

万维网数据质量评估:挑战与进展

清华大学智能技术与系统国家重点实验室

刘奕群 张敏

2010年9月18日





Information Retriever @ Tsinghua University

万维网上有多少页面？

- * Surface Web contains 167 Terabytes (**10 billion pages est.**); Deep Web contains 91,850 Terabytes data. (*How Much Info project, UCLA, 2003*)
- * The Indexable Web is more than **11.5 billion pages.** (*WWW poster, Antonio Gulli et al., 2005*)
- * 截止2009年底，中国网页数目达到**336亿个**，年增长率超过100%. (*CNNIC, 2010*)
- * Google: Now We Know About **1 Trillion Pages.** (*Google official blog, 2008*)



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

* 2010年上半年，中国注册网站数减少了13.7%；

全球注册网站数减少了11.5%

The screenshot displays two Internet Explorer browser windows. The left window is titled '恶作剧~的收藏 - QQ书签' and shows a list of bookmarks. The right window is titled '天涯问答' and shows a forum thread titled '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 been browsing on Taobao for a long time and doesn't know what to buy. I'm begging for a good weight loss medicine.).

Left Window: QQ Bookshelf

- Address: <http://shuqian.soso.com/296321560/?p=2>
- Navigation: 首页 | 热门收藏 | 最新收藏 | 我的收藏 | 收藏新网址
- Collection: 恶作剧~的收藏 (208)
- Item 1: 就去色色—97sese, 11sss, 11mmm, 22ccc, 自拍偷拍, 美腿玉足, 人...
分类: 就去色色—97sese 11sss 11mmm 22ccc 自拍偷拍 美腿玉足 人..
描述:
- Item 2: 97sese.97ai.77bbb.97mo.11sss.xiao77激情美女
分类: 97sese.97ai.77bbb.97mo.11sss.xiao77激情美女
描述:
- Item 3: 97sese.97ai.77bbb.97mo.11sss.xiao77激情美女
分类: 97sese.97ai.77bbb.97mo.11sss.xiao77激情美女
描述:
- Item 4: 电影 黄网 97爱 97ai 趁关系秘照

Right Window: Tianya Forum

- Address: <http://wenda.tianya.cn/wenda/thread?tid=15c9n3k8jp0ard3e9jgphpe6jvk8a1e6siss>
- Search: 输入要查询的文字
- Navigation: 首页 > 提问列表 > 提问
- Thread Title: 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. [健康]
- Post by: 蹠起脚尖说爱你 (2010-09-15 11:30:47) 121.33.96.*
- Answers:
 - Post by: xw384uy29 (2010-09-15 11:31:15) 119.109.45.*
现在很流行网购，网上的东西真的便宜很多，选购也很方便，我收集了很多热卖好评的商品和店铺，并做了统计排行，很多都是专家通过比较店铺信誉和销售记录以及网友的评价，做出的排行榜，当然也有很多是我通过购买和网友的交流统计出来的，都是热卖好评的，网购这么多年了，现在才知道，原来这样统计下，真的方便很多，现在分享给大家，^_^，地址: www.bang123.net，这些都是偶平时在淘宝网经验总结出来的好产品了，淘宝上的东西也是很便宜很方便的咯，偶还专门收集了很多热卖好评的产品和高信誉的店铺，还专门做了统计做出了排行榜，用起来真的方便很多了，现在分享给大家收藏起来挺实用的!
- More: 更多→【淘宝网热销产品排行榜】请访问淘宝精华榜
 - 淘宝减肥 排行榜 www.bang123.net/jianfei.html
 - 左旋 肉碱排行榜 www.bang123.net/zuoxuanroujian.html
 - 丰胸排行榜 www.bang123.net/fengxiong.html
 - 瘦腿排行榜 www.bang123.net/shoutui.html



万维网数据质量评估

- * 对于搜索引擎:

- * Two of the top 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.**

- * One of the top two challenges proposed by Amit Singhal (in SIGIR 2005, ECIR 2008)

- * **Search Engine Spam**, Evaluation

- * 对于其他研究领域:

- * 数据挖掘、情感分析、自然语言处理、...



Information Retriever @ Tsinghua University

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

量排名

WW

WWW

W

lat

WW

WW

WWW

bl

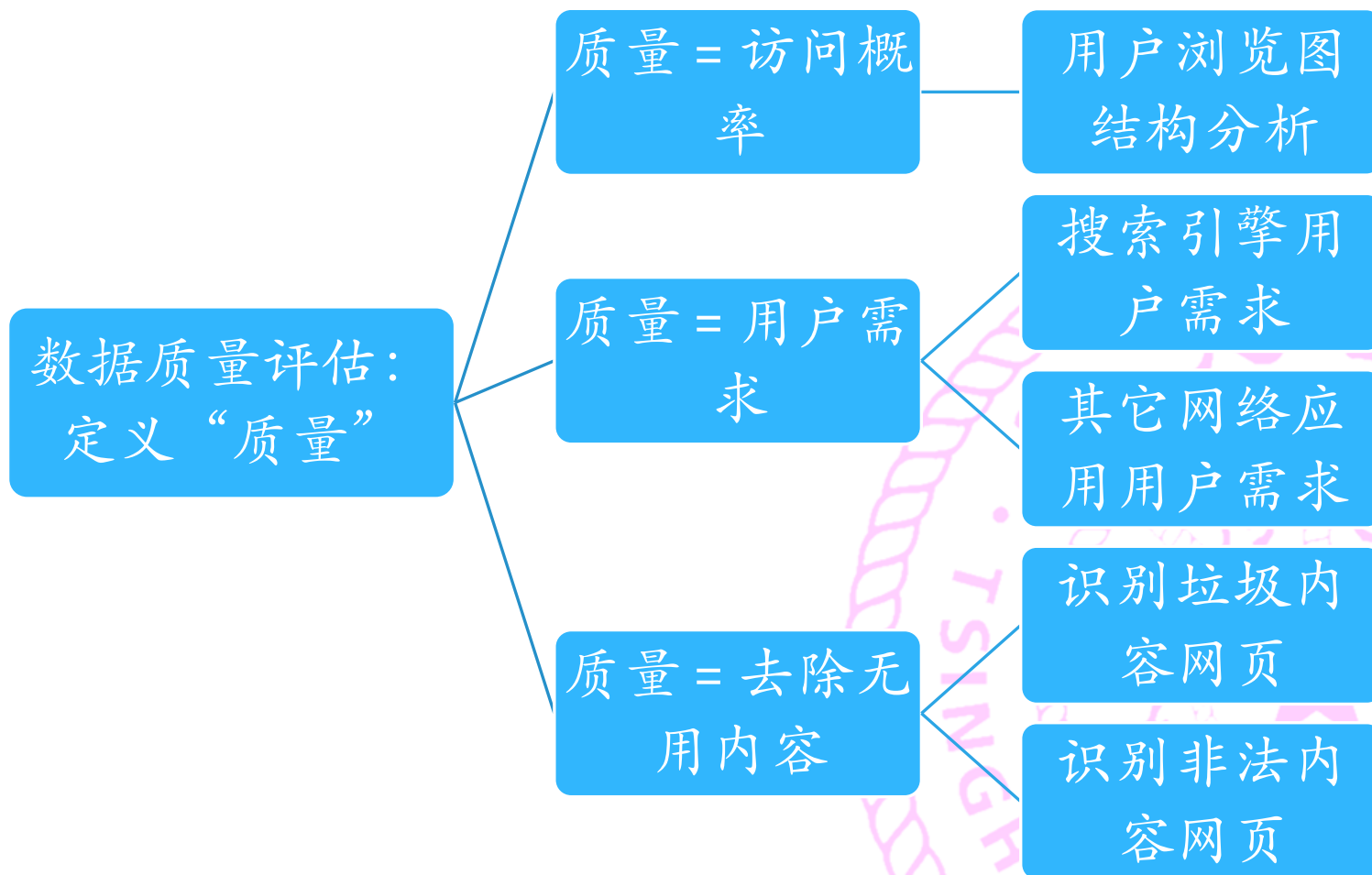
WW

WWW



Information Retriever @ Tsinghua University

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





Information Retrieval @ Tsinghua University

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

* 例如

* 3

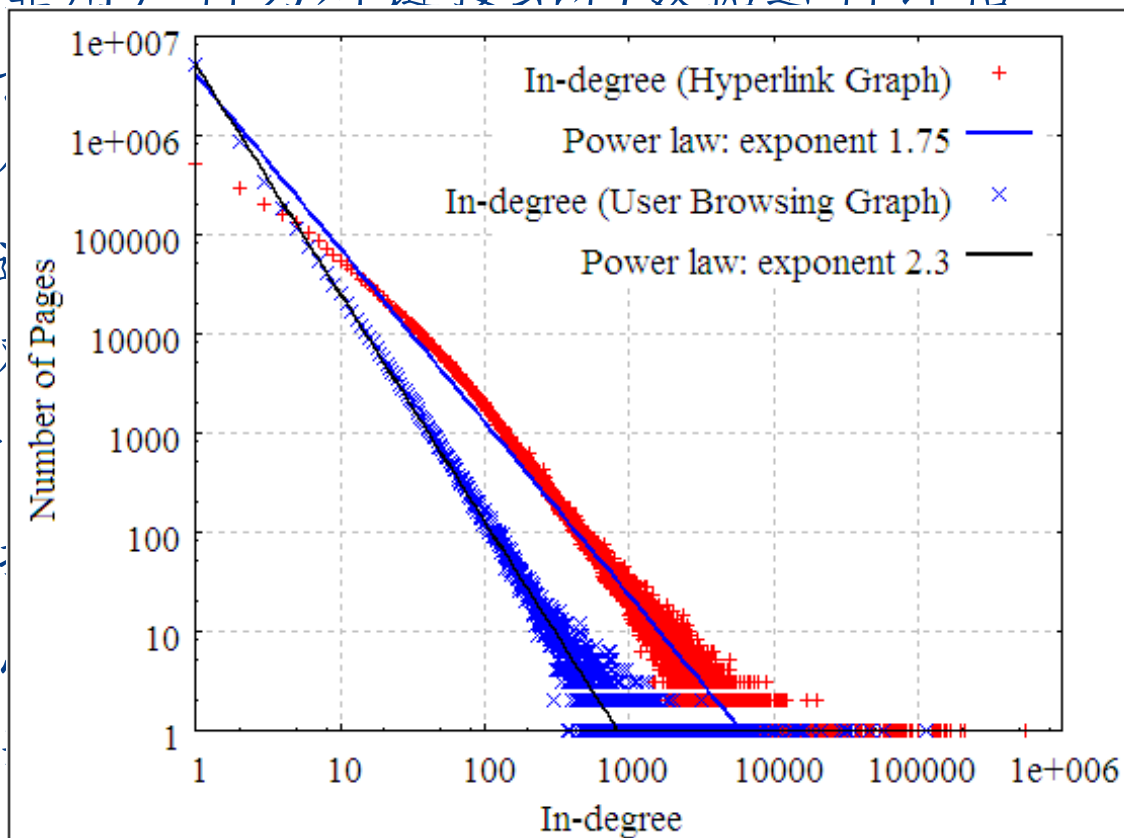
* 客户

* 用户

* 数据

* 顶点

* 边缘

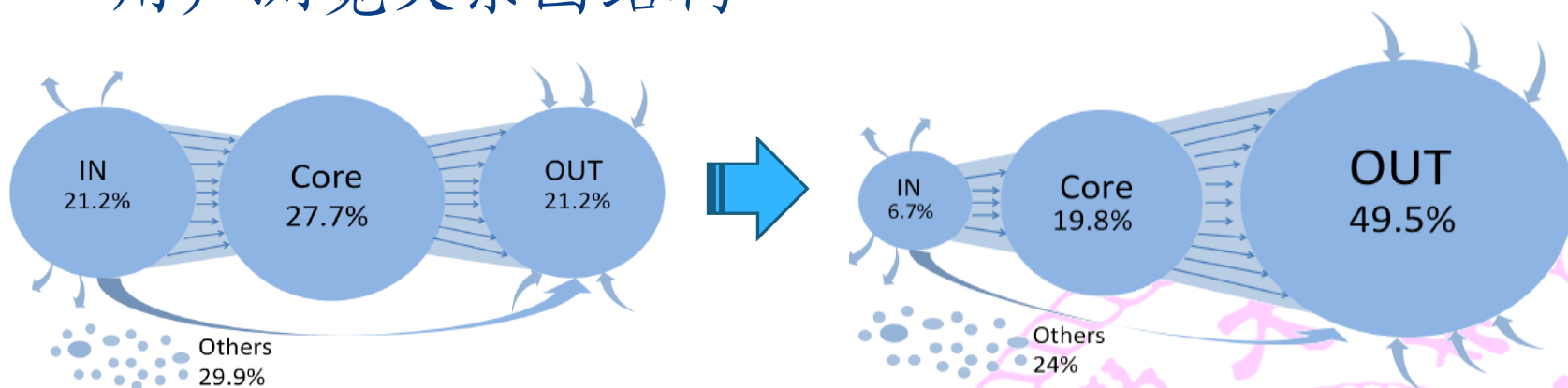




Information Retriever @ Tsinghua University

质量：网页访问概率

* 用户浏览关系图结构



Alexa.com访问 质量排名	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	9	14
5	www.hd315.gov.cn	3	16



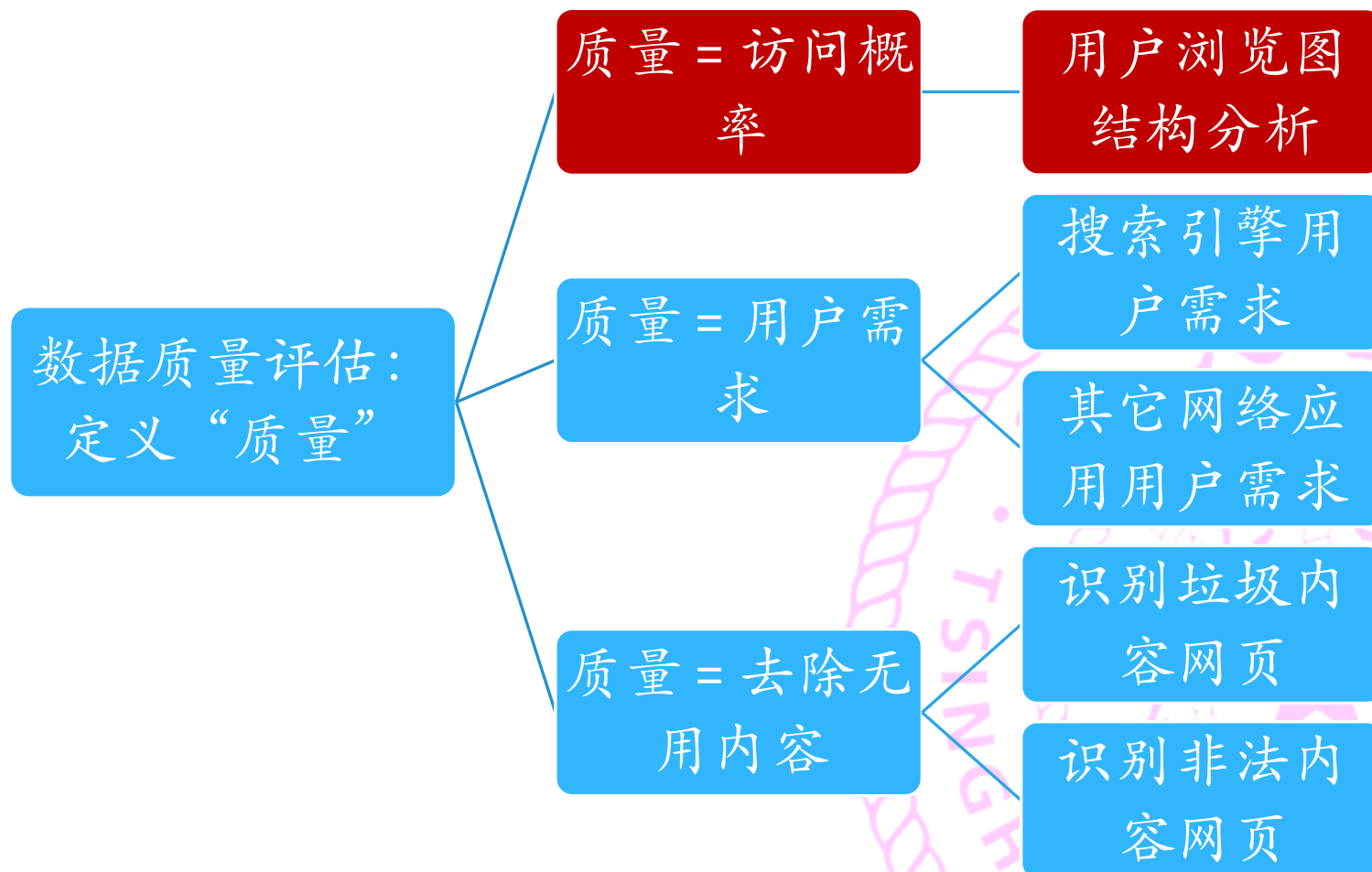
Information Retriever @ Tsinghua University

参考文献

- * Yiqun Liu, Yijiang Jin, Min Zhang, Shaoping Ma, Liyun Ru, User Browsing Graph: Structure, Evolution and Application. Late breaking result session in Second ACM International Conference on Web Search and Data Mining (WSDM 2009).
- * Yiqun Liu, Yufei Xue, Ronwei Cen, Min Zhang, Shaoping Ma, Liyun Ru, Constructing a Reliable Web Graph with Browsing Behavior Information. Conditional accepted by Decision Support Systems (DSS). 2010.
- * 薛宇飞, 刘奕群, 张敏, 马少平, 茹立云. 基于用户浏览图的网页质量评估方法的比较分析. 全国第十届计算语言学学术会议(CNCCL-2009), 482-488.
- * 王晓光, 刘奕群, 岑荣伟, 马少平, 茹立云, 基于用户行为的页面质量评估. 第十二届中国机器学习会议(CCML 2010), 计算机科学与探索, 2010.



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

查询信息需求类型	检索任务	用户需求比例 (Broder 2002)	用户需求比例 (Rose 2004)
导航类	命名网页定位	20%	15%
	主页定位		
信息类	主题过滤	50%	61%
事务类	垂直搜索	30%	24%

- * 检索目标页面是否存在与查询无关的特征？

- * PageRank!



Information Retrieval @ 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.”
Udi Manber, Google's head of search

- * 用户访问量大小，PageRank，入链接个数，入链接文本长度等

- * 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

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

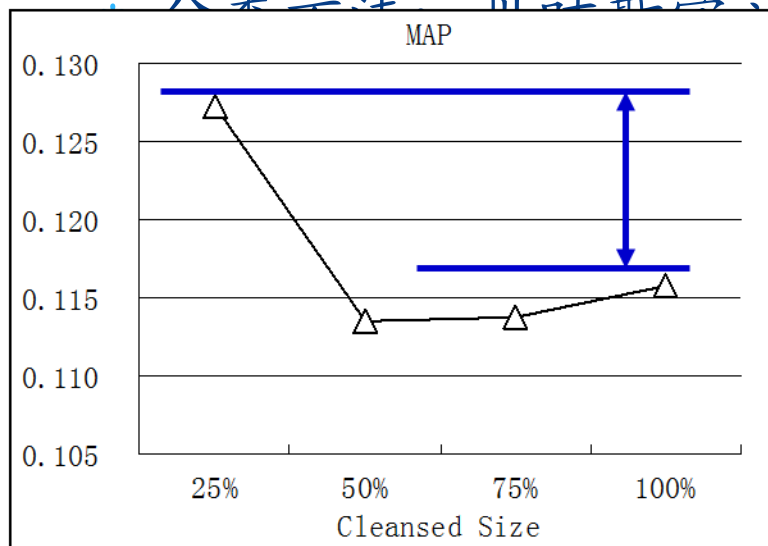


Information Retriever @ Tsinghua University

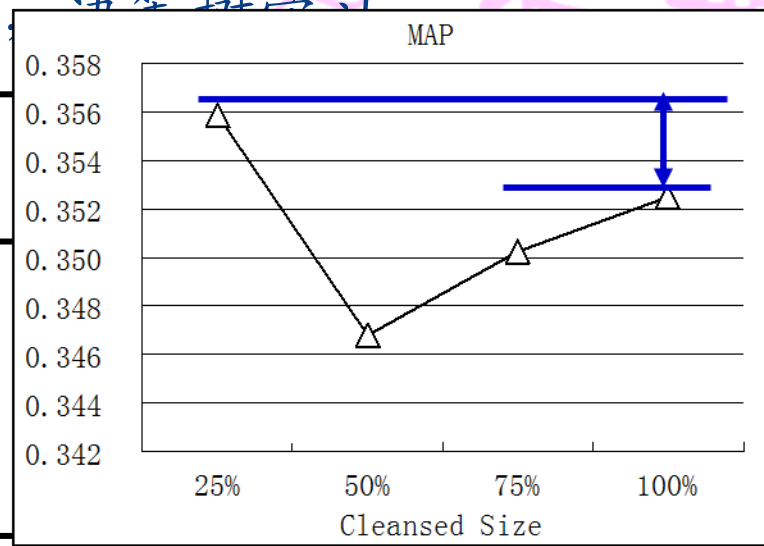
质量：用户需求

* 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



Information Retriever @ Tsinghua University

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.

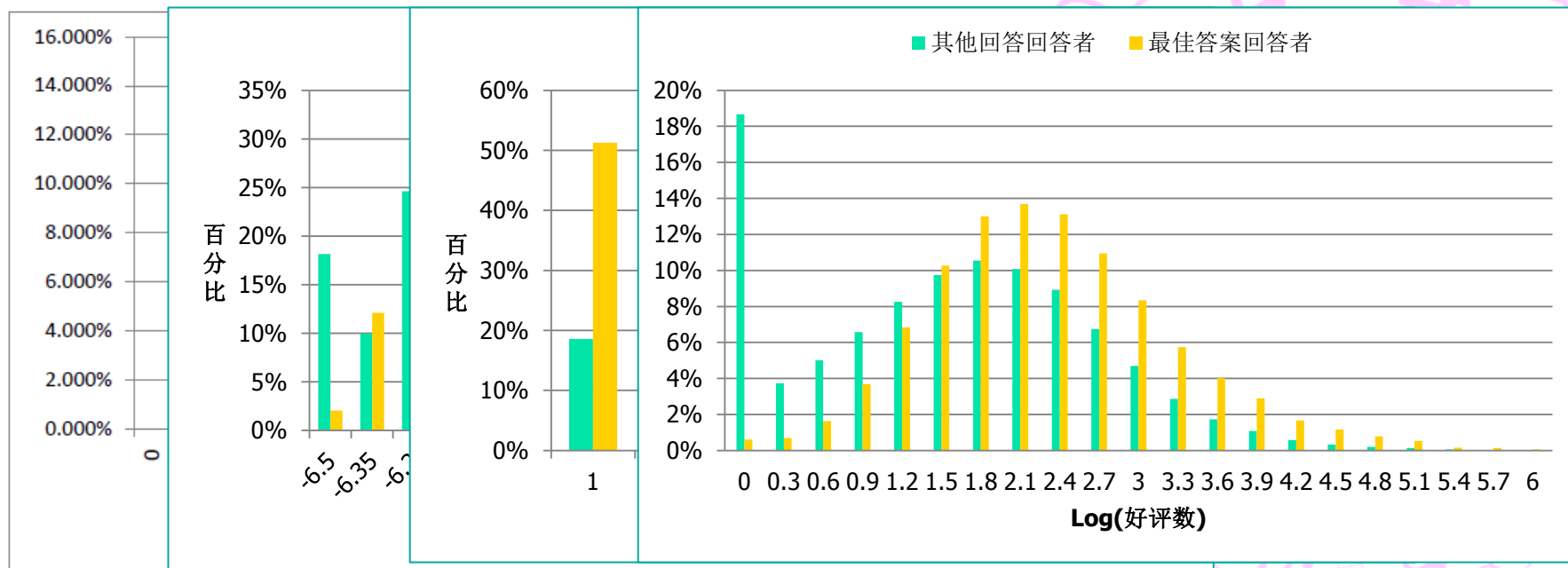


质量：用户需求

* 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。





Information Retriever @ Tsinghua University

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%



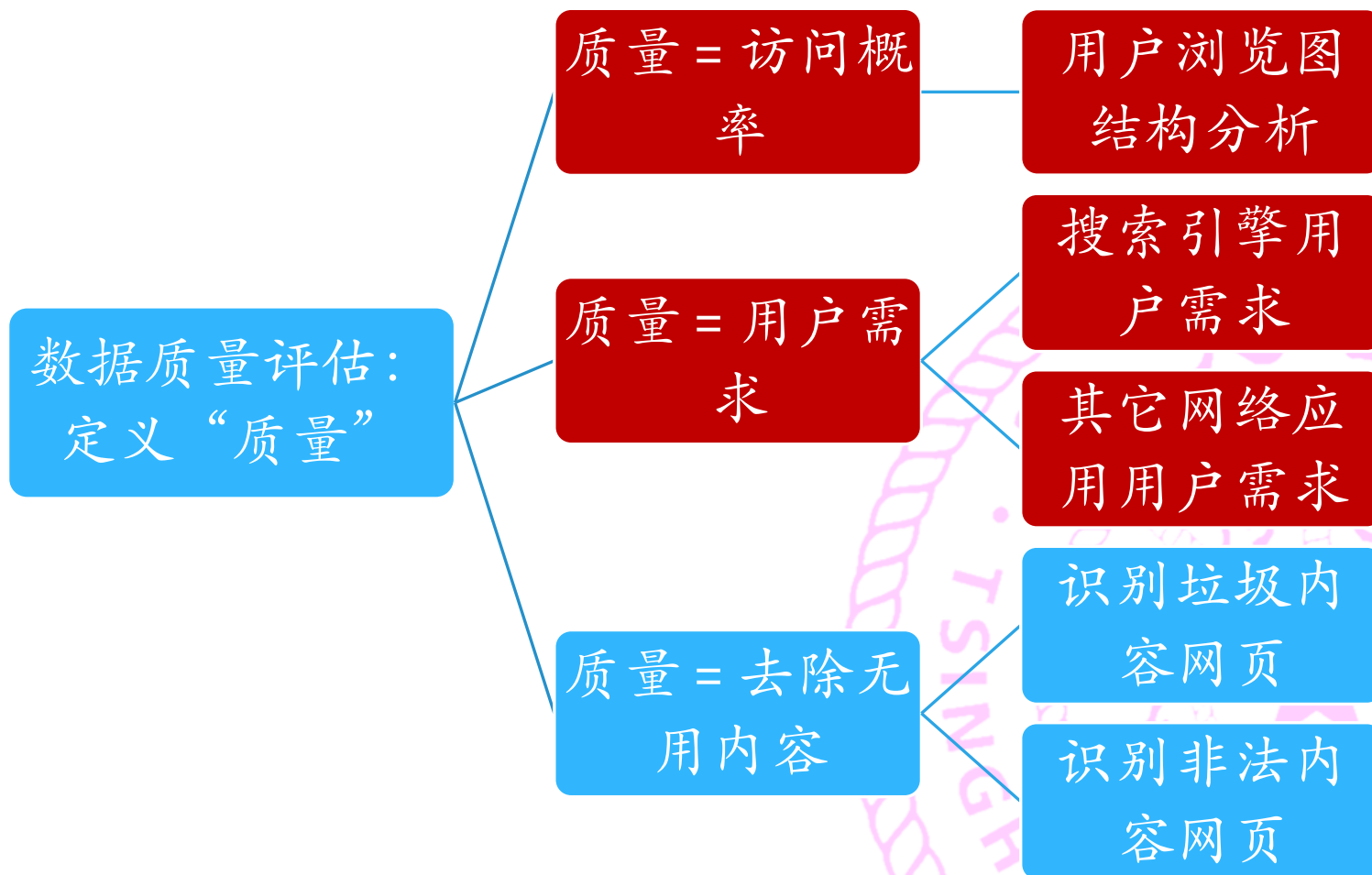
Information Retriever @ Tsinghua University

参考文献

- * Data Cleansing for Web Information Retrieval using Query Independent Features. Yiqun Liu, Min Zhang, Liyun Ru, Shaoping Ma. Journal of the American Society for Information Science and Technology (JASIST), Volume 58, Issue 12, Pages 1884-1898, 2007.
- * Yiqun Liu, Canhui Wang, Min Zhang, Shaoping Ma, Web Data Cleansing for Information Retrieval using Key Resource Page Selection, Proceedings of the 14th International World Wide Web conference, Pages 1136-1137, 2005.
- * 刘奕群,张敏,马少平.基于改进决策树算法的网络关键资源有效定位. 软件学报, Vol. 16(11): pp. 1958-1966, 2005.
- * 刘奕群,张敏,马少平.面向信息检索需要的网络数据清理研究. 中文信息学报 Vol. 20(3): pp. 70-77, 2006.
- * 孔维泽, 刘奕群, 张敏, 马少平, 问答社区中回答质量的评价方法研究. 第六届全国信息检索学术会议, 2010.



数据质量评估的研究思路





Information Retriever @ Tsinghua University

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 Retrieval @ Tsinghua University

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。



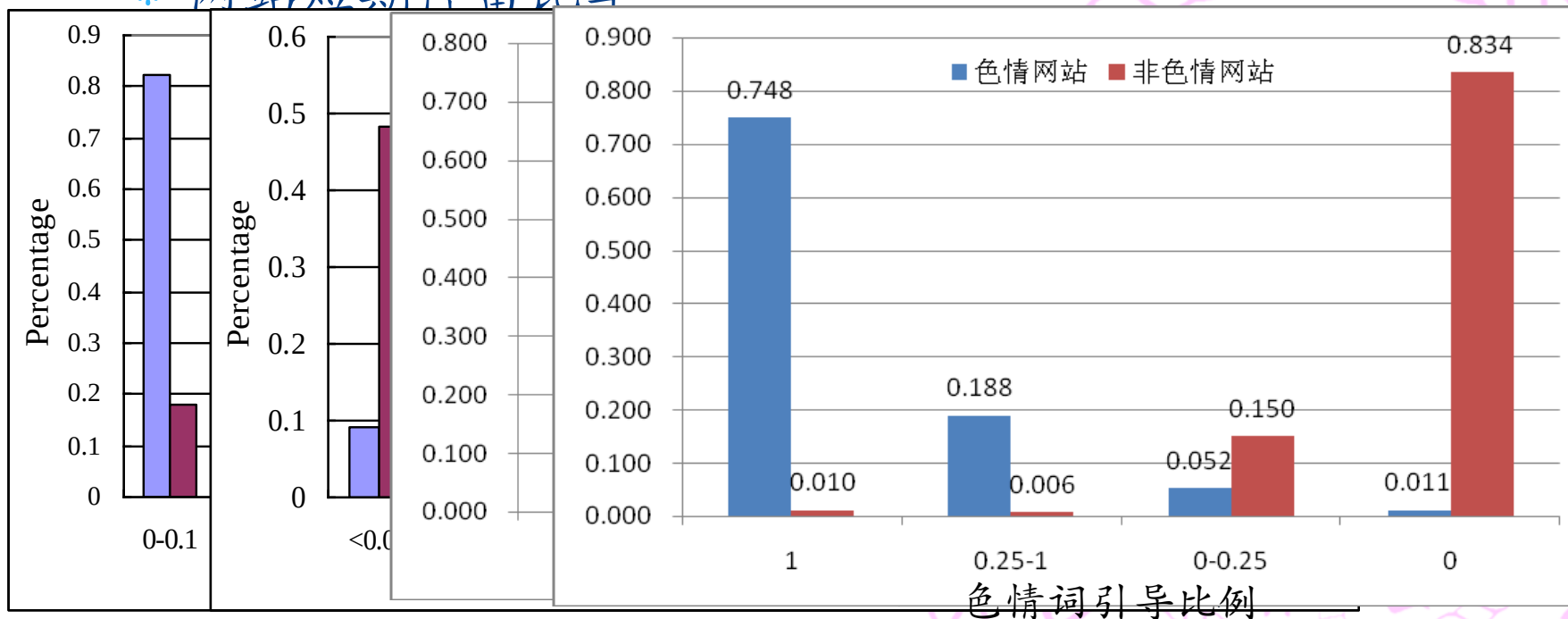
质量：去除无用内容

* 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





Information Retrieval @ Tsinghua University

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 Retriever @ Tsinghua University

参考文献

- * Yiqun Liu, Rongwei Cen, Min Zhang, Shaoping Ma, Liyun Ru. Identifying Web Spam with User Behavior Analysis. The Fourth International Workshop on Adversarial Information Retrieval on the Web. 2008.4.
- * Liu, Y., Zhang, M., Ma, S., and Ru, L. 2008. User behavior oriented web spam detection. In Proceeding of the 17th international Conference on World Wide Web (Beijing, China, April 21 - 25, 2008). WWW '08. ACM, New York, NY, 1039-1040.
- * 岑荣伟, 刘奕群, 张敏, 茹立云, 马少平, 基于日志挖掘的搜索引擎用户行为分析, 中文信息学报Vol.24(3): pp49-54. 2010。
- * Huijia Yu, Yiqun Liu, Min Zhang, Shaoping Ma and LiYun Ru. Web Spam Identification with User Browsing Graph. Asia Information Retrieval Symposium 2009 (AIRS 2009), also in Springer LNCS 5839, pp. 38-49, 2009.11.
- * 曹建勋, 刘奕群, 岑荣伟, 马少平, 茹立云, 基于用户行为的色情网站识别。第六届全国信息检索学术会议, 2010。(优秀论文)
- * Yiqun Liu, Weize Kong, Huijia Yu, Min Zhang, Shaoping Ma, Liyun Ru. Submitted to ACM Transaction on the Web (TWeb), 2010.



数据质量评估研究展望

- * 数据质量评估的粒度问题
 - * 网站：粒度过粗（作者不统一、代管问题）
 - * 网页：粒度过细（变化快，新出现网页的评估难题）
- * 多层次质量评估体系的制定
 - * “高质量”与“低质量”容易定义
 - * “中等质量”网页的评估标准难以制定
- * 如何利用自然语言处理技术改进质量评估
 - * 质量评估大都利用非内容特征，缺乏对内容的理解
 - * 如何引入适用于大规模网络数据处理的NLP技术？



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](#)