面向搜索引擎的互联网用户行为分析

智能技术与系统国家重点实验室

信息检索课题组

2009年11月15日

From Alexa.com

Global Top Sites

The top 500 sites on the web.

1. 
Yahoo!
Personalized content and search options. Chatrooms, free e-mail, clubs, and pager.
www.yahoo.com
Site info for yahoo.com

2. 
Google
Enables users to search the Web, Usenet, and images. Features include PageRank, caching and translation of results, and an option to find similar pages. The company's focus is developing search technology.
www.google.com
Site info for google.com

3. 
YouTube
YouTube is a way to get your videos to the people who matter to you. Upload, tag and share your videos worldwide!
www.youtube.com
Site info for youtube.com

4. **Windows Live**
Search engine from Microsoft.
www.live.com
Site info for live.com

5. **Facebook**
A social utility that connects people, to keep up with friends, upload photos, share links and videos.
www.facebook.com
Site info for facebook.com

Top Sites in China

The top 100 sites in China.

1. 
Baidu.com
The leading Chinese language search engine, provides â€œsimple and reliableâ€? search experience, strong in Chinese language and multi-media content including MP3 music and movies, the first to offer WAP and PDA-based mobile search in China.
baidu.com
Site info for baidu.com

2. 
QQ.COM
中国最大的门户网站，提供即时通讯、新闻资讯、网络游戏以及在线拍卖业务，
qq.com
Site info for qq.com

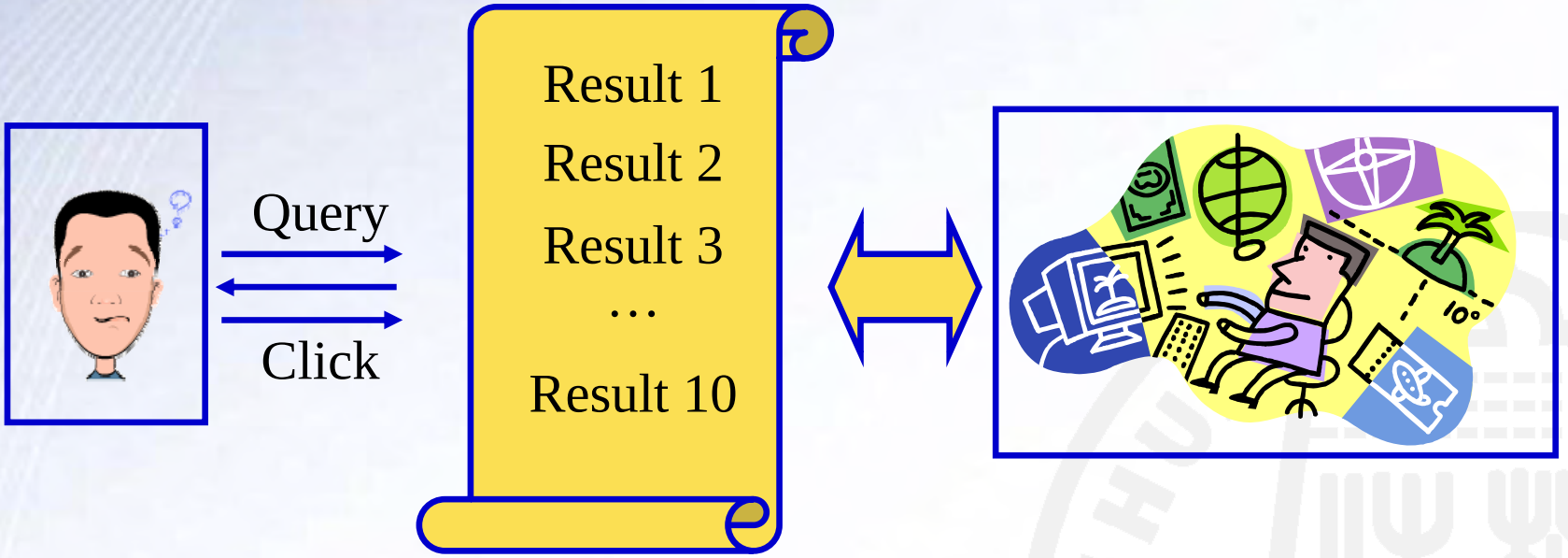
3. 
新浪新闻中心
包括即日的国内外不同类型的新闻与评论，人物专题，图库。
sina.com.cn
Site info for sina.com.cn

4. **Google**
网页、图片、新闻搜索，支持个性化搜索及本地搜索，提供论坛、邮箱、日历服务和桌面搜索工具。
google.cn
Site info for google.cn

5. **淘宝网**
包括电脑通讯、数码、男装、女装、童装、化妆品、书籍音像、运动用品、游戏装备等各种商品的买卖，还有相关的社区交流，同时提供支付宝网上交易安全保证系统。
taobao.com

搜索引擎面临的技术挑战

- 用户＆搜索引擎＆万维网



Google: I'm feeling lucky

搜索引擎面临的技术挑战

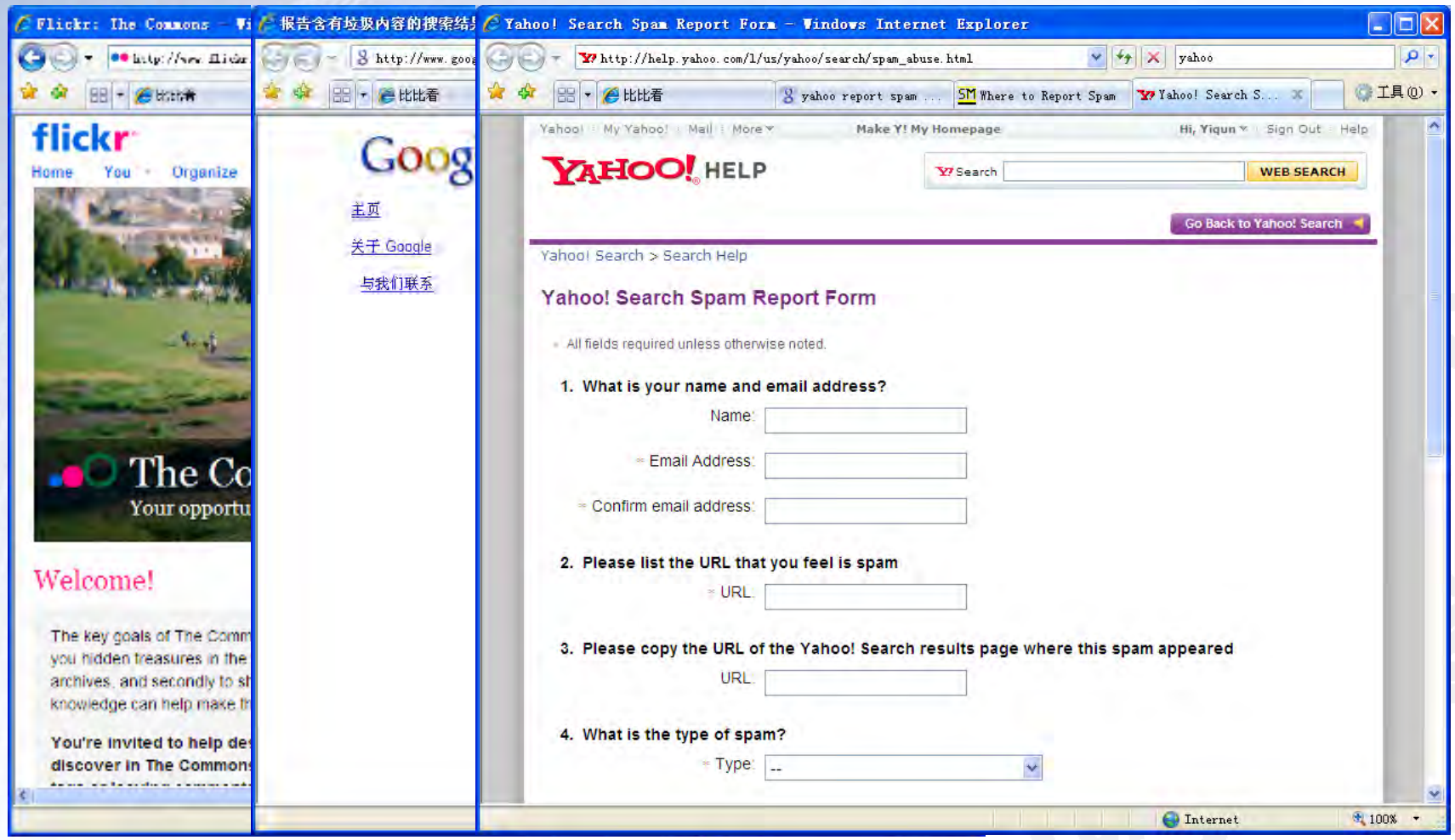
- 用户层面
  - 丰富的信息需求只能通过简短的查询来表示
    - 查询的平均长度为2-3个词
    - 构建复杂查询的尝试(W3QL, WebSQL等)以失败告终
- 万维网层面
  - 数据繁杂，质量参差不齐
    - 2002年，Web上所存储的数据超过500,000 TB
    - 2008年，Google索引量声称超过1 trillion 网页
    - 冗余、过期、低质量乃至垃圾数据层出不穷

如何解决？

- 借助用户的力量
  - 用户查询：如何查询高考分？
  - 传统思路：查询分析与分类，关键词提取...
  - 依靠用户的思路：百度知道

如何解决？

- 借助用户的力量



如何解决？

- 借助用户的力量
  - 搜索质量与经济利益密切相关
  - 群众的话不能不信，也不能全信
  - 需要借助标注人员的过滤，反馈速度慢

举报严重作弊网站 百度数周未见处理

2008-08-04 12:32:54 来源:CHINAZ月用户投稿 作者:小张 【大 中 小】 评论：2 条

收藏到：

近日，搜索众多自己网站关键词发现，域名为idc38.cn的网站出现结果中，而自己的网站被过滤掉，严重影响了搜索引擎的结果。

idc38.cn网站就是利用了类似百度快照的形式，自己替换别人网站的链接为自己网站，引导搜索引擎抓取，让搜索引擎认为其它网站的内容是自己网站的内容，显然在做一个搜索引擎库。

如何更好的借助用户的力量？

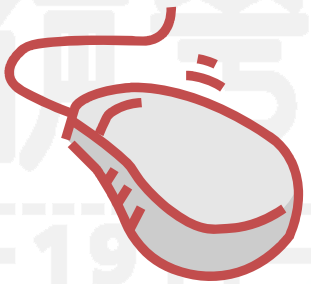
- 解决思路：用户群体的行为分析
  - 隐式反馈与显式反馈 (implicit / explicit feedback)
    - 显式反馈
      - 用户主动反馈
      - 直接，对用户行为产生影响，少量
    - 隐式反馈
      - 用户被动反馈
      - 间接，不对用户行为产生影响，大量

如何更好的借助用户的力量？

- 用户的点击都是有目的的
- 从统计角度分析，用户点击背后所隐藏的是用户的语义信息

如何更好的借助用户的力量？

- 用户行为的载体：日志数据
  - 查询与点击日志
    - 用户提交的查询
    - 用户点击了哪些结果
    - 其他辅助信息
      - 结果对应的排序
      - 时间戳
      - 用户点击的序列关系
      - 用户Session ID（记录在Cookie里）



如何更好的借助用户的力量？

- 用户行为的载体：日志数据
  - 互联网访问日志
    - 用户当前正在访问的网页
    - 用户从此网页出发下一步访问的网页
    - 辅助信息
      - 时间戳
      - 用户Session ID（记录在Cookie里）
      - 用户停留时间

面向搜索引擎的用户行为分析方法

- 用户层面
  - 利用用户的查询行为信息识别信息需求类别
  - 利用用户的查询行为信息进行查询推荐
- 万维网层面
  - 利用用户访问信息评估数据质量
  - 利用用户行为模式识别垃圾网页
- 搜索引擎层面
  - 利用用户查询行为进行搜索引擎查询性能评估
  - 利用用户访问信息构建网络信息检索语料

面向搜索引擎的用户行为分析方法

- 用户层面
  - 利用用户的查询行为信息识别信息需求类别
  - 利用用户的查询行为信息进行查询推荐
- 万维网层面
  - 利用用户访问信息评估数据质量
  - 利用用户行为模式识别垃圾网页
- 搜索引擎层面
  - 利用用户查询行为进行搜索引擎查询性能评估
  - 利用用户访问信息构建网络信息检索语料

基于用户行为分析的信息需求识别

- 用户信息需求分类
  - 目的：依照信息需求对查询进行不同处理
  - 用户查询分类体系(Broder & Rose et al.)
    - 面向导航类需求的用户查询
      - 用户检索时具有确定的检索目标页面
      - 查找某个已知存在的页面/资源
    - 面向信息事务类需求的用户查询
      - 用户检索时没有确定的检索目标页面
      - 查找与某个主题相关的页面/资源

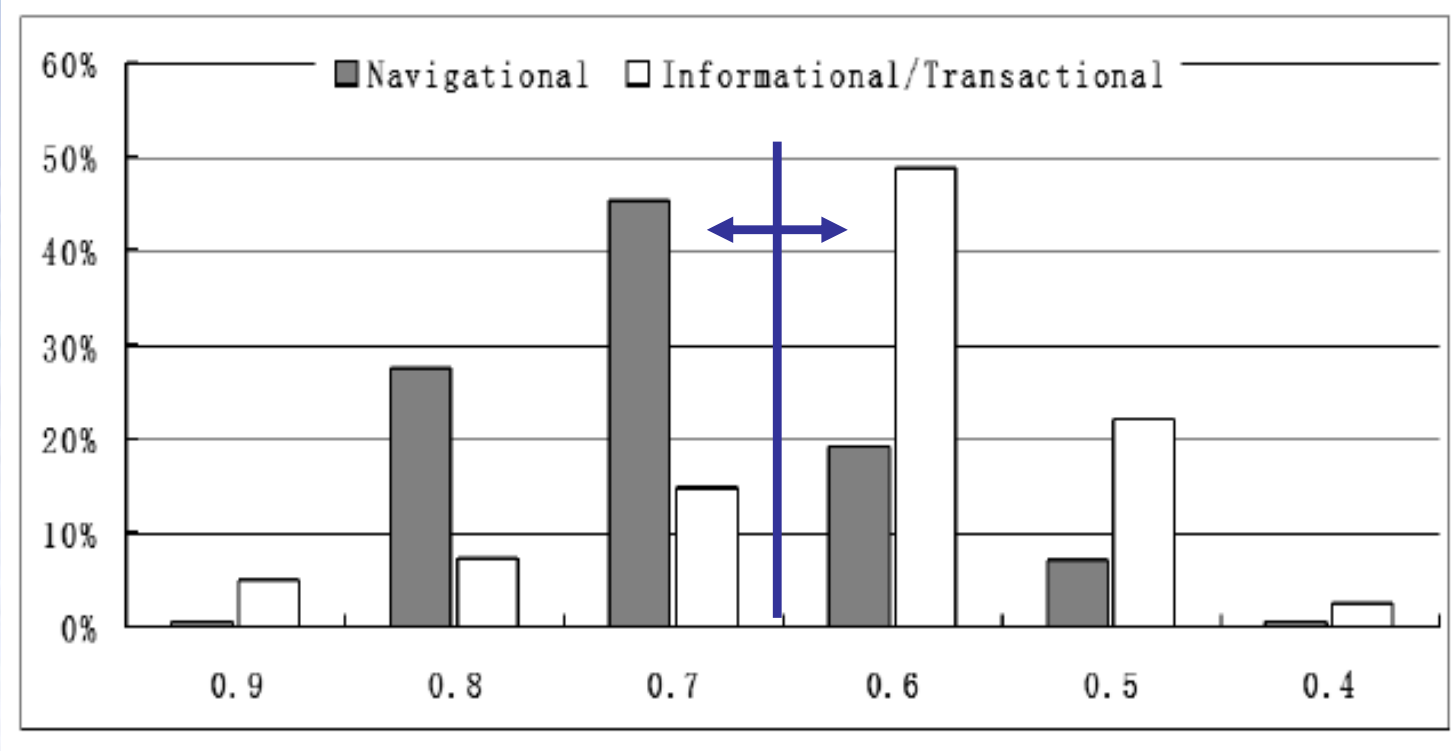
基于用户行为分析的信息需求识别

- 针对查询历史行为信息的特征提取
  - 假设1（懒鬼假设）：用户的检索需求是导航类型时，一般他只会点击很少数的几个答案
    - 进行导航类检索时，用户意识中有一个比较明确的查找目标
    - 他只会在结果页面中重点浏览与这个查找目标非常相关的URL或摘要内容，而不会点击其他的结果。
  - 特征：点击n次就满足的比例（n clicks satisfied）

$$nCS(Query\ q) = \frac{\#(Session\ of\ q\ that\ involves\ less\ than\ n\ clicks)}{\#(Session\ of\ q)}.$$

基于用户行为分析的信息需求识别

- nCS的分布情况



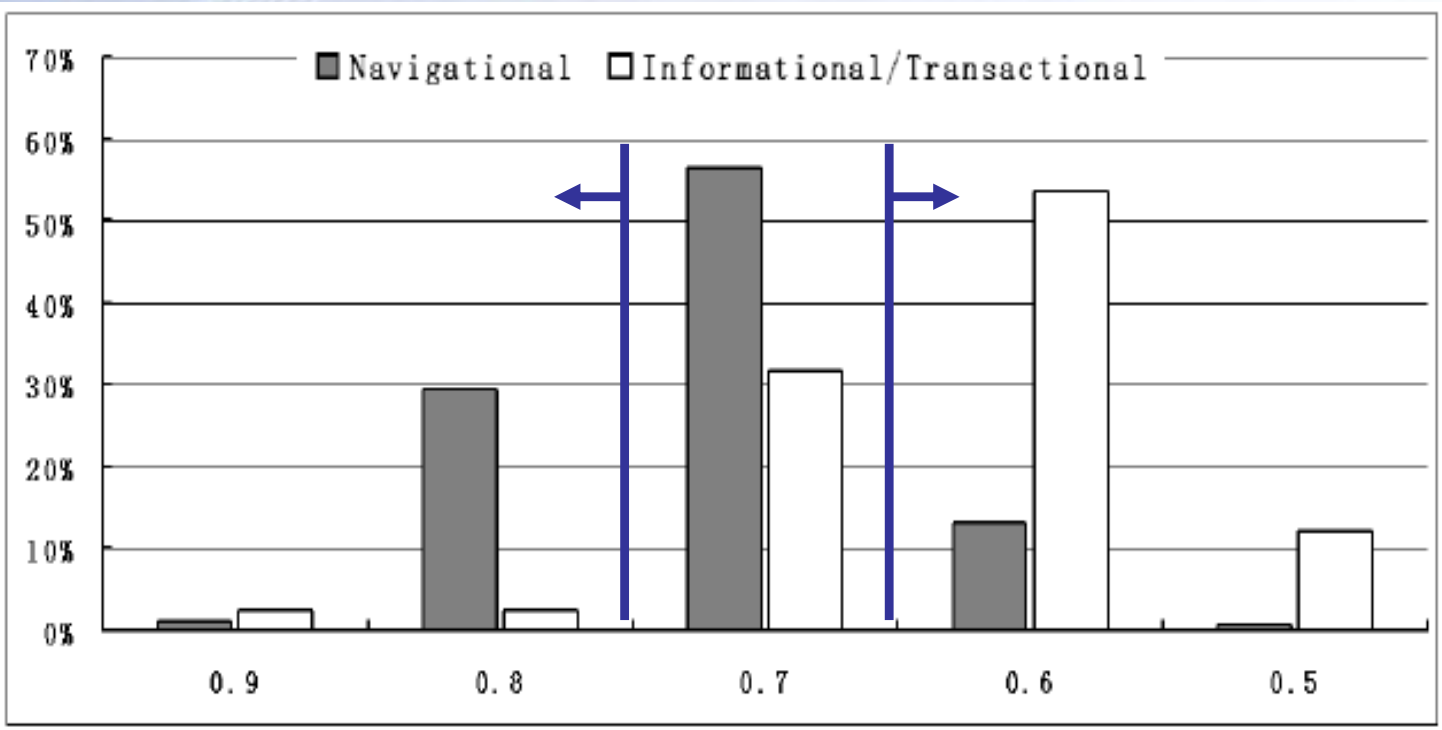
基于用户行为分析的信息需求识别

- 针对查询历史行为信息的特征提取
  - 假设2（封面假设）：用户的检索需求是导航类型时，一般他只会点击排名最靠前的几个答案
    - 检索系统导航类检索的性能一般都较高（MRR在80%以上）
    - 他很少有必要点击前几位之后的答案。
  - 特征：点击前n位就满足的比例（top n results satisfied）

$$nRS(Query\ q) = \frac{\#(Session\ of\ \boldsymbol{q}\ that\ involves\ clicks\ only\ on\ top\ \boldsymbol{n}\ results)}{\#(Session\ of\ \boldsymbol{q})}.$$

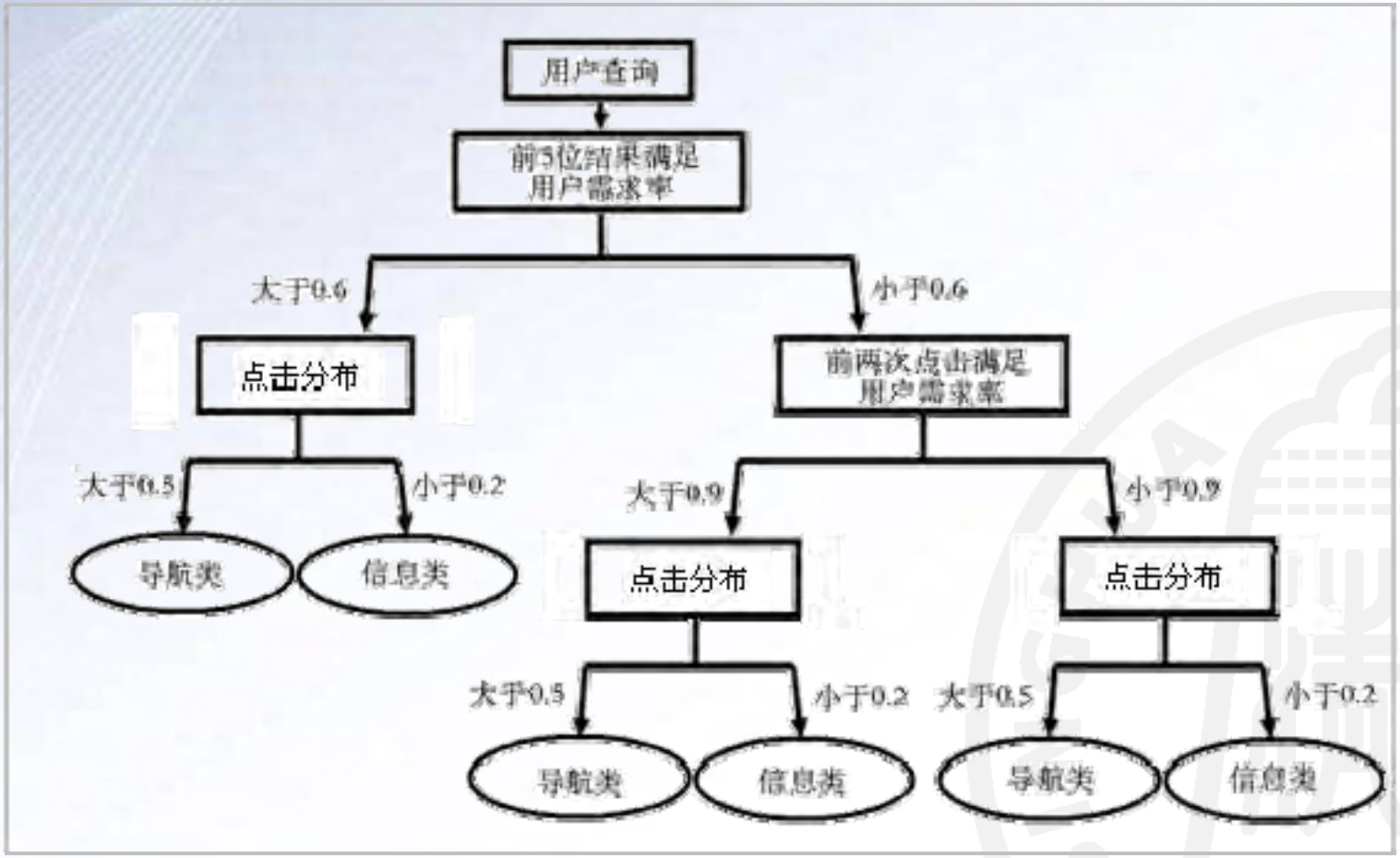
基于用户行为分析的信息需求识别

- nRS的分布情况



基于用户行为分析的信息需求识别

- 基于决策树学习的分类算法



基于用户行为分析的信息需求识别

- 识别结果
  - Sogou 2006年2月全月查询和点击日志数据
  - 共86,538,613条点击，涉及26,255,952个用户session
  - 训练集：198个查询；测试集：233个查询

	训练集合			测试集合		
	信息事务类	导航类	综合	信息事务类	导航类	综合
精确率	76.00%	91.07%	87.65%	73.74%	85.62%	81.49%
召回率	66.67%	90.71%	85.25%	72.84%	86.18%	81.54%
F-measure	0.71	0.91	0.86	0.73	0.85	0.81

基于用户行为分析的查询推荐

- 用户查询 V.S. 信息需求
  - 长度短：英文搜索平均长度不超过3个单词
  - 内容意义混淆不明：打字、俱乐部
  - 信息需求不明确：魔兽争霸(下载? 资讯? 主页?)
- 查询推荐
  - 协助用户重新组织查询，明确信息需求。
  - 当前主要思路：从已有用户查询中查找与当前查询相似(内容、点击)的查询

基于用户行为分析的查询推荐

- 问题
  - 缺乏对用户信息需求的明确理解
  - 死结？　　查询词：WWW2010

#	Baidu	Google	Sogou
1	pes2010	2010国家公务员职位表	2010年国家公务员
2	qq2010	2010年国家公务员报名	2010发型
3	实况2010	2010国家公务员报名	2010年考研报名
4	实况足球2010	2010国家公务员报名	2010公务员报名
5	卡巴斯基2010	大相径庭	2010公务员考试

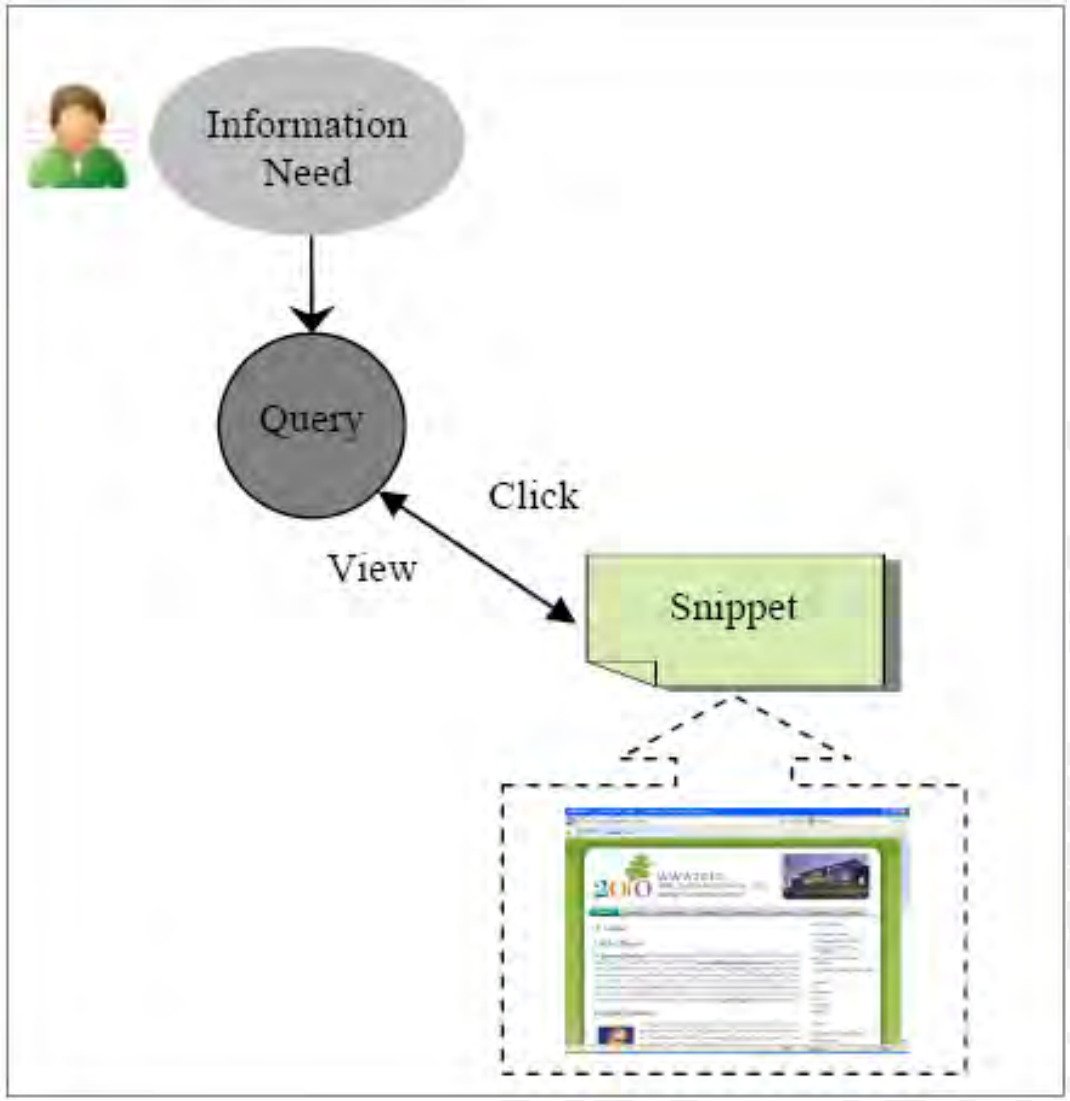
查询Q无法准确表述信息需求

搜索引擎查找与Q相似的查询推荐给用户

排序的结果与用户的需求有可能

基于用户行为分析的查询推荐

- 解决思路
  - 用户信息需求
    如何表达？
  - 用户进行点击时，
    并未阅读过页面
    的真实内容
  - 用户点击=>
    对结果页面摘要
    内容的兴趣



基于用户行为分析的查询推荐

- 实验结果
  - 结果样例：搜狗搜索引擎的查询推荐点击日志(2005.09)
  - 针对用户点击到的结果摘要进行关键词提取，生成的查询推荐内容能够吸引更多的用户点击
  - 评价指标：点击比率

搜狗搜索引擎推荐结果	是否被算法推荐	实际用户点击
死亡先知	0	0
d o t a先知	0	0
先知出什么	0	0
先知装备	0	0
先知的圣物		
灾难先知		
电影先知		
先知下载		
永恒先知之戒		
先知出什么装备		

	CTR improvement
Baidu	+32.80%
Sogou	+34.15%

其他推荐结果：

尼古拉斯 凯奇

高清

纪伯伦 （西方著名预言家）

用户的查询行为信息分析

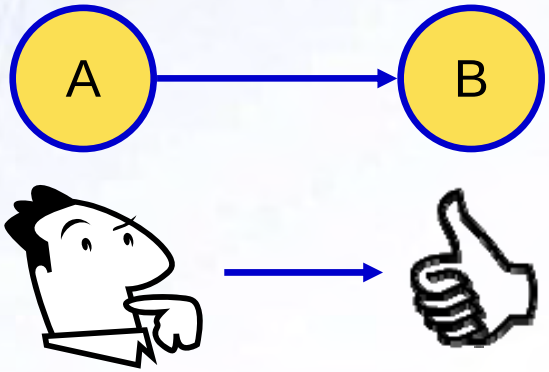
- Yiqun Liu, Min Zhang, Liyun Ru and Shaoping Ma，Automatic Query Type Identification Based on Click Through Information, Asia Information Retrieval Symposium, AIRS 2006,
- Bo Zhou, Min Zhang, Shaoping Ma, Yiqun Liu, Liyun Ru, Log-Mining Based Query Spelling Correction for Chinese Search Engines, Journal of Computational Information Systems, Volume 5, Number 3, pp1225-1234，2009.
- Bo Zhou, Min Zhang, Shaoping Ma, Yiqun Liu, Liyun Ru, Query Spelling Correction For Multi-Language Search Engines, Journal of Computational Information Systems, Volume 5, Number 3, pp1521-1528，2009.

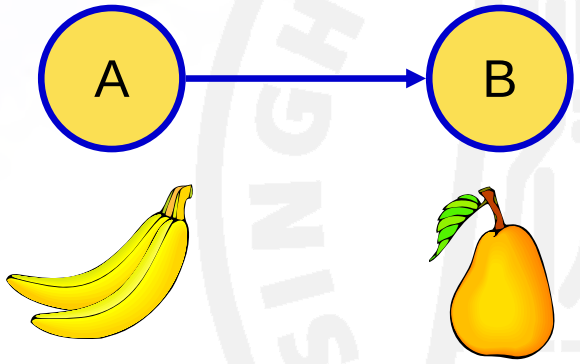
面向搜索引擎的用户行为分析方法

- 用户层面
  - 利用用户的查询行为信息识别信息需求类别
  - 利用用户的查询行为信息进行查询推荐
- **万维网层面**
  - **利用用户访问信息评估数据质量**
  - **利用用户行为模式识别垃圾网页**
- 搜索引擎层面
  - 利用用户查询行为进行搜索引擎查询性能评估
  - 利用用户访问信息构建网络信息检索语料

基于用户行为分析的数据质量评估

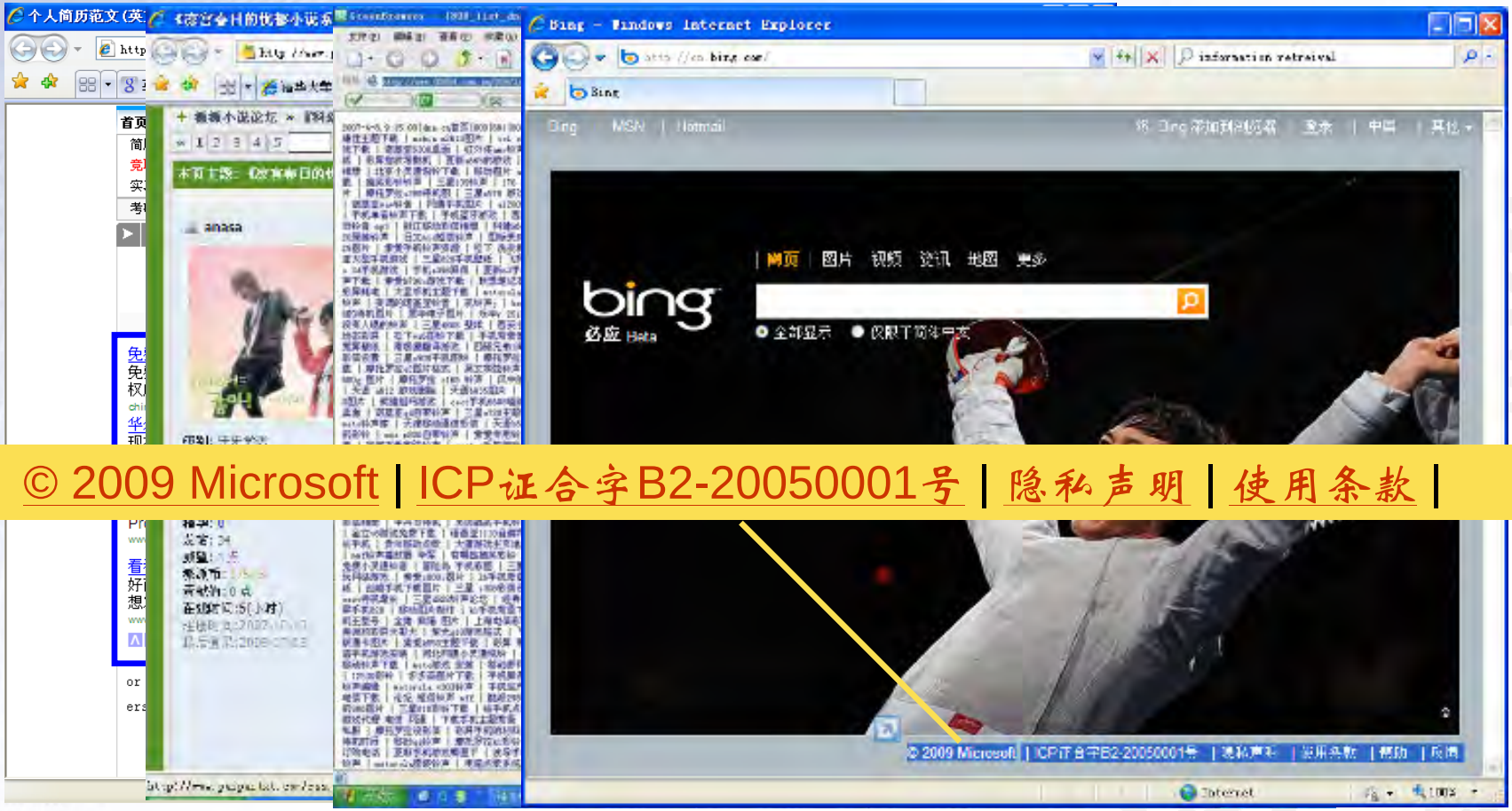
- 现状：链接结构分析算法为主
  - 超链接在被链接的两个网页之间建立如下关系：
    - 内容推荐关系：页面A的作者推荐页面B的内容，且利用L的链接文本内容对B进行描述。
    - 主题相关关系：被超链接连接的两个页面A与B比随机抽取的两个页面有更大的概率有内容相关性。





基于用户行为分析的数据质量评估

- 以链接结构分析为基础的质量评估
  - 链接结构数据本身质量存在问题



基于用户行为分析的数据质量评估

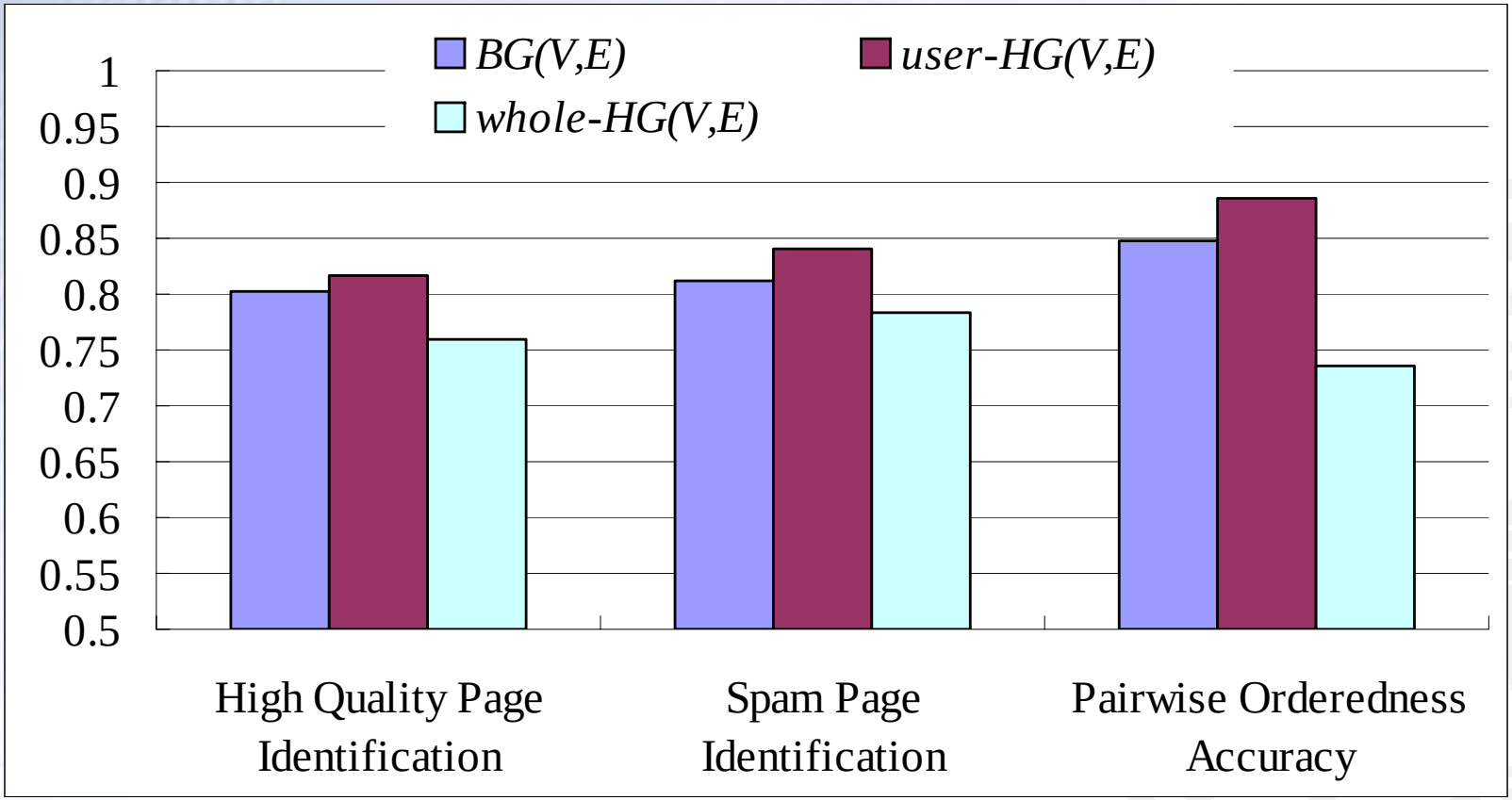
Web Site	Ranked by PageRank on SogouT	Ranked by Alexa.com traffic rank in China
www.adobe.com	1	139
www.hd315.gov.cn	北京市工商行政管理局	1,655
www.qq.com	3	2
labs.adobe.com	4	139
www.tencent.com	5	1,062
www.baidu.com	6	1
www.miibeian.gov.cn	信息产业部ICP/IP地址信息备案管理	179
blog.sohu.com	8	8
www.sina.com.cn	9	3
www.xinhuanet.com	10	42

基于用户行为分析的数据质量评估

- 解决思路
  - 依靠用户行为对链接结构数据进行清理
    - 用户点击：个人兴趣、信息需求
    - 被用户点击的网页/链接比未被点击的部分更可靠
  - 构建方式
    - User Browsing Graph：只保留用户访问过的网页和用户访问过的链接
    - User-oriented Hyperlink Graph：只保留用户访问过的网页，以及这些网页之间原始的链接关系

基于用户行为分析的数据质量评估

- PageRank性能测试
  - ROC/AUC测试、网站对质量测试

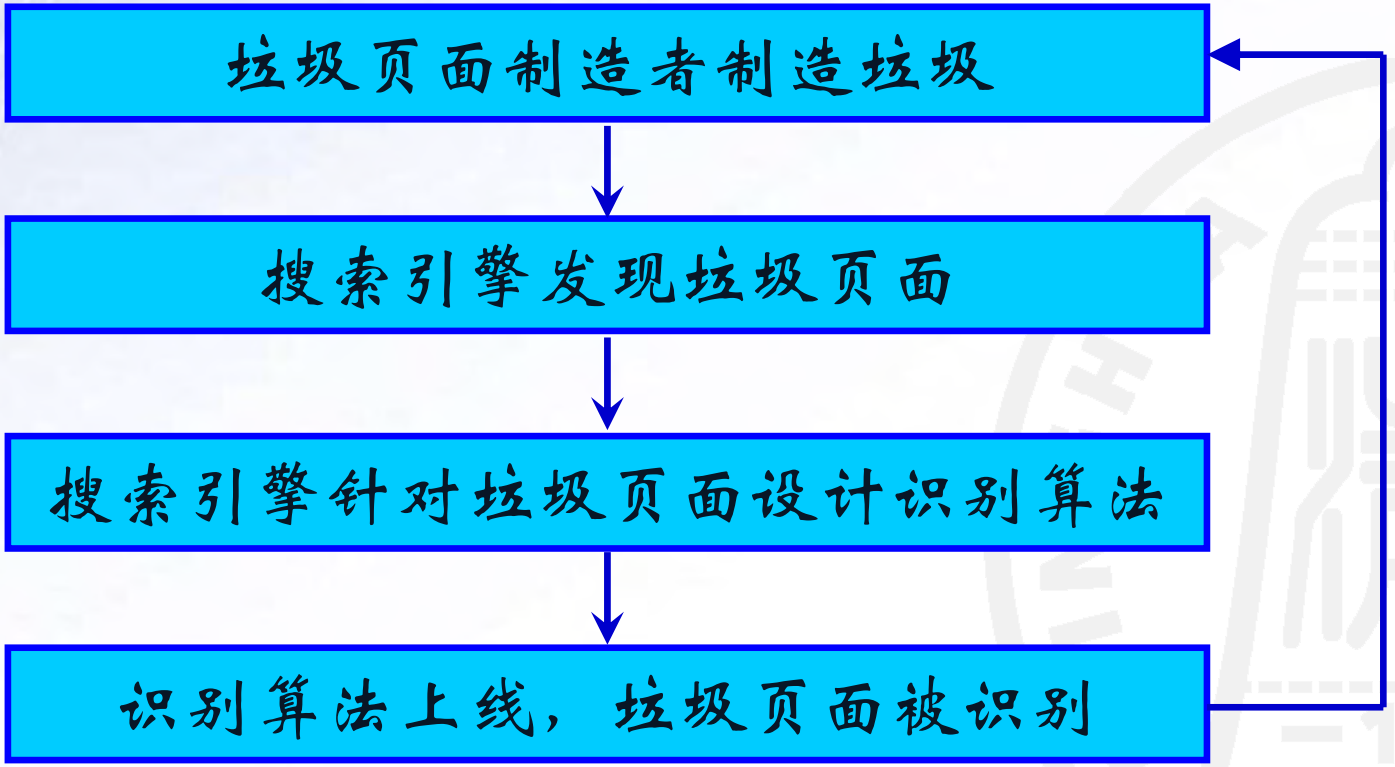


基于用户行为分析的数据质量评估

- Yiqun Liu, Yijiang Jin, Min Zhang, Shaoping Ma and Liyun Ru. User Browsing Graph: Structure, Evolution and Application. Late breaking result session in WSDM '09.
- 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).
- Yiqun Liu, Yufei Xue, Rongwei Cen, Min Zhang, Shaoping Ma and Liyun Ru, Web Page Quality Estimation with User Behavior Analysis. Submitted to ACM Tran. Web.

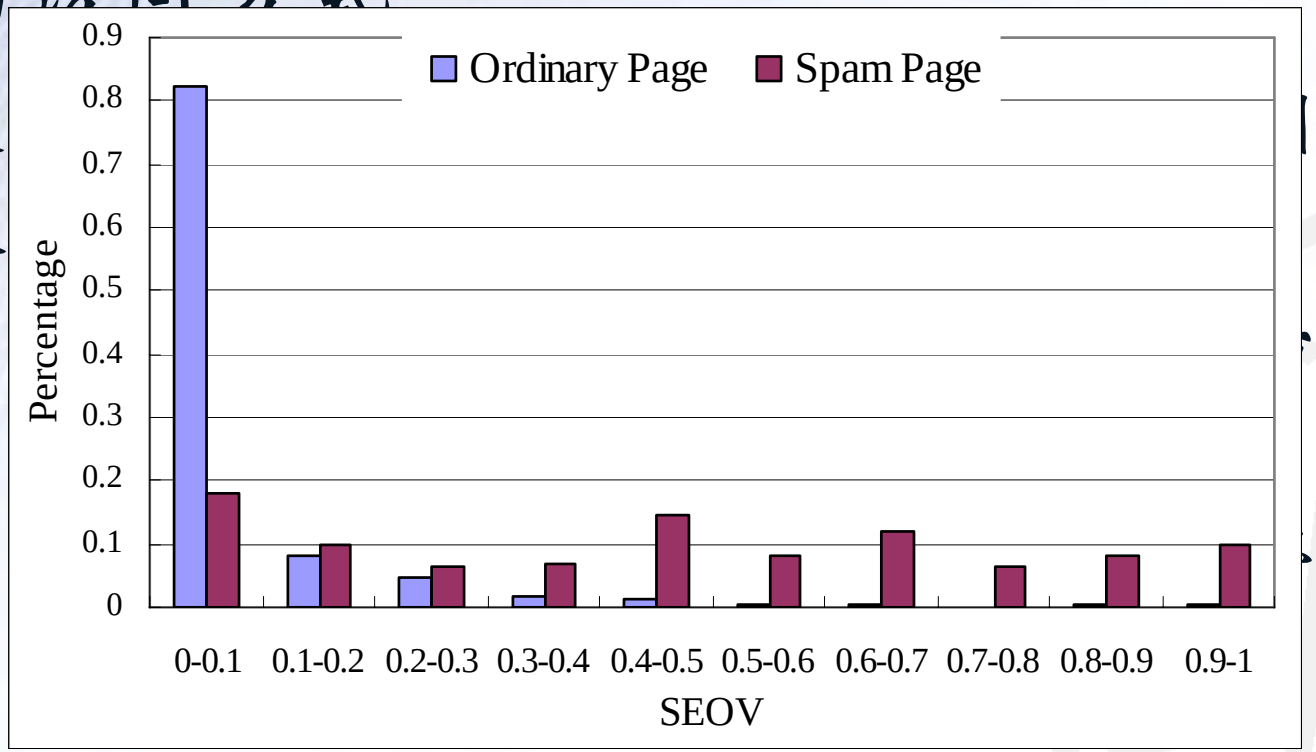
基于用户行为分析的垃圾页面识别

- 垃圾页面：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
- 传统识别方法：针对垃圾页面作弊手段



基于用户行为分析的垃圾页面识别

- 解决思路: 用户是垃圾网页最直接的受害者
- 用户对垃圾的访问方式不同于其对正常页面的访问方式
  - 搜 量主要是
  - 页 垃圾网页
  - 网 留较长时
  - ……



基于用户行为分析的垃圾页面识别

- 准确性
  - P@300: 94.0%，ROC/AUC: 0.9150
- 通用性
  - 不局限作弊形式，能够发现新出现的作弊形式
- 时效
  - 算
  - 3月
  - 59

Page Type	Percentage
Non-spam pages	6.00%
Web spam pages (Term spamming)	21.67%
Web spam pages (Link spamming)	23.33%
Web spam pages (Other spamming)	10.67%
Pages that cannot be accessed	38.33%

及网站

日索引量：

基于用户行为分析的垃圾页面识别

- Yiqun Liu, Min Zhang, Shaoping Ma, Liyun Ru. User behavior oriented web spam detection. In Proceeding of WWW '08.
- 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.
- Huijia Yu, Yiqun Liu, Min Zhang, Liyun Ru, Shaoping Ma, Web Spam Identification with User Browsing Graph. To be appeared in AIRS ‘09

面向搜索引擎的用户行为分析方法

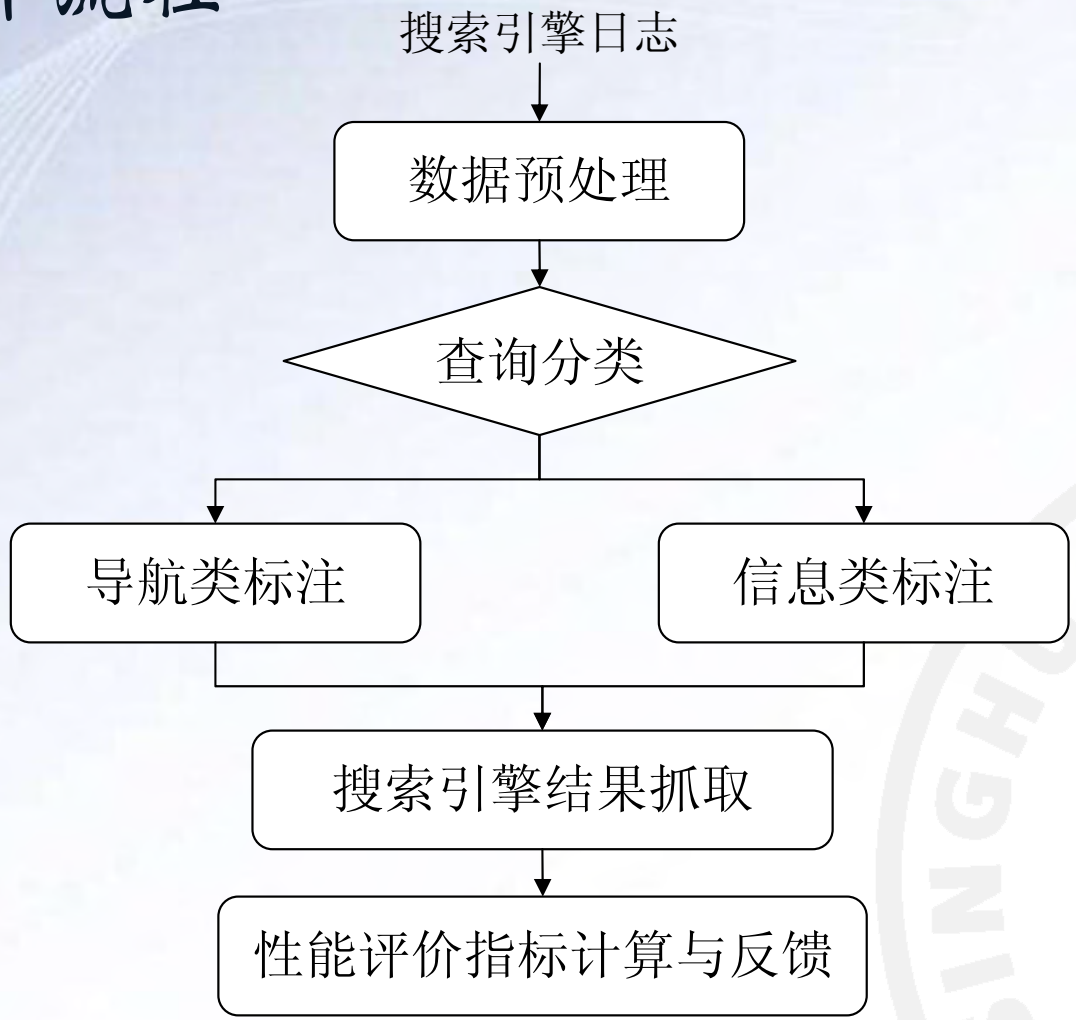
- 用户层面
  - 利用用户的查询行为信息识别信息需求类别
  - 利用用户的查询行为信息进行查询推荐
- 万维网层面
  - 利用用户访问信息评估数据质量
  - 利用用户行为模式识别垃圾网页
- 搜索引擎层面
  - 利用用户查询行为进行搜索引擎查询性能评估
  - 利用用户访问信息构建网络信息检索语料

基于用户行为分析的搜索引擎评价

- 传统方式：基于手工进行答案标注
  - 不同标注人员的判定标准差异： TREC 2008某个子任务中，有58%的文档标注人员观点不一致
  - 人力资源成本问题：一个规模为800万的文档集合，针对1个查询主题的相关性评判，需耗费1名标注人员9个月的工作时间
- 解决方案：
  - 通过对用户行为日志进行分析，无需标注
  - 使用群体，而不是个体的点击行为作为依据

基于用户行为分析的搜索引擎评价

- 自动评价流程



基于用户行为分析的搜索引擎评价

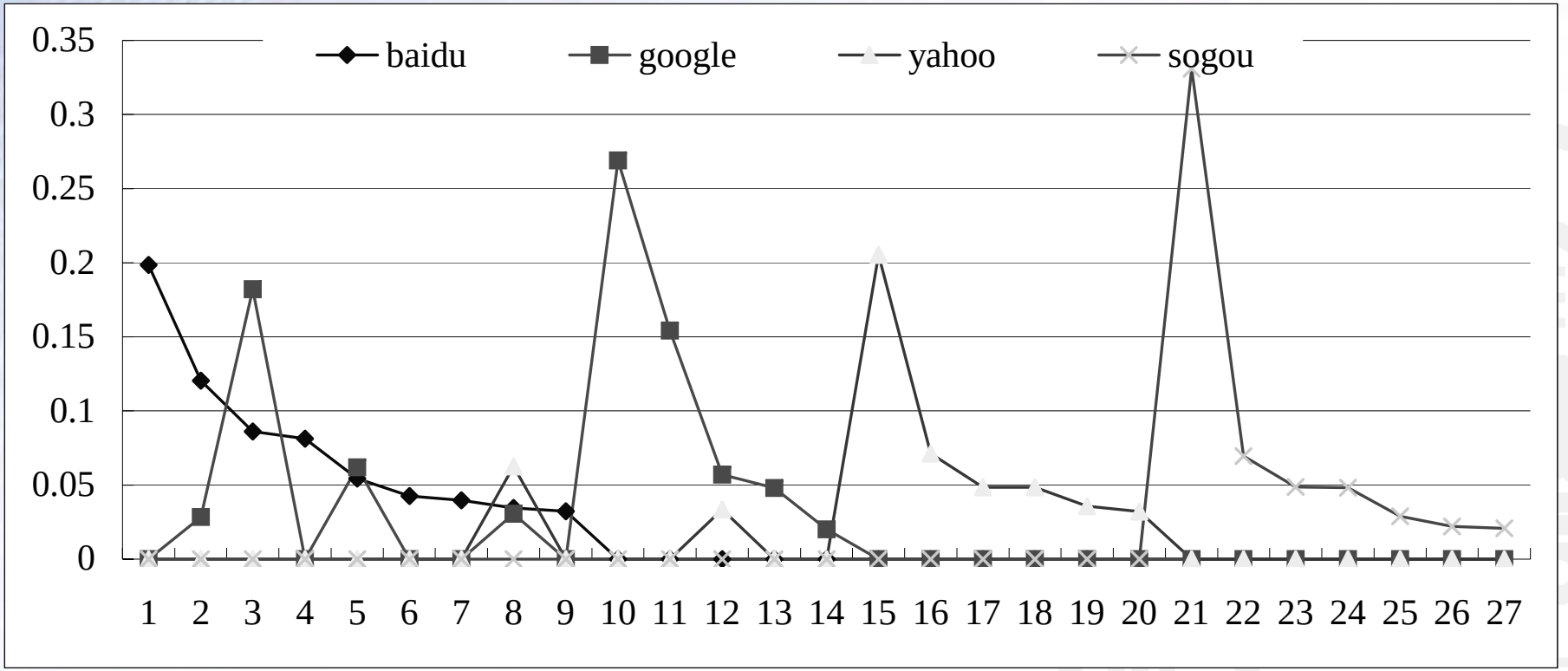
- 针对导航类查询的结果自动标注
  - 利用单个搜索引擎的点击信息即可完成
  - 焦点假设：不同用户具有相同的导航类别检索需求时，他们的点击都会集中在其检索目标网页（或其镜像）上。
  - 网页r针对查询q的点击集中度

$$ClickFocus(Query\ q, Result\ r) = \frac{\#(Session\ of\ \ q\ that\ clicks\ r)}{\#(Session\ of\ \ q)}$$

  - q的点击集中度最高的r即为其检索目标页面

基于用户行为分析的搜索引擎评价

- 针对信息事务类查询需求的答案自动标注
  - 以查询词"电影"为例
    - 不同搜索引擎的点击分布差异大



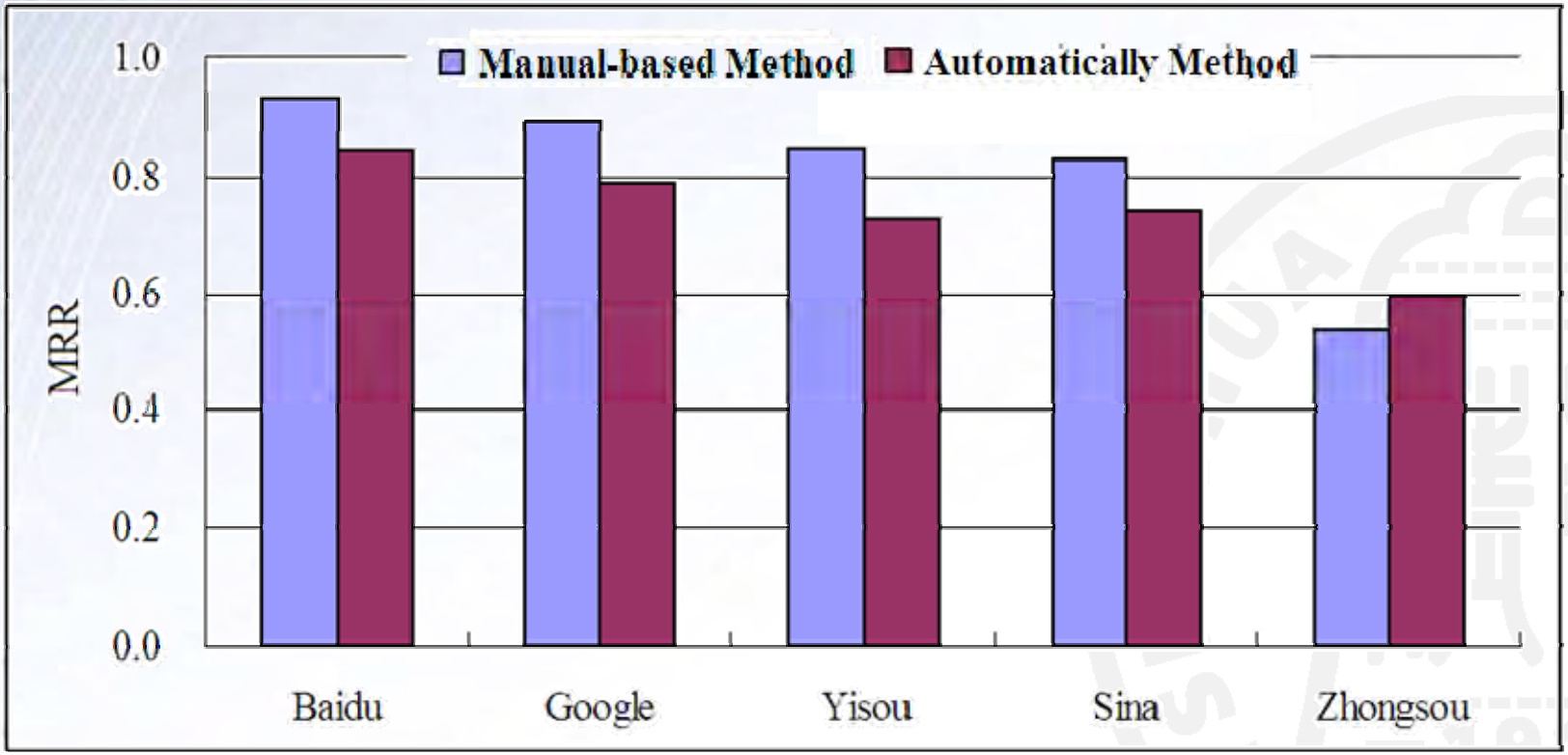
基于用户行为分析的搜索引擎评价

- 针对信息事务类查询需求的答案自动标注
  - 基于多搜索引擎用户行为挖掘
    1. 利用单搜索引擎用户行为进行各自独立的标注
    2. 借鉴Pooling做法，综合不同标注者（这里为搜索引擎用户的宏观行为）的意见
  - 需要考虑的因素
    - 用户点击行为差异、用户访问量差异、搜索引擎相对重要性的差异

$$P(url_i \mid q) = \sum_j P(url_i \mid SE_j, q) P(SE_j \mid q)$$

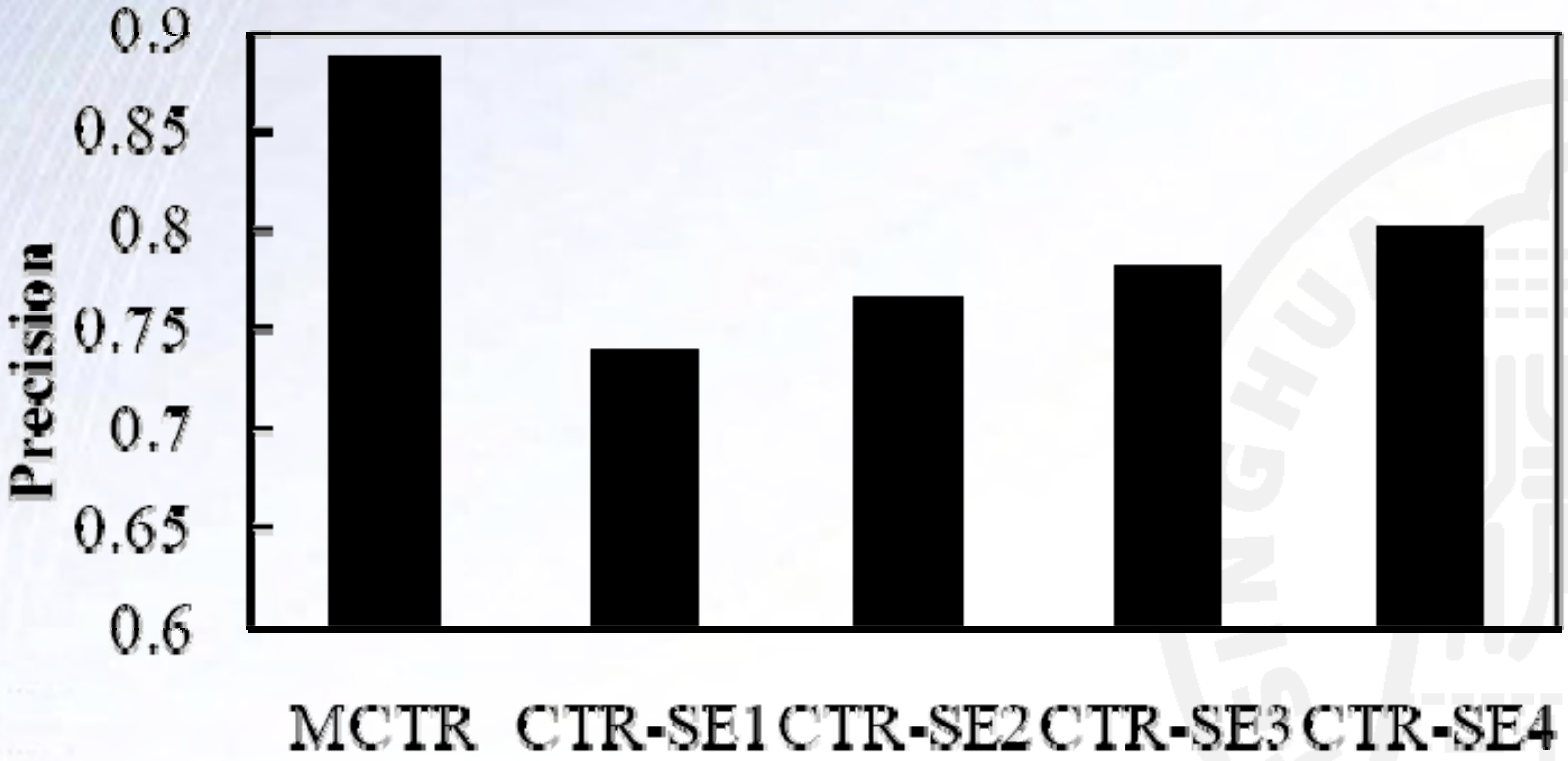
基于用户行为分析的搜索引擎评价

- 实验数据
  - Sogou搜索8个月查询日志(超过7亿条日志信息)
  - 导航类评测结果：相关系数达到0.965



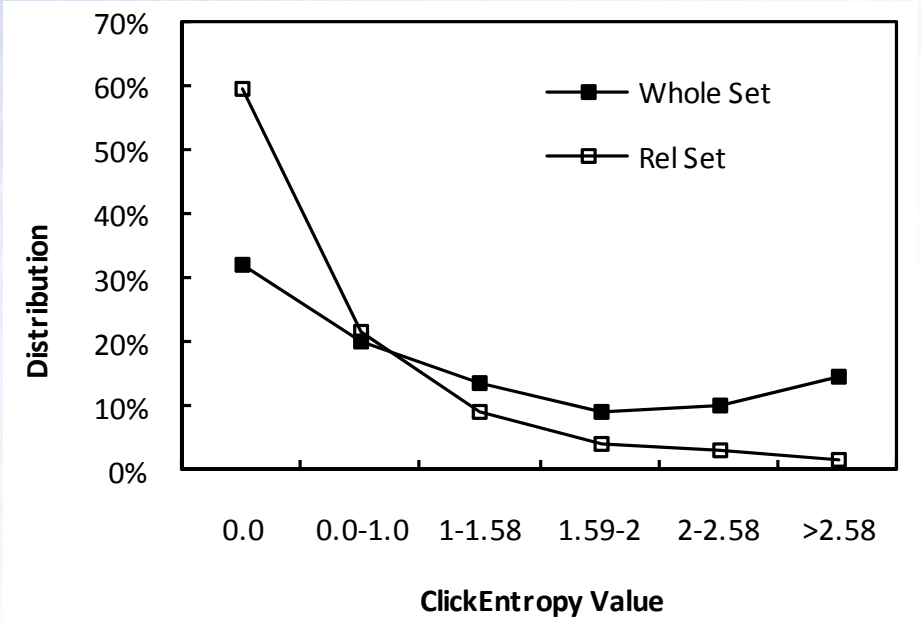
基于用户行为分析的搜索引擎评价

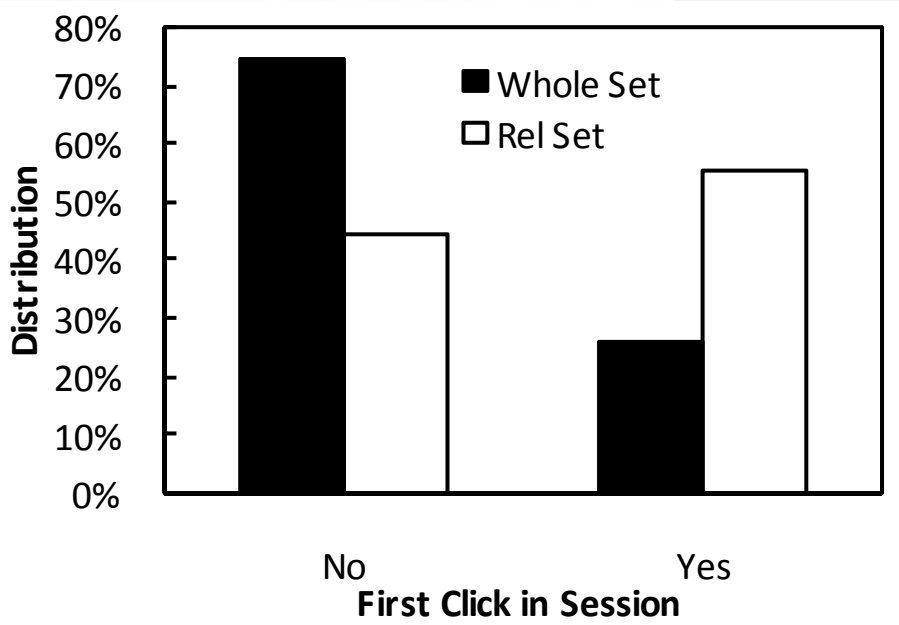
- 实验数据
  - 信息事务类评测结果
  - 查询—结果对标注准确率近90%



基于用户行为分析的搜索引擎评价

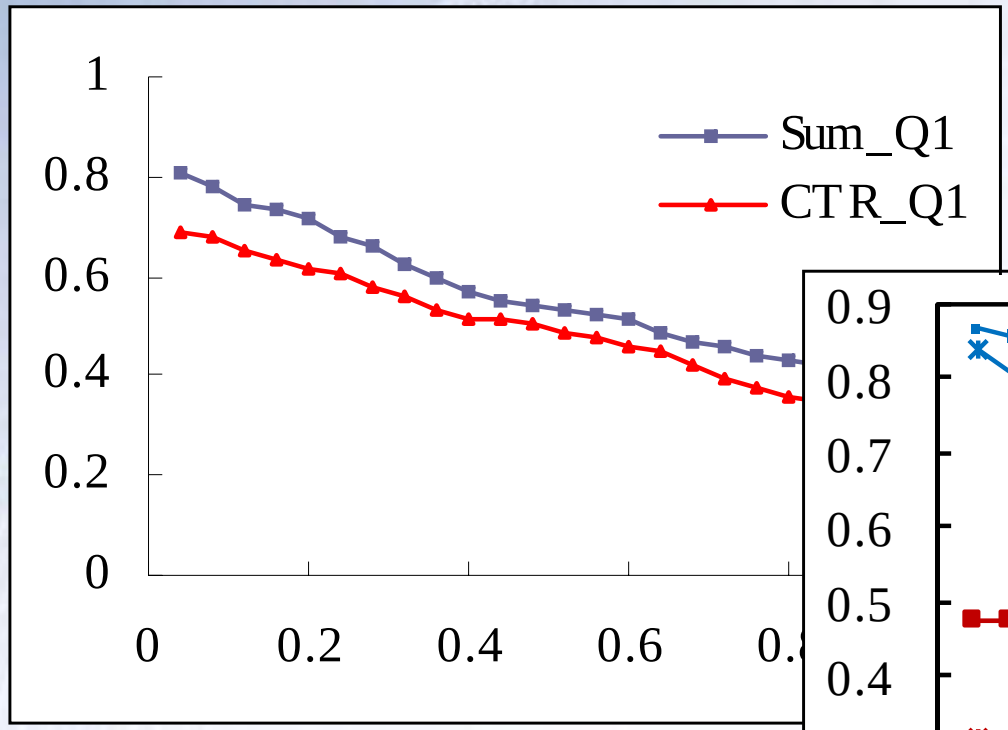
- 如何处理只有较少量用户点击的冷门查询？
  - 去除非正常的用户点击（abnormal user sessions）
  - 评价用户点击的可靠性



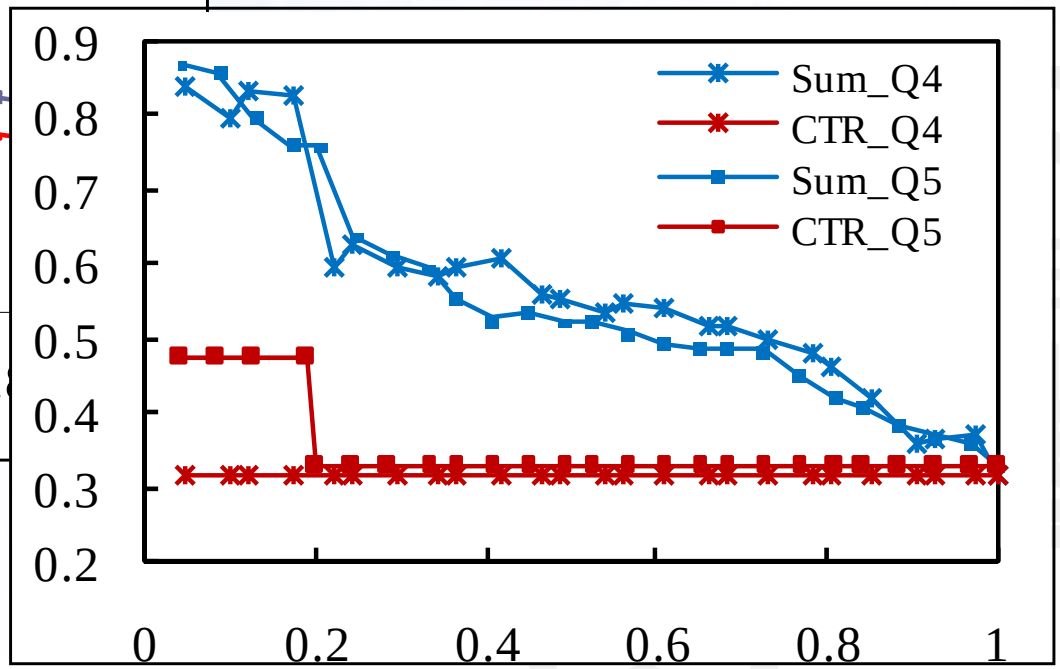


基于用户行为分析的搜索引擎评价

- 基于用户可靠性分析的冷门查询性能评价



热门查询性能



冷门查询性能

基于用户行为分析的搜索引擎评价

- 中文搜索引擎性能评价平台“搜索仪”

网络信息检索语料库构建

- 评测语料构成
  - 文本语料：1.387亿网页，存储空间约5 Terabyte，对应的链接关系数据和SogouRank数据。
  - 查询语料：2008年6月查询量最大的10000个用户查询，占当月用户查询总量的56%。
  - 标注语料：利用用户行为分析技术，自动标注65465个答案，抽样检查发现正确率在95%左右
  - 满足中文互联网信息检索研究各方面需求

网络信息检索语料库构建

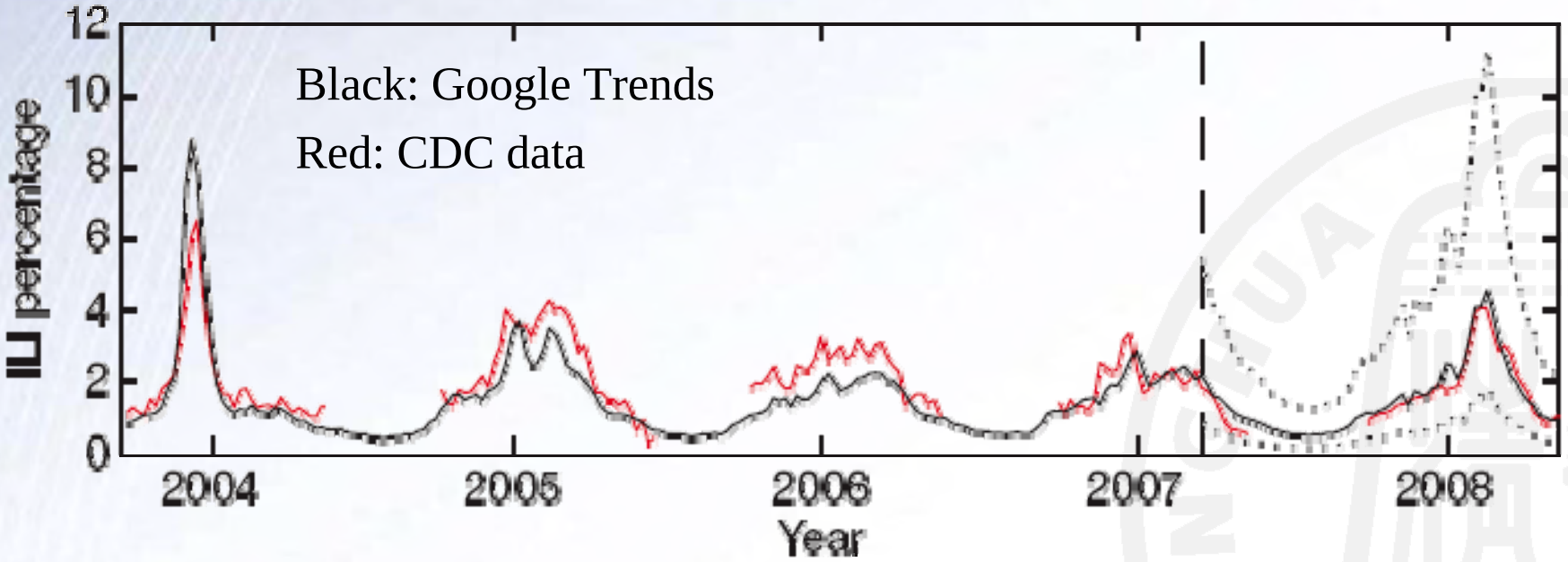
- 目前已经发放近40个拷贝

基于用户行为分析的搜索引擎评价

- Yiqun Liu, Yupeng Fu, Min Zhang, Shaoping Ma, Liyun Ru. Automatic Search Engine Performance Evaluation with Click-through Data Analysis. Proceedings of WWW ’07.
- Rongwei Cen, Yiqun Liu, Min Zhang, Liyun Ru, Bo Zhou, Shaoping Ma. Exploring Relevance for Clicks. Proceedings of CIKM ‘09
- Rongwei Cen, Yiqun Liu, Min Zhang, Liyun Ru, Shaoping Ma. Study on the Click Context of Web Search Users for Reliability Analysis. To Proceedings of AIRS ‘09.
- Rongwei Cen, Yiqun Liu, Min Zhang, Liyun Ru, Shaoping Ma. Automatic Search Engine Performance Evaluation with the Wisdom of Crowds. Proceedings of AIRS ‘09.

站在搜索引擎的角度观察世界

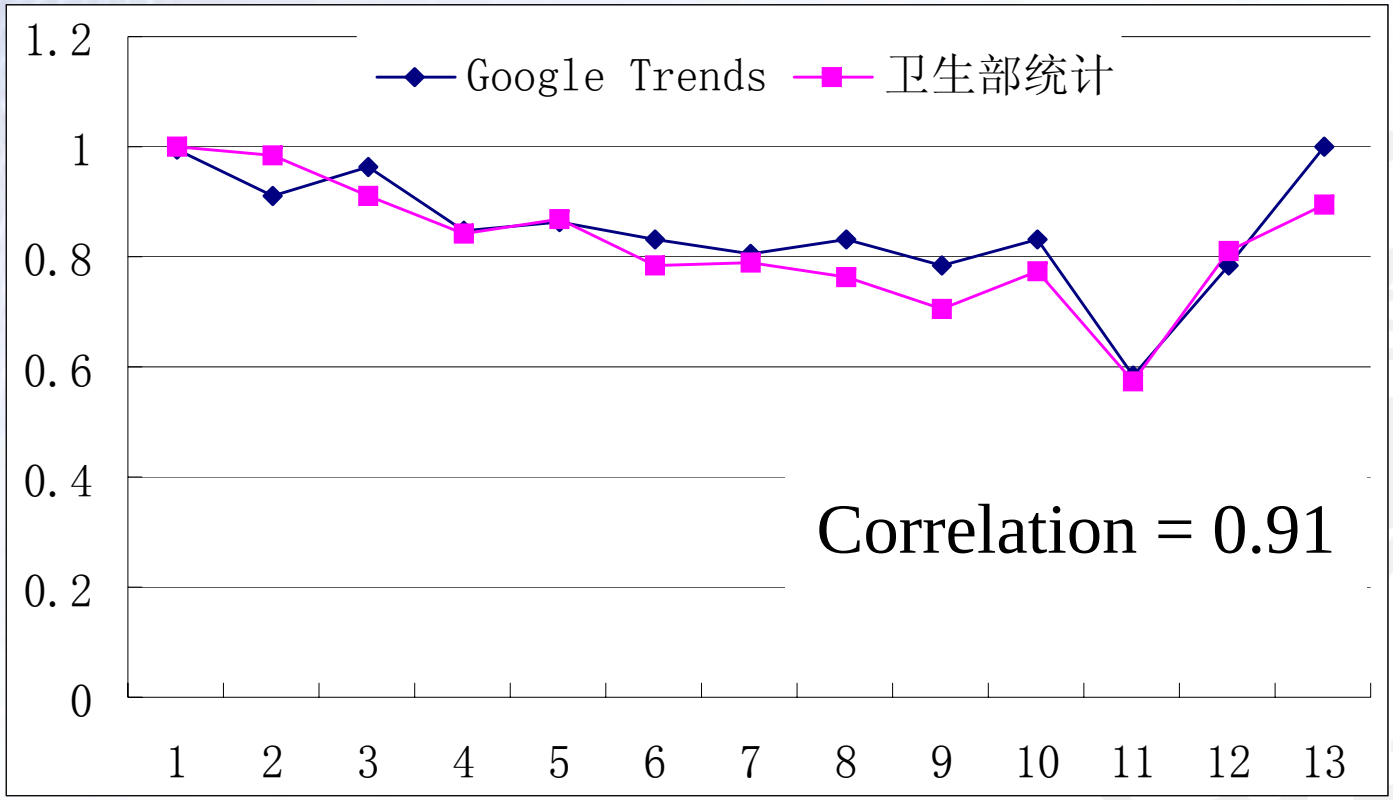
- 流感发病趋势预测
  - http://www.google.com/trends/flu
  - 当地查询日志可以用于预测此地流感发病趋势



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

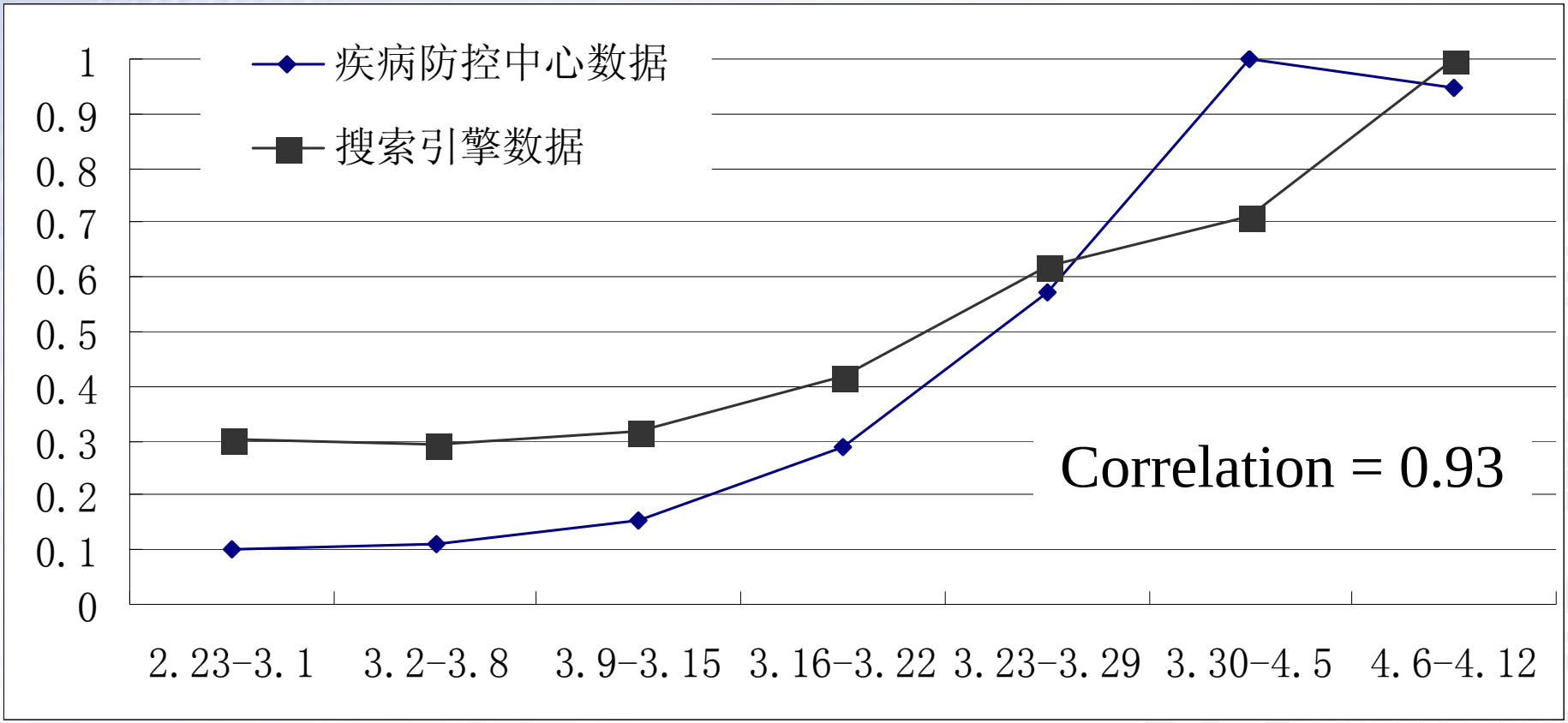
站在搜索引擎的角度观察世界

- Google trends实验
  - 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



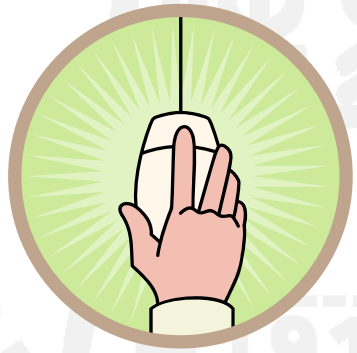
站在搜索引擎的角度观察世界

- 手足口病发病趋势预测
  - 北京市疾病防控中心数据 vs.
    来自北京IP的Sogou疾病症状查询数据



面向搜索引擎的用户行为分析

- 总结
  - 改进搜索引擎算法
  - 评价搜索引擎性能
- 未来工作
  - 用户信息需求分析
    - 心理学模型，社会文化模型的融合
  - 基于用户行为方法理解互联网数据
  - 站在搜索引擎的角度观察世界



THE END

Thank you!

Questions or comments?

http://www.thuir.cn/