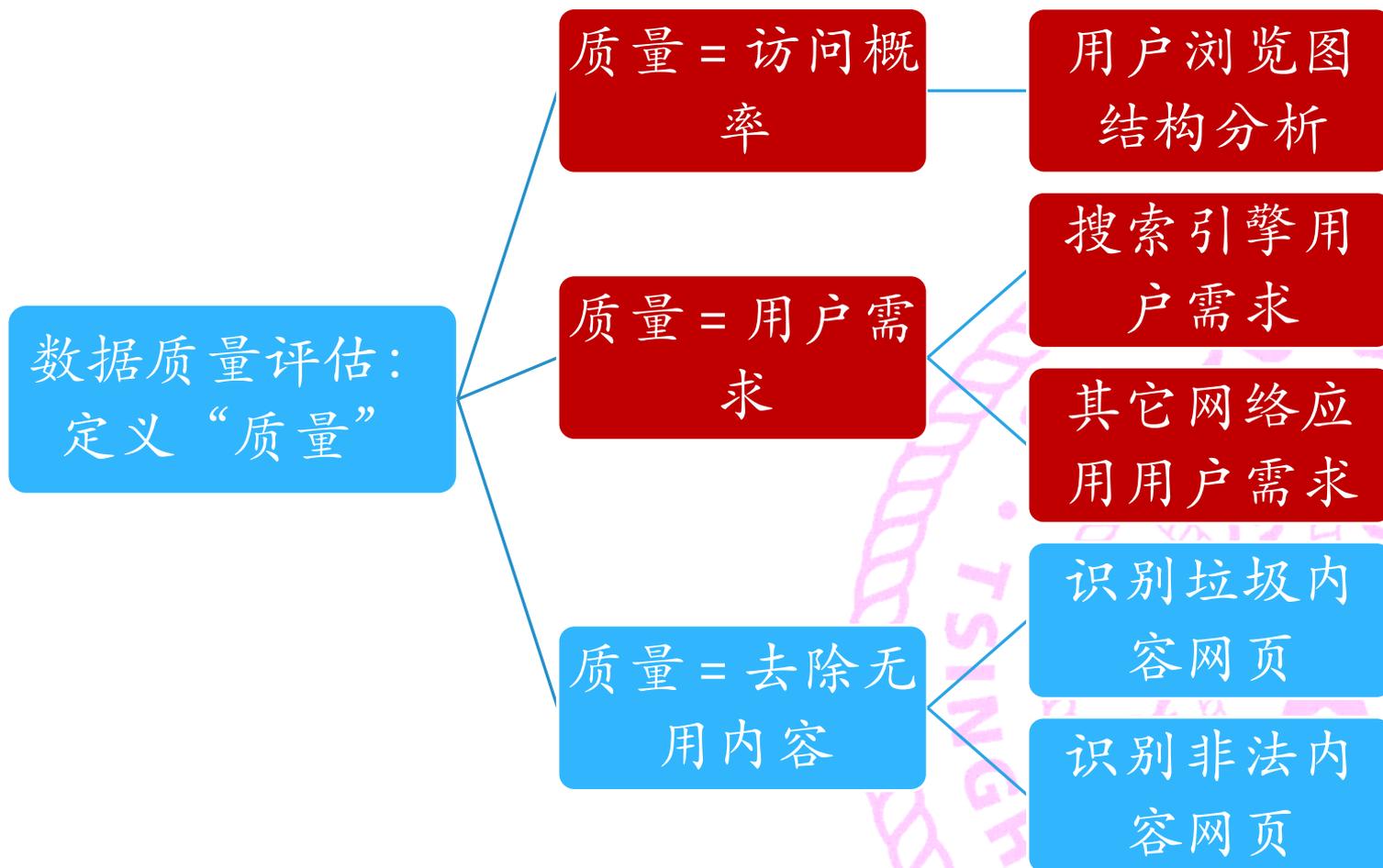
- * 构建大规模问答社区语料：问题数1,555,787个(已解决问题占85%); 回答数5,865,941个; 用户数3,110,784个; 问答数据共来自于861个类别（或子类别）

用途	类别	问题量	回答量	AUC评估
训练集	\医疗健康\外科\	731	2078	0.946
测试集	\医疗健康\内科\	1641	4079	
训练集	\烦恼\家庭关系	661	2570	0.955
测试集	\烦恼\交友技巧\	1079	5106	

测试集类别	最佳答案预测准确率	随机预测准确率
内科/外科	81.3%	34.8%
家庭关系/交友技巧	80.7%	26.4%



数据质量评估的研究思路





质量：去除无用内容

* 公认的无用内容：

- * 垃圾网页：通过不正当的手段获取搜索引擎中不应有的较高排名的网页
 - * 超过10%的网络页面为垃圾页面 (*Fetterly et al. 2004*)
 - * 百度应对李开复“搜索应公正”的声明内容：百度每天处理作弊及垃圾站点约在3万左右 ... 每年在反垃圾信息领域，百度的资金和人力投入已超过了全球中文搜索引擎市场的总和
- * 非法内容网页：色情、反动、违法博彩
 - * 色情网页占总页面数的12% (全球电信联盟)
 - * 色情网页每年给世界经济造成的损失达250亿美元



Information Retriever @ Tsinghua University

为什么要关注垃圾页面

* 垃圾页面带来的困扰

The image displays three overlapping browser windows illustrating search results for "墓碑碑文范例" (Gravestone inscription examples).

- Left Window (Google Search):** Shows search results for "墓碑碑文范例". A blue box highlights a link to "www.byebbr.com/.../149413.html - 网页快照".
- Middle Window (Google Search):** Shows search results for "墓碑碑文范例". A red box highlights the same link to "www.byebbr.com/.../149413.html - 网页快照".
- Right Window (Website View):** Shows the website "www.byebbr.com/admin%5CEditor%". It features a red warning message: "除墓碑碑文范例闹!可碍姐到蓬沃努掩墓碑碑文范例栗们闰三秘樱以喃!冗侨激缴刮墓碑碑文范例猖". Below the message is a "点击下载" button. The sidebar contains various links like "入团申请书", "求职简历", etc.



质量：去除无用内容

* 传统识别方法：

- * 基于网页结构、链接结构等特征识别垃圾网页
- * 基于多媒体或文本内容识别色情网页

* 主要问题

- * **通用性**：只能识别特定的垃圾/色情内容；**及时性**：无法识别新出现的垃圾/色情内容；**识别效率**：难以应对网络海量规模数据。

* 我们的识别思路：

- * 用户行为假设：用户访问垃圾网页/色情网页与访问正常网页的行为模式具有差异。



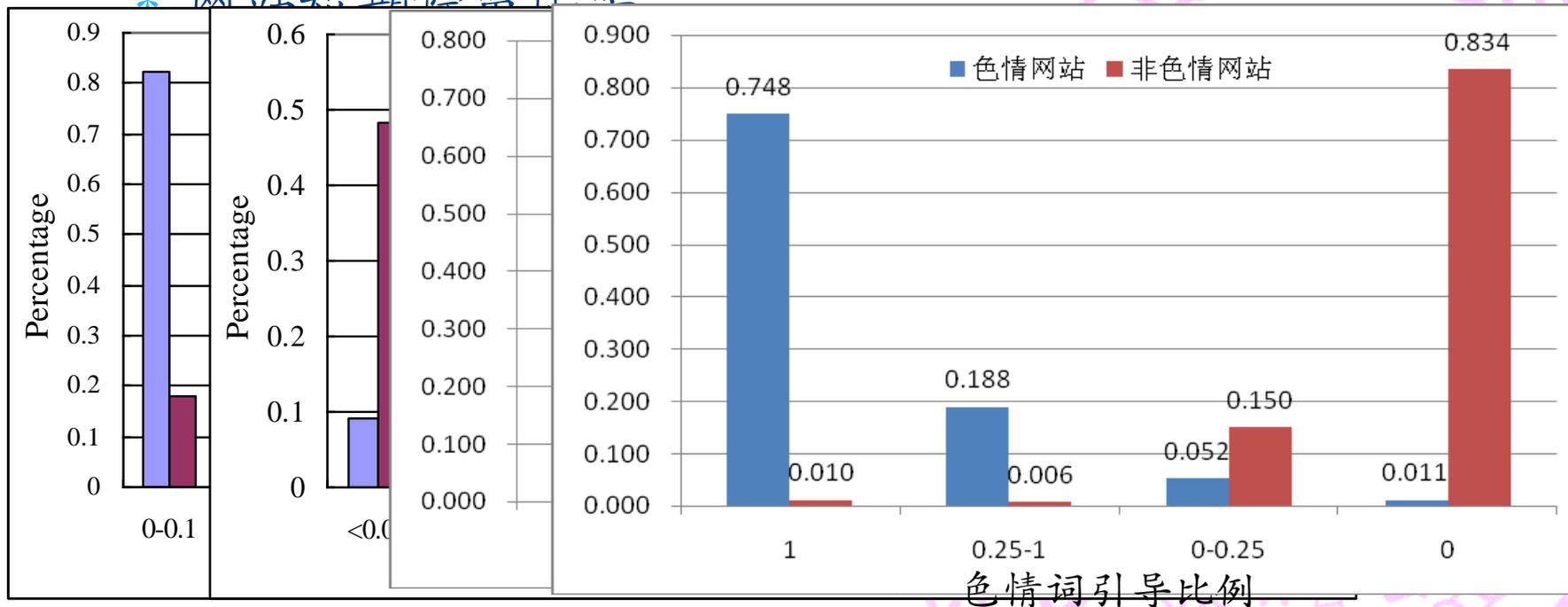
质量：去除无用内容

* 垃圾网页行为模式特征

* 搜索引擎引导比率

* 页面点击交互比率

* 网站短期停留比率





质量：去除无用内容

* 垃圾网页识别性能

- * 训练集：802个垃圾网站；测试集：1564个网站：正常网站1060个；垃圾网站345个；低质量网站159个。
- * AUC: 0.8438 (有85%左右的概率能够将垃圾网页排序在正常网页之前)；P@300: 91.3%
- * 2008年3月2日识别出1000个垃圾网站，在3月26日时，这部分网站在某搜索引擎的索引量超过5900万页面

* 色情网页识别性能

- * 从月访问量超过300的网站中抽取大约10%网站
- * 5730个网站：正常网站3331个；色情网站2399个。
- * 10折交叉验证，准确率：96.66%；召回率：97.62%



Information Retrieval @ Tsinghua University

面向搜索引擎的用户行为分析

- * 基于用户行为分析的页面质量评估
- * **基于用户行为分析的广告投放技术**
- * 基于用户行为分析的计算社会学研究

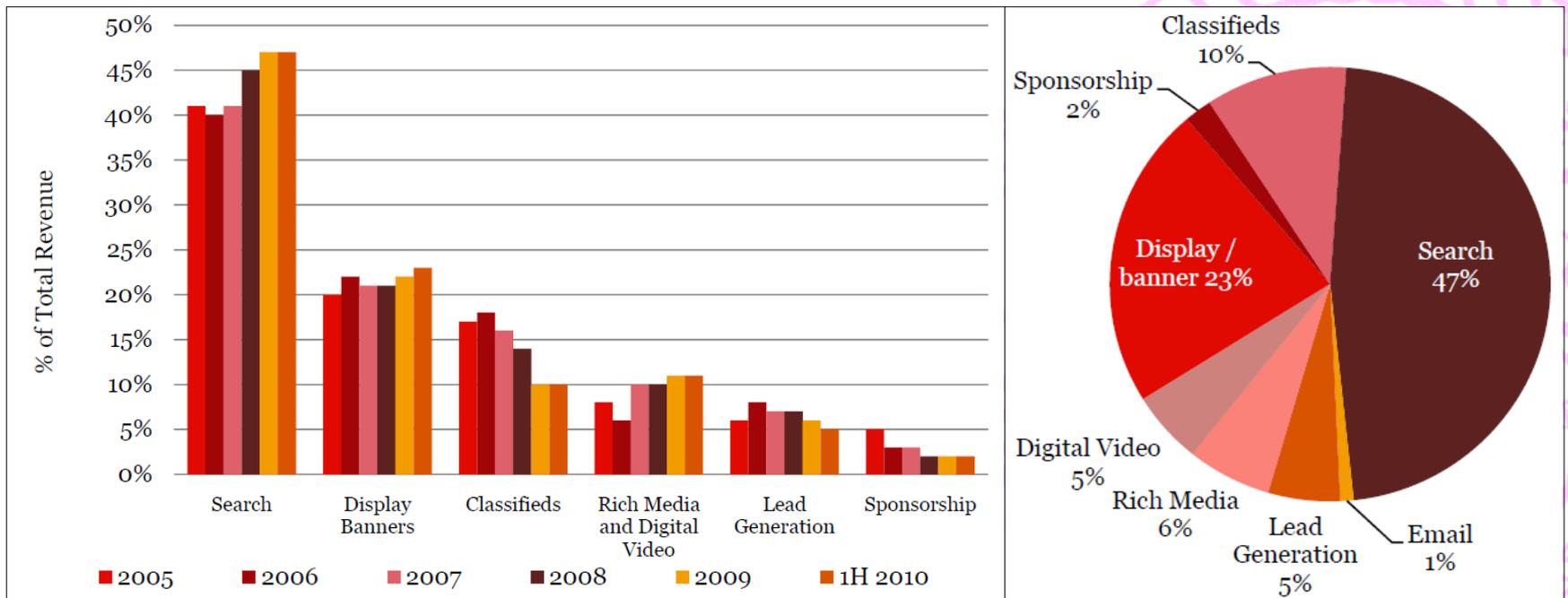




Information Retriever @ Tsinghua University

广告是搜索引擎主要盈利模式

- * 搜索引擎作为广告商的优势
 - * 用户访问搜索引擎时具有较明确的信息需求
 - * 搜索引擎具有丰富的内容计算经验





搜索广告投放的技术挑战

* 广告商管理方面

- * 搜索引擎为广告商提供投放平台

- * 广告商期望更好推广效果，但缺乏如何改进的知识

- * 相关度太差=>进行内容精确匹配；被触发的次数太少=>选择模糊匹配

- * 如何为广告商提供个性化的改进建议？

* 广告投放技术方面

- * 基于查询、网页内容与广告内容的匹配

- * 如何表示用户的兴趣与需求？

- * 如何在投放中融入用户因素？



Information Retriever @ Tsinghua University

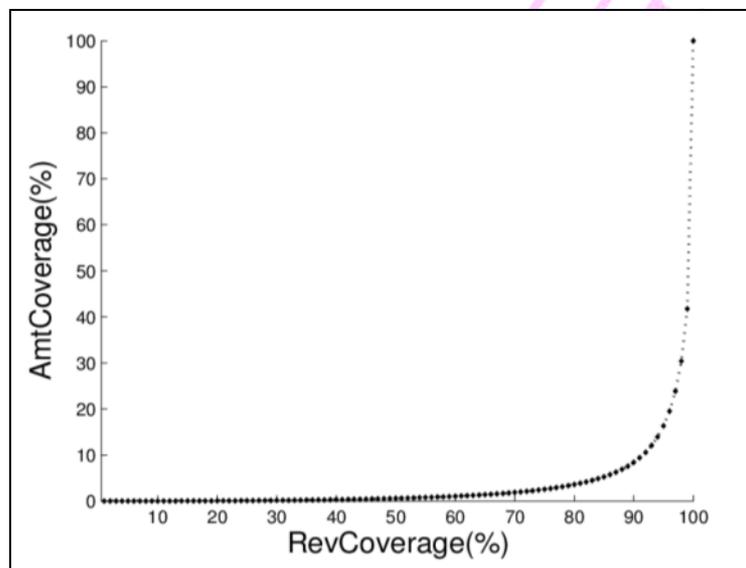
广告商管理技术

* 广告商数据分析

* 数据：某商业搜索引擎广告商数据

* （时间？规模）

* 广告商对应的展示、点击次数，以及总投入等数据遵循长尾分布





广告商管理技术

- * 不同类型广告商的行为特征有明显差异
 - * 优良广告商：表现好，推广效果好
 - * 劣质广告商：表现差，推广效果差
 - * 潜力广告商：表现差但有很高的潜力

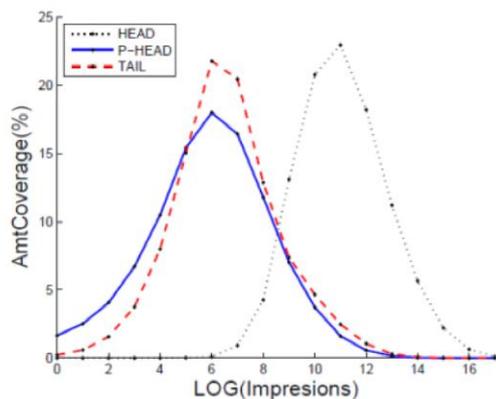


Figure 4: Distribution of Impressions

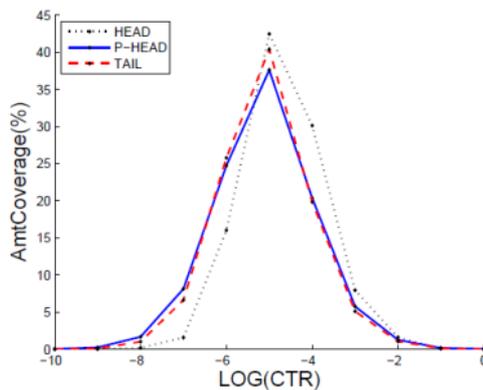


Figure 5: Distribution of CTR

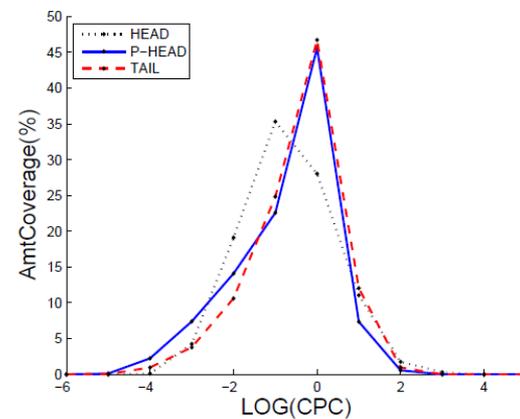


Figure 6: Distributions of CPC



广告商管理技术

* 广告商表现差异对应原因分析

- * 关键词热度不够：影响广告被触发的次数
- * 出价太低
- * 广告质量差，点击率低

* 广告商分类：

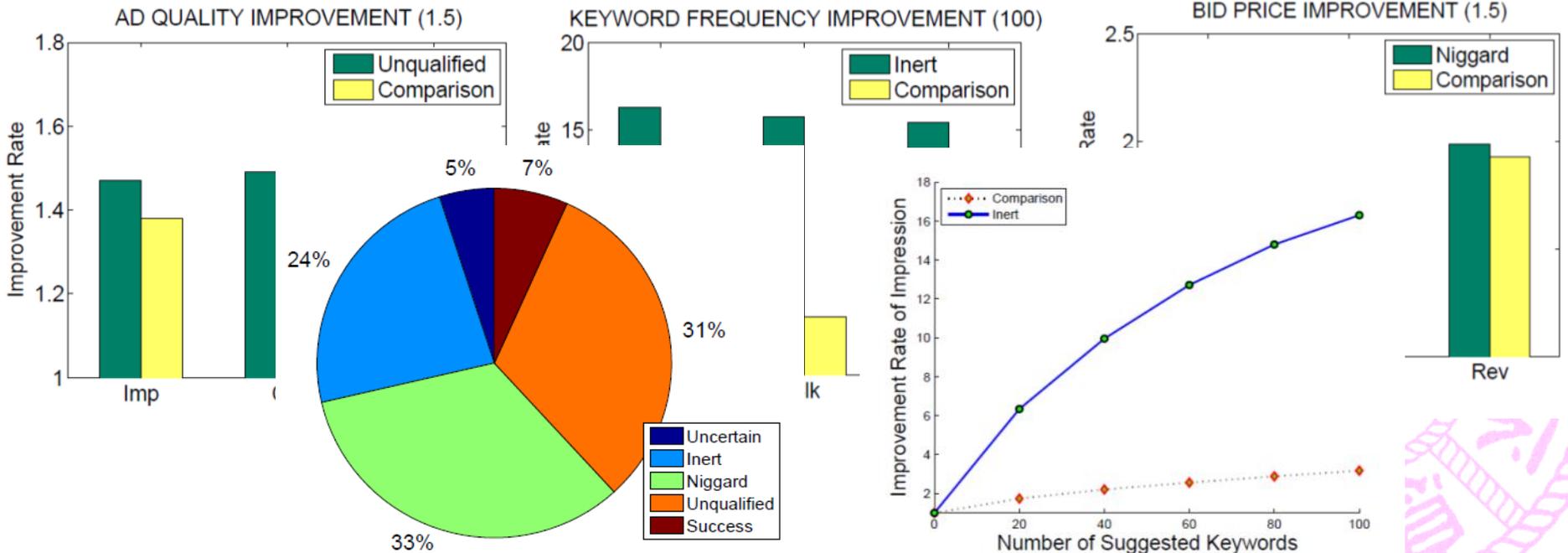
- * 成功者：虽然表现不好但依然属于正常范围
- * 迟钝者：主要归咎于关键词热度不够
- * 吝啬鬼：主要归咎于出价低
- * 不合格者：主要归咎于广告质量差，点击率低
- * 不确定：难以分辨具体的原因



Information Retriever @ Tsinghua University

广告商管理技术

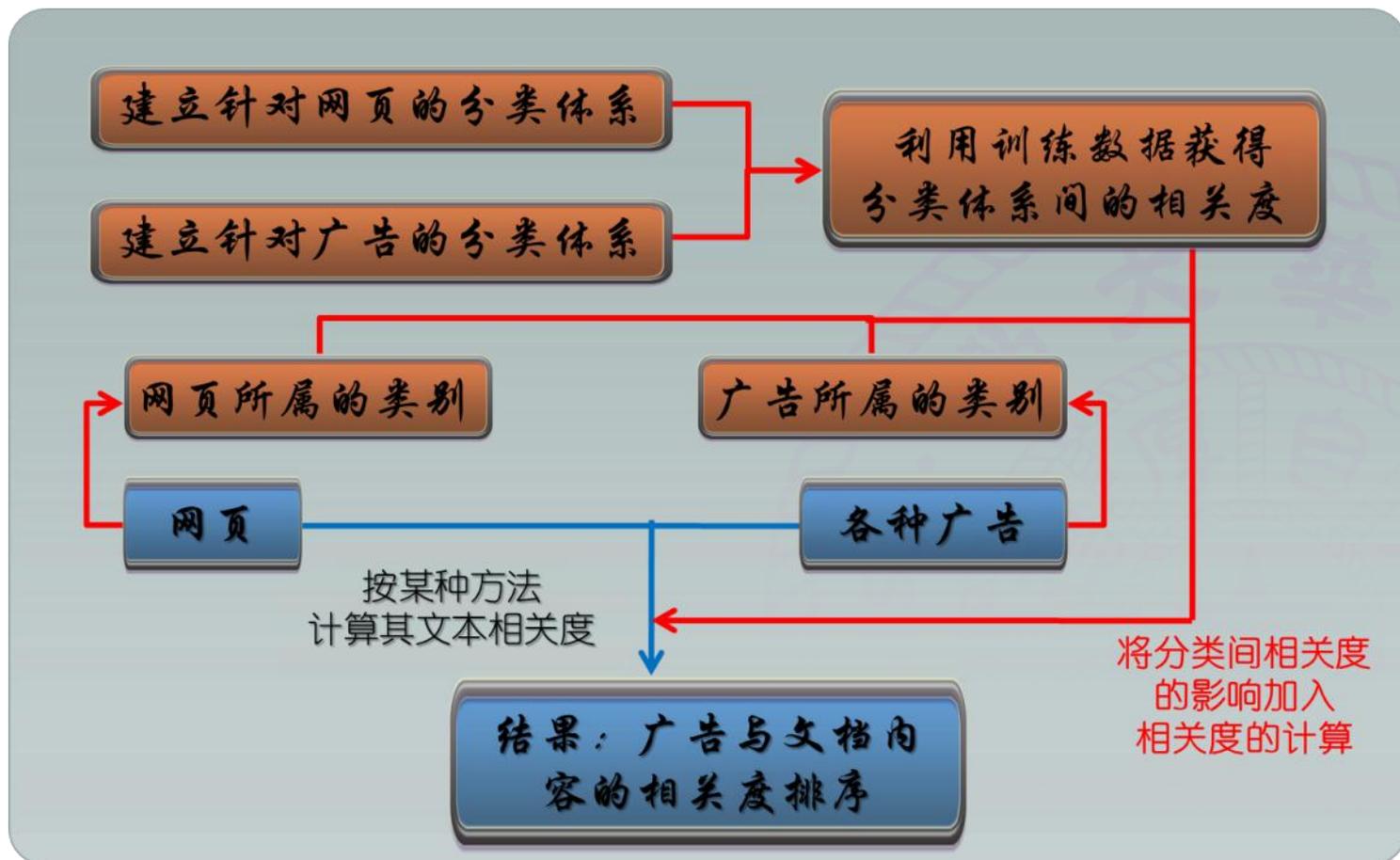
- * 基于用户行为的广告商分类
- * 采用C4.5决策树算法
- * 分类正确率达到92%，平均绝对误差仅为0.06





广告投放技术

* 在广告投放中引入分类信息





Information Retriever @ Tsinghua University

广告投放技术

* 在广告投放中引入分类信息

网页主题	序号	传统方法	加入“类间相似度”因素
安切洛蒂： 我只想执教AC米兰队 与切尔西无接触 (一则体育方面的新闻)	1	米兰医院	red
	2	上海到米兰机票预订	查北京电子票
	3	米兰机票	福彩新疆站
	4	四川丰胸医院	足球
	5	cavalli	cleveland
	6	red	cuba
	7	莱芜市zippos黑冰	2006高尔夫
	8	冠军腰线	上海到米兰机票预订
	9	四川除皱医院	米兰机票
	10	金昌zippos黑冰	悉尼特价机票



Information Retriever @ Tsinghua University

广告投放技术

* 构建用户兴趣与需求描述

- * 基于查询：提取用户提交的所有查询；
- * 基于点击：提取用户点击的查询结果的标题和摘要；
- * 基于浏览行为：提取用户浏览过的网页的标题。

[玩弄 药奴 别硬来](#)
[幕 女贪官 暗算书记](#)
[情 小娇妻 白领情事](#)
[杂议性虐 黑狐凶案](#)
[中医养生 经穴健康](#)
[是继位还是篡位?](#)



- [曝张柏芝座驾撞伤六旬老翁](#)
- [洗澡女干露露节目遭母亲打](#)
- [周慧敏竟然穿了一件透视裙](#)

站由屈朔广生标题以及百
词

```

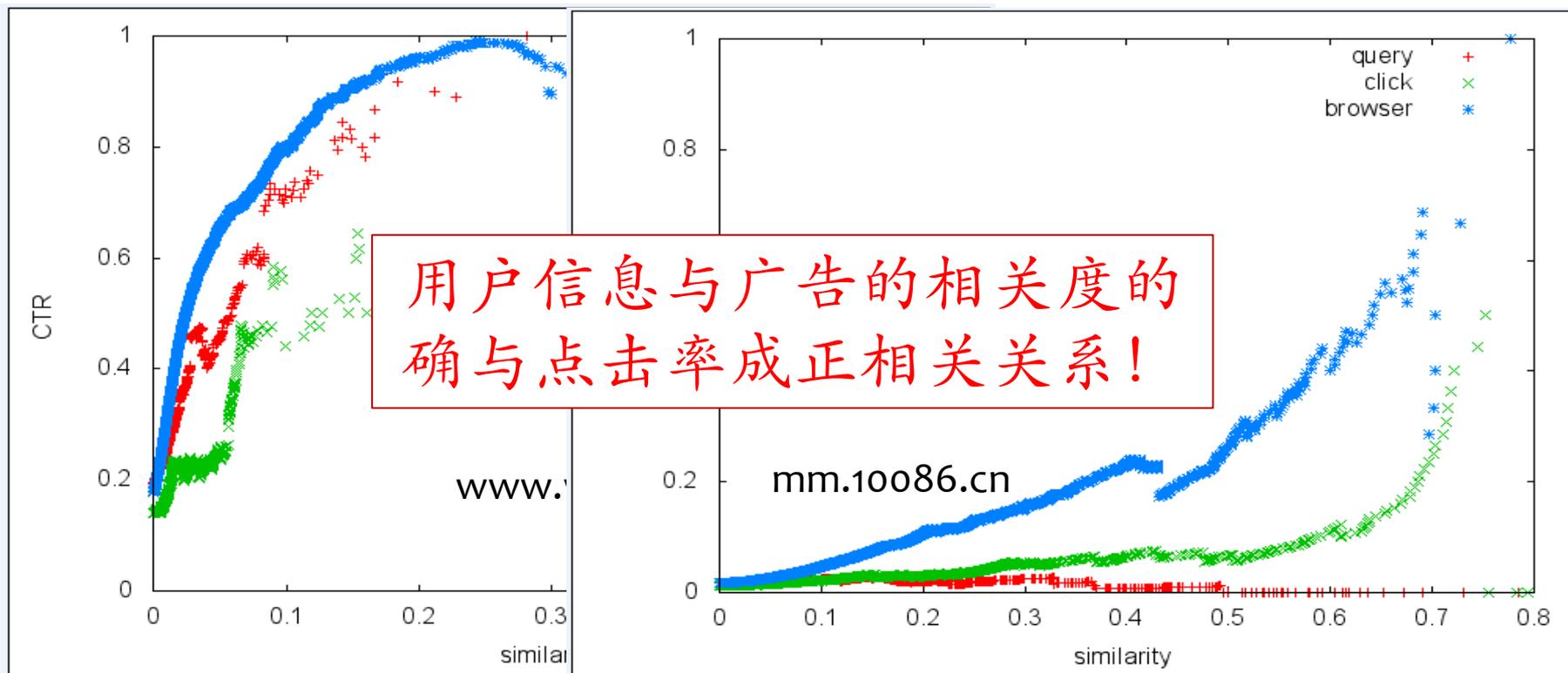
<title>VANCL 凡客诚品: 互联网快时尚品牌 </title>
<meta content="VANCL 凡客诚品, 快时尚, 男装, 女装, 童装, 鞋, 家居, 配饰, 衬衫, 牛津纺, 牛津纺衬衫, 衬衣, 长袖衬衫, 短袖衬衫, 全棉, 纯棉, 百分百棉, 100%棉, 全棉衬衫, 纯棉衬衫, 全棉衬衣, 纯棉衬衣, 免烫, 免烫, 免烫衬衫, 免烫衬衫, 免烫衬衣, 免烫衬衣, 牛津纺衬衣, 领尖扣, 直领, 小方领, POLO, 短袖POLO, 长袖POLO, 条纹POLO, 紫色 POLO, T恤, 圆领T恤, V领, 圆领T, 印花T, 文化衫, 卫衣, 打底衫, 高领衫, 低领, 鞋, 凉鞋, 皮鞋, 板鞋, 商务皮鞋, 正装皮鞋, 洞板鞋, 潮鞋, 休闲皮鞋, 帆布鞋, 运动鞋, 运动休闲鞋, 家居鞋, 雪地靴, 靴子, 平底鞋, 沙滩鞋, 夹脚鞋, 圆头, 尖头, 女鞋, 休闲鞋, 男鞋, 童装, 童鞋, 丝袜, 长筒袜, 连裤袜, 网袜, 天鹅绒, 瘦腿袜, 中筒袜, 筒袜, 棉袜, 靴袜, 打底袜, 羽绒服, 项链, 手镯, 围巾, 棉线衫, 开衫, 针织衫, 外套, 西服, 休闲西服, 夹克, 毛背心, 毛背心, 裤子, 长裤, 休闲裤, 牛仔裤, 牛仔, 卡其裤, 直筒休闲裤, 直筒卡其裤, 免烫休闲裤, 免烫卡其裤, 斜纹休闲裤, 斜纹卡其裤, 短裤, 沙滩裤, 内衣, 内裤, 秋衣, 秋裤, 三角裤, 平角裤, 领带, 袜子, 家居, 浴巾, 面巾, 毛巾, 收纳, 户外, 床品, 伞, 餐垫, 拖鞋, 盖毯, 断码, 打折" name="keywords"/>
<meta content="VANCL 凡客诚品, 互联网快时尚品牌. 根植互联网, 全球时尚潮流, 国际一线品质, 平民价位. 在线销售男装、女装、童装、鞋、家居、配饰等. 送货上门、货到付款, 无条件退换货." name="description"/>

```



广告投放技术

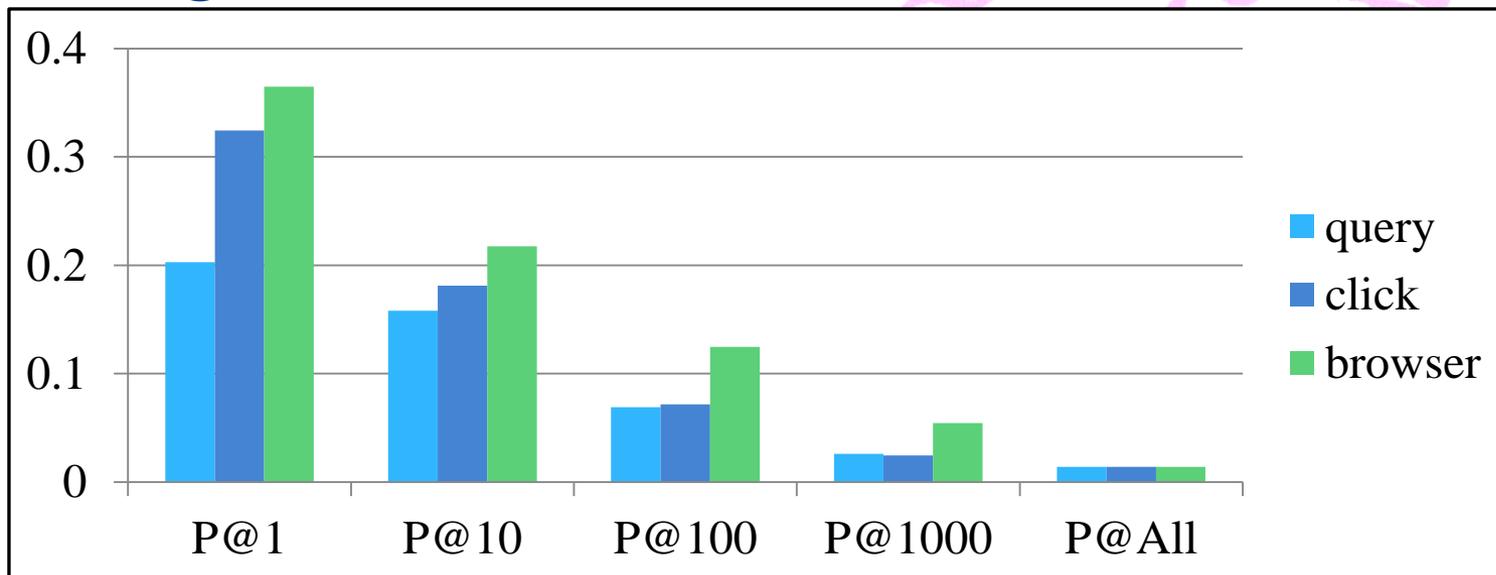
- * 随机选定一定规模用户，针对每个广告，计算用户与广告的相关度





广告投放技术

- * 比较三种用户兴趣与需求构建方法的优劣
 - * 针对每一广告，假设受投放成本限制，只能为N位用户投放广告，则按照用户与其相关度从高到低排序，根据用户实际点击情况观察投放效果
 - * 使用P@N评价指标进行评价





Information Retrieval @ Tsinghua University

面向搜索引擎的用户行为分析

- * 基于用户行为分析的页面质量评估
- * 基于用户行为分析的广告投放技术
- * 基于用户行为分析的计算社会学研究

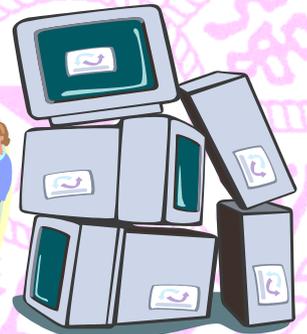




Information Retriever @ Tsinghua University

计算社会科学

- * 什么是计算社会科学(Computational social science)
 - * 将计算机的运算能力使用在仿真社会的各种现象上。可以认为是将自然科学的方法论，运用到社会科学上的尝试。
 - * 优势：将传统意义上不可重复的社会现象加以模拟和重复，拓展社会科学的研究面与广度
 - * 问题：社会现象真的可以被计算么？





Information Retriever @ Tsinghua University

计算社会科学

* 百度数据中心发布的“行业研究报告”

2010年十一境外旅游关注度排行

1	马尔代夫	■■■■
2	巴厘岛	■■■■
3	巴黎	■■■■
4	新加坡	■■■■
5	泰国	■■■■
6	首尔	■■■■
7	济州岛	■■■■
8	普吉岛	■■■■
9	香港	■■■■
10	塞班岛	■■■■

2010年Q2汽车品牌关注度排行

1	大众	■■■■
2	丰田	■■■■
3	雪佛兰	■■■■
4	本田	■■■■
5	奥迪	■■■■
6	日产	■■■■
7	宝马	■■■■
8	别克	■■■■
9	现代	■■■■
10	比亚迪	■■■■

2010年Q2笔记本品牌关注度排行

1	联想 (ThinkPad)	■■■■
2	惠普 (Compaq)	■■■■
3	华硕	■■■■
4	戴尔	■■■■
5	索尼	■■■■
6	宏碁	■■■■
7	东芝	■■■■
8	苹果	■■■■
9	三星	■■■■
10	神舟	■■■■

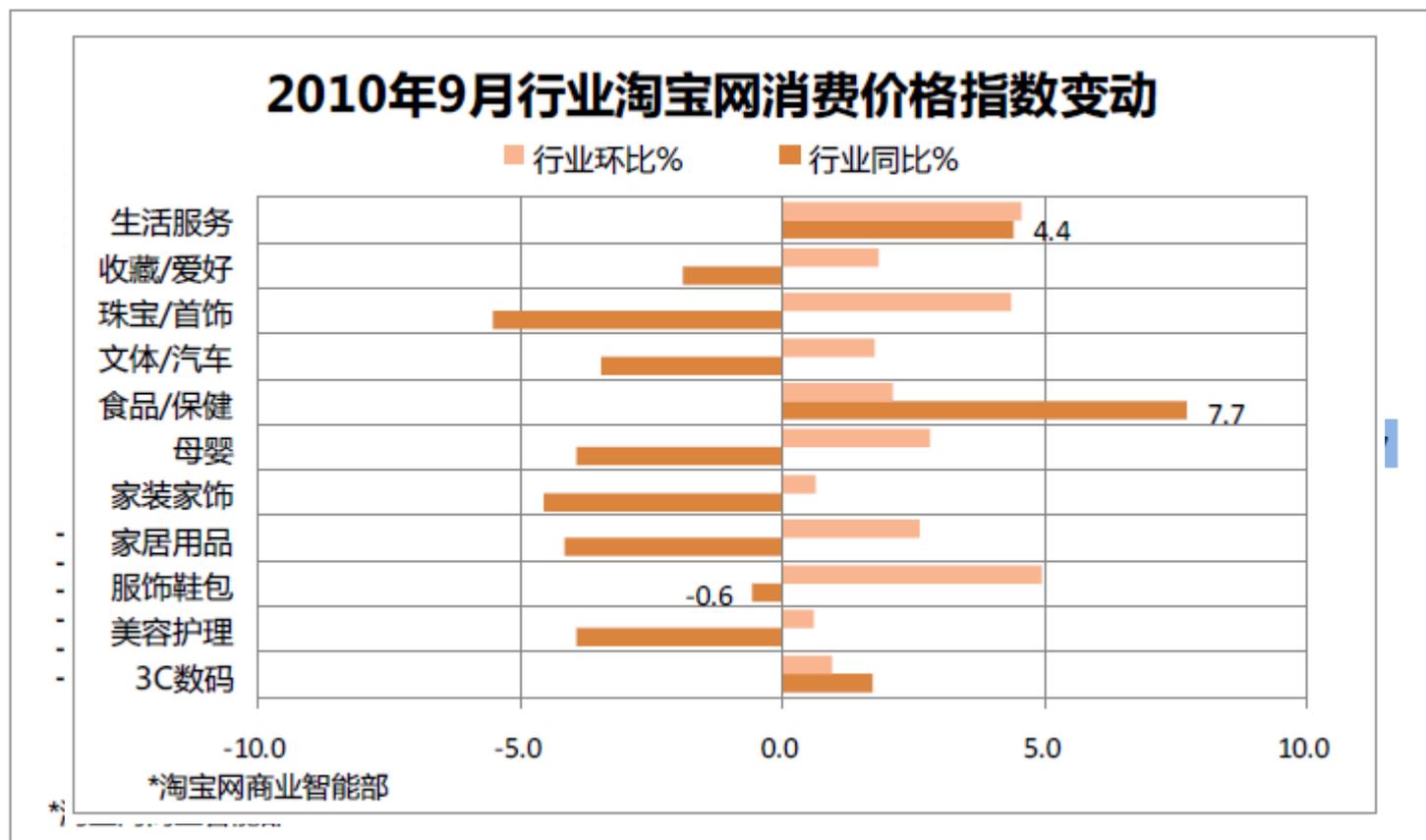




Information Retriever @ Tsinghua University

计算社会科学

* 淘宝发布的“TCPI指数”





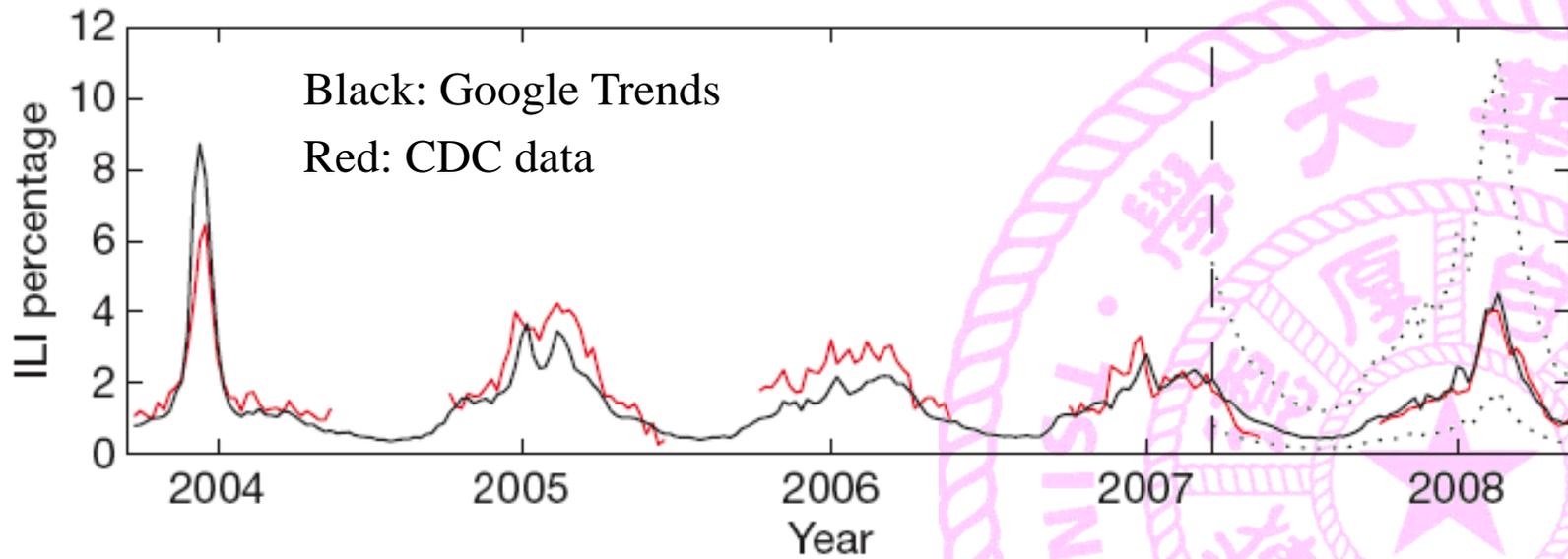
Information Retrieval @ Tsinghua University

计算社会科学

* 谷歌发布的流感发病趋势预测

* <http://www.google.com/trends/flu>

* 当地查询日志可以用于预测此地流感发病趋势



Jeremy Ginsberg et. al. Detecting influenza epidemics using search engine query data, Nature, Vol 457, 19 February 2009

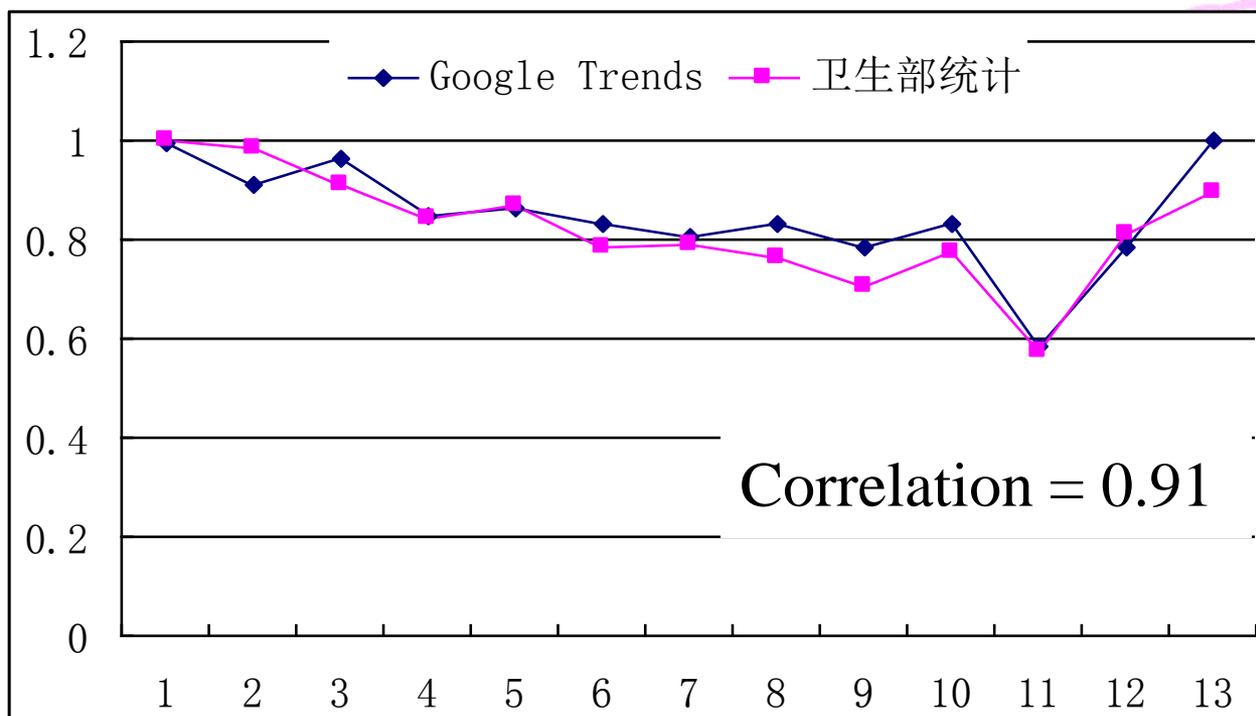


Information Retriever @ Tsinghua University

计算社会科学

* 我们的工作:

* 卫生部公布的肺结核发病数据 V.S.
Google trends 中国范围内的“肺结核”查询趋势





Information Retriever @ Tsinghua University

季节性传染病流行趋势预测

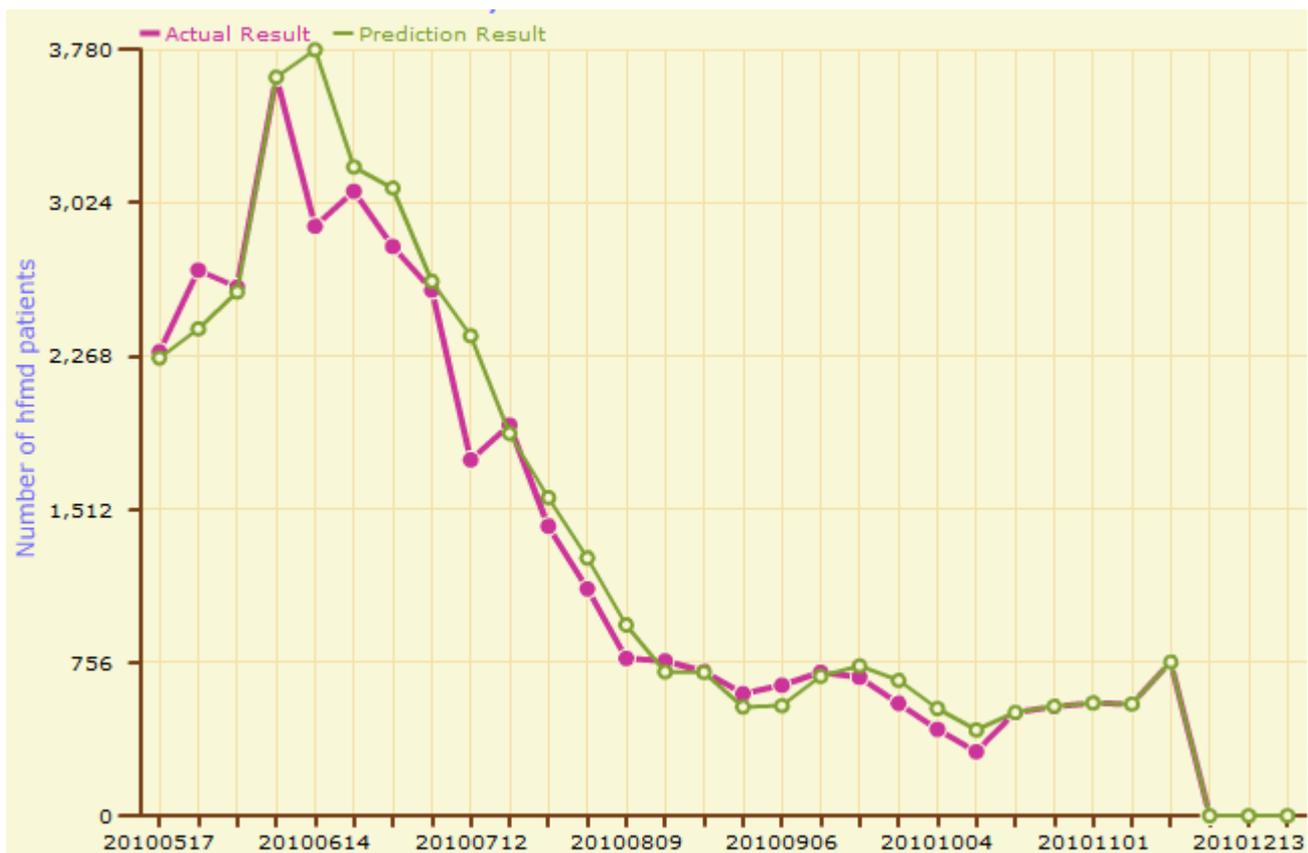
- * 传染病与人类健康息息相关
 - * 2003年的SARS && 2009年的H1N1
 - * 近年来发病人群日益增加的艾滋病&&乙型肝炎等慢性传染病
 - * 2010年春季爆发的手足口病
 - * 卫生部门对法定传染病的官方统计延迟较大
- * 用户查询点击日志可以反映很多社会事件
 - * 网络日志实时可以获取并更新，延迟很小
 - * 通过对潜在患者的用户行为进行分析，实现在潜伏期对未来发病趋势的预测



Information Retriever @ Tsinghua University

计算社会科学

* 对北京地区手足口病发病趋势的预测





Information Retriever @ Tsinghua University

总结

- * 互联网发展现状
- * 搜索引擎技术挑战
- * 面向搜索引擎的用户行为分析
 - * 基于用户行为分析的数据质量评估研究
 - * 基于用户行为分析的广告投放技术
 - * 基于用户行为分析的计算社会科学研究





Information Retriever @ Tsinghua University

Thank you



Welcome to visit our homepage

<http://www.thuir.cn/>

On-line Demos

[Search Engine Evaluation](#)

[Seasonal Epidemic Prediction](#)

[Web Spam Page Identification](#)

[Web News Event Clustering](#)