



基于用户群体行为分析的 网络信息检索研究

智能技术与系统国家重点实验室

答辩人：刘奕群

导师：马少平 教授

2007年6月12日





提纲

- 研究背景
- 基于用户群体行为分析的网页质量评估
- 基于用户群体行为分析的查询需求分类
- 基于用户群体行为分析的检索效果评价
- 网络信息检索系统架构设计的改进尝试
- 结论





Information Systems © Tsinghua University

研究背景

- Web的发展带来了什么？

- 信息量的急剧增长

- 1994年，面向个人用户的浏览服务诞生
 - 2002年，Web上所存储的数据超过500,000 TB



- 知识的获取空前简单与繁荣

- “在信息化时代，**知识实际上已经不是资源，智慧才是资源。**”

- 从Web中有效的获取知识成为人们的必需技能

- 高科技企业员工1/3的时间用于查找资料



Information Systems © Tsinghua University

研究背景

- 网络信息检索工具成为人们生活和工作中的重要信息获取手段
 - 全球范围内84%的互联网用户使用搜索引擎，其中超过一半的人几乎每天使用。
 - 2007年1月，我国网民总人数为13700万人。
 - 95%以上的中国网民使用过搜索引擎，84.5%的用户将搜索引擎作为得知新网站的主要途径。
 - 商用搜索引擎竞争日趋激烈。





研究背景

• 搜索引擎发展面临的困境

- 无用信息过多？

liuyiqun03@gmail.com | [我的帐户](#) | [退出](#)

网页 图片 资讯 地图新! 更多»

Google 清华大学 Google 搜索 高级搜索 | 使用偏好

☒ 所有网页 ☐ 中文网页 ☐ 简体中文网页 ☐ 中国的网页

网页 约有7,790,000项符合清华大学的查询结果，以下是第1-10项 （搜索用时 0.04 秒）

[欢迎光临清华大学](#)
招生信息、院系设置、新闻动态。
www.tsinghua.edu.cn/ - 10k - [类似网页](#)

赞助商链接
[企业接班人](#)
清华大学企业继承人高级研修班

- 有用信息过少？

liuyiqun03@gmail.com | [我的帐户](#) | [退出](#)

网页 图片 资讯 地图新! 更多»

Google m608c待机时间 Google 搜索 高级搜索 | 使用偏好

☒ 所有网页 ☐ 中文网页 ☐ 简体中文网页 ☐ 中国的网页

网页 约有110,000项符合m608c待机时间的查询结果，以下是第1-10项 （搜索用时 0.34 秒）

[索爱m608c待机时间 voogaa](#)
索爱m608c待机时间... 名称: 索爱m608c待机时间. 作者: 佚名. 格式: 支持多种格式. 大小: 123 KB. 下载, 直接发送到手机. 支持: 索爱m608c待机时间支持移动, 联通. 上: 索爱k750待机图片(312224); 下: 索爱w810c待机图片(212226) ...
www.voogaa.cn/voowww/voo.asp?g=12225 - 6k - [补充材料](#) - [类似网页](#)





Information Partner of Tsinghua University

研究背景

- 搜索引擎发展面临的困境
 - 结果排序不合理

Baidu 百度 新闻 网页 贴吧 知道 MP3 图片

加拿大签证申请

百度一下 结果中找 帮助 | 高级搜索

把百度设为首页 百度一下，找到相关网页约149,000篇，用时0.060秒

[中旅新途旅行网 优惠多多惊喜多多](#) 赞助商链接
www.sintu.com/visa/C专业级的旅游网站.专业办理加拿大签证,最新加拿大特惠酒店推荐...

[金东方自在移民网助您快速办理加拿大签证](#) 赞助商链接
www.egofuture.com本公司长期致力于加拿大、澳大利亚、新西兰、马来西亚等国家的移...

[父母赴加拿大探亲签证申请所需材料的清单](#)
加移民部长:没听说中国学生签证难 中“父母赴加拿大探亲签证申请所需材料的清单”相关内容 中“父母赴加拿大探亲签证申请所需材料的清单”相关内容 中“父母赴加拿大探亲签证申请所需材料的清单”相关内容 中“父母赴加拿大探亲签证申请所需材料...”
www.visawang.com/visa/Class22/200503/723.html 40K 2007-5-28 - 百度快照

[加拿大探亲访友和商务签证申请资格及收费标准 \(存档\) - Wayabroad](#)
二、商务签证申请资格 1. 申请人有加拿大商业往来的邀请信或传真。 2、有资金证明能够支付在加拿大的开销,或有证明...赴加留学签证被拒申诉有捷径 赴加留学签证慢但签证率上升至63% 申请留学加拿大签证 7月13日赴加留学签证纪实(高中...
www.chuguolu.com/pack/pack02/3932_2_2.html 12K 2007-3-10 - 百度快照

[科智留学 专业办理加拿大](#)
拥有十年丰富经验的科智留学是中国最著名的办理出国留学服务机...
www.kezhi.com.cn

[中国天鹅 加拿大移签证和](#)
中国天鹅 加拿大签证和移民的专业机构!国内唯一国家部委的中介...
immiworld.com

[枫叶之都专业的加拿大](#)
加拿大梦幻旅游集团中国总代理枫叶之都,协助办理各种加拿大签证
www.canada-travel.cn

[中旅新途旅行网 优惠多多](#)
专业级的旅游网站.专业办理加拿大签证,最新加拿大特惠酒店推荐...
www.sintu.com



研究背景

- 搜索引擎发展面临的困境

- 信息相对过剩：无用信息多，有用信息少
- 结果排序不合理：错误理解用户需求

- 关键问题：

网络数据处理过程与查询需求处理过程没有充分考虑检索用户的行为特点与信息需求

- 与搜索引擎的用户交互形式有关



研究背景

- 审视搜索引擎的用户交互方式
 - 传统文本信息检索系统的交互方式
 - 自然语言交互
 - MIT Start系统(<http://www.ai.mit.edu/projects/infolab/>)
 - AskJeeves搜索引擎(<http://www.askjeeves.com/>)
 - 特定查询语言交互
 - W3QL系统(Konopnicki 1995), WebSQL系统(Mendelzon 1997)
 - 搜索引擎：“关键词查询+选择性浏览”的方式
 - 提高了易用性
 - 符合当前搜索技术的发展水平





Information Systems @ Tsinghua University

研究背景

- 现有用户交互方式带来的问题
 - 关键词无法传递用户复杂的信息需求
 - 例：魔兽争霸
 - 导航类：定位游戏官方主页
 - 信息类：获取最新相关资讯
 - 资源类：查找游戏下载站点
 - 现有的检索系统用户交互方式带来了用户需求信息的传递瓶颈
 - 如何解决：针对检索用户群体行为信息进行分析 and 挖掘



研究背景

- 用户群体行为分析方法推动信息检索技术的发展
 - 指导搜索引擎算法改进
 - PageRank算法与其“随机浏览模型”
 - 进行搜索引擎性能评估
 - TREC的pooling方法，性能评价指标的确定等
- 基于日志进行用户群体行为特征分析
 - 方便：利用搜索引擎Web服务收集的日志信息
 - 可靠：减少个体行为的偶然性影响
 - 真实：采集过程对用户使用没有影响





研究背景

- 主要研究对象

- 网络信息检索系统的处理对象

- 网络数据：在网络数据的组织和处理过程中考虑到用户群体普遍的信息需求
 - 用户查询：在用户查询需求的处理过程中，充分利用用户行为的先验知识

- 网络信息检索系统的效果评价方法

- 效果评价：在性能评价策略、指标的制定中，充分考虑用户行为特点并使用用户群体的行为知识。



Information Systems © Tsinghua University

研究思路

瓶颈：关键词查
询+选择性浏览



用户群体行为
分析

研究重点：

使用用户群体行为分析方法，对网络信息检索系统处理**网络数据**和**查询需求**的过程进行改进和整合，并更准确的实现**效果评价**



Information Engineering @ Tsinghua University

提纲

瓶颈：关键词查
询+选择性浏览



用户群体行为
分析



网页质量评估





网页质量评估

- 研究背景：搜索引擎索引规模的竞争
 - 索引规模竞争是早期搜索引擎竞争的主战场

Search Engine	索引量	页面最大大小
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+

From Danny Sullivan, SearchEngineWatch web site





网页质量评估

- 搜索引擎索引规模的竞争（续）
 - 没有任何一个搜索引擎可以覆盖互联网上的所有资源

	Google	Yahoo!	MSN	Teoma
Round 1	76.30%	69.28%	62.03%	57.58%
Round 2	76.09%	69.29%	61.90%	57.69%
Round 3	76.27%	69.37%	61.87%	57.70%
Round 4	76.05%	69.30%	61.73%	57.57%
Average	76.16%	69.32%	61.90%	57.62%

A. Gulli, et al., The Indexable Web is More than 11.5 billion pages

- 2005年9月，Google从首页去除了页面索引数量的信息，并解释说：“绝对的数量已经不再重要”



网页质量评估

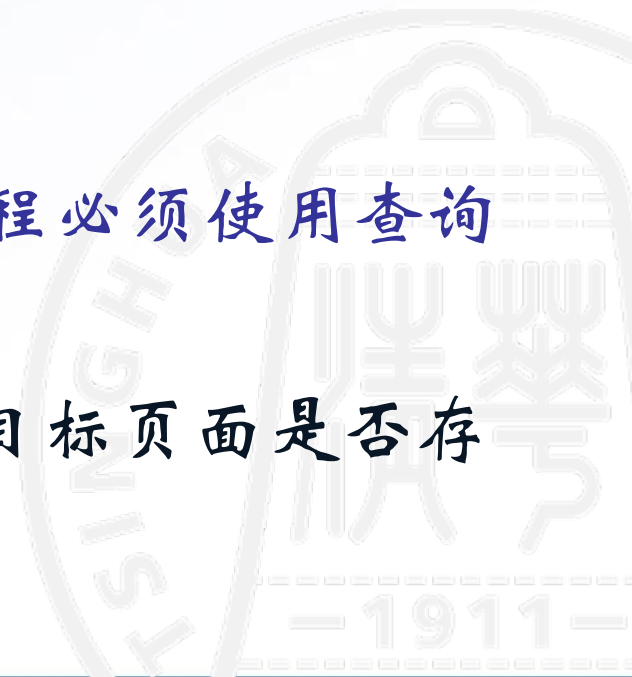
- 网络信息检索工具的存储与计算负担巨大，必须对资源予以合理利用
 - Yahoo!报告的索引量超过190亿页面
 - Google每天处理的用户需求超过2.7亿个
- 网页质量评估：搜索引擎面临的最大挑战之一。
- 质量评估当前工作：对链接结构的分析
 - PageRank, Topic-sensitive Pagerank, TrustRank...
- 问题
 - 链接数据本身的可靠性受到挑战
 - 忽略对用户真实需求的关注



Information Systems © Tsinghua University

网页质量评估

- 页面质量的定义
 - 以用户需求为中心：有可能满足用户群体的检索需求的页面（查询目标页面）才是高质量的
- 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？





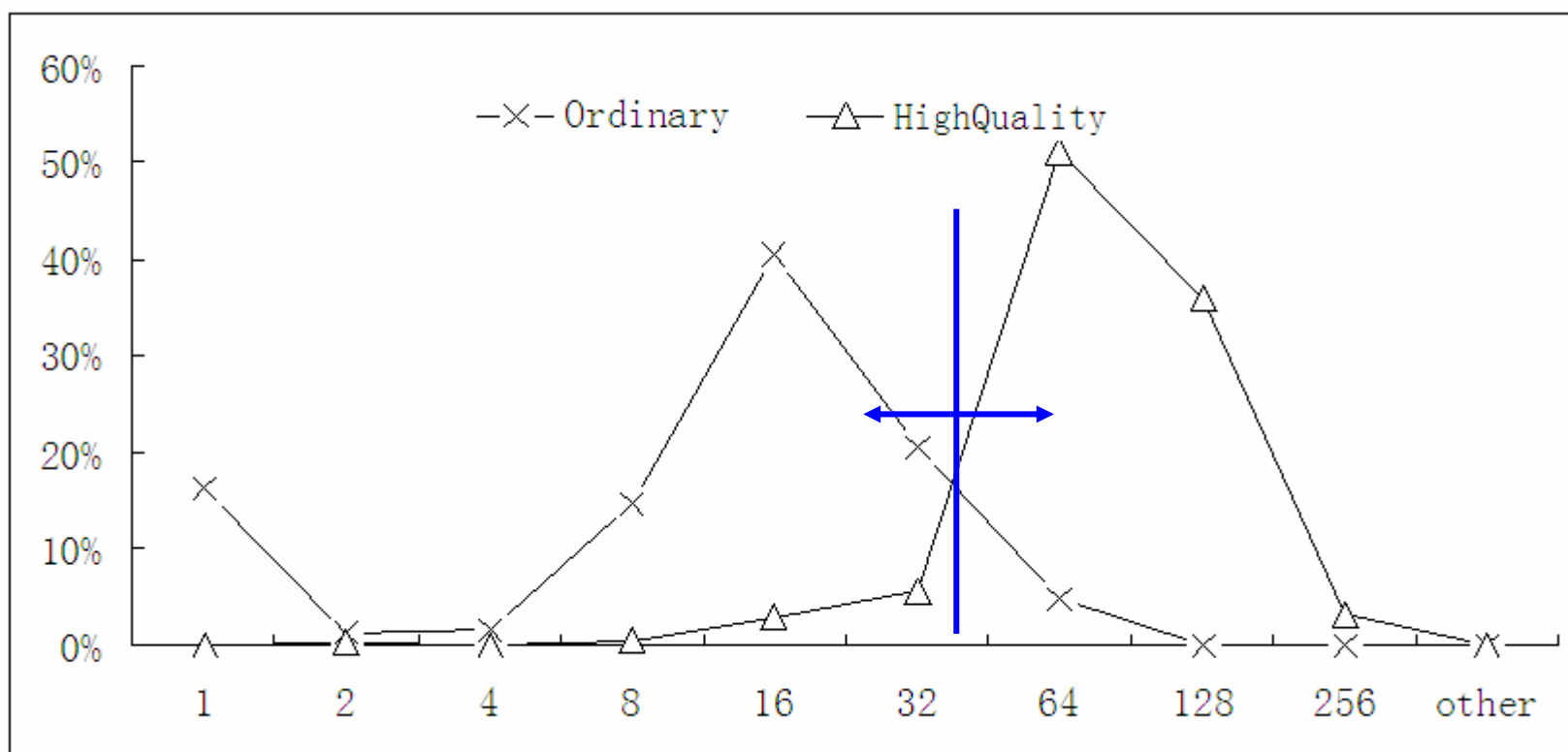
高质量页面的查询无关特征分析

- 基于真实网络语料库进行查询目标页面的查询无关特征分析
 - 语料库：采集于2005年11月的超过3700万中文网页
 - 高质量页面采样：训练集规模超过1600个网页
 - 基于超链接结构分析的特征
 - PageRank值，出、入链接个数，出、入链接文本长度，内容比例等
 - 其他类型的查询无关特征
 - 文档长度/大小/包含多媒体数据数目，页面镜像数目，URL长度与类型，页面编码等



高质量页面的查询无关特征分析

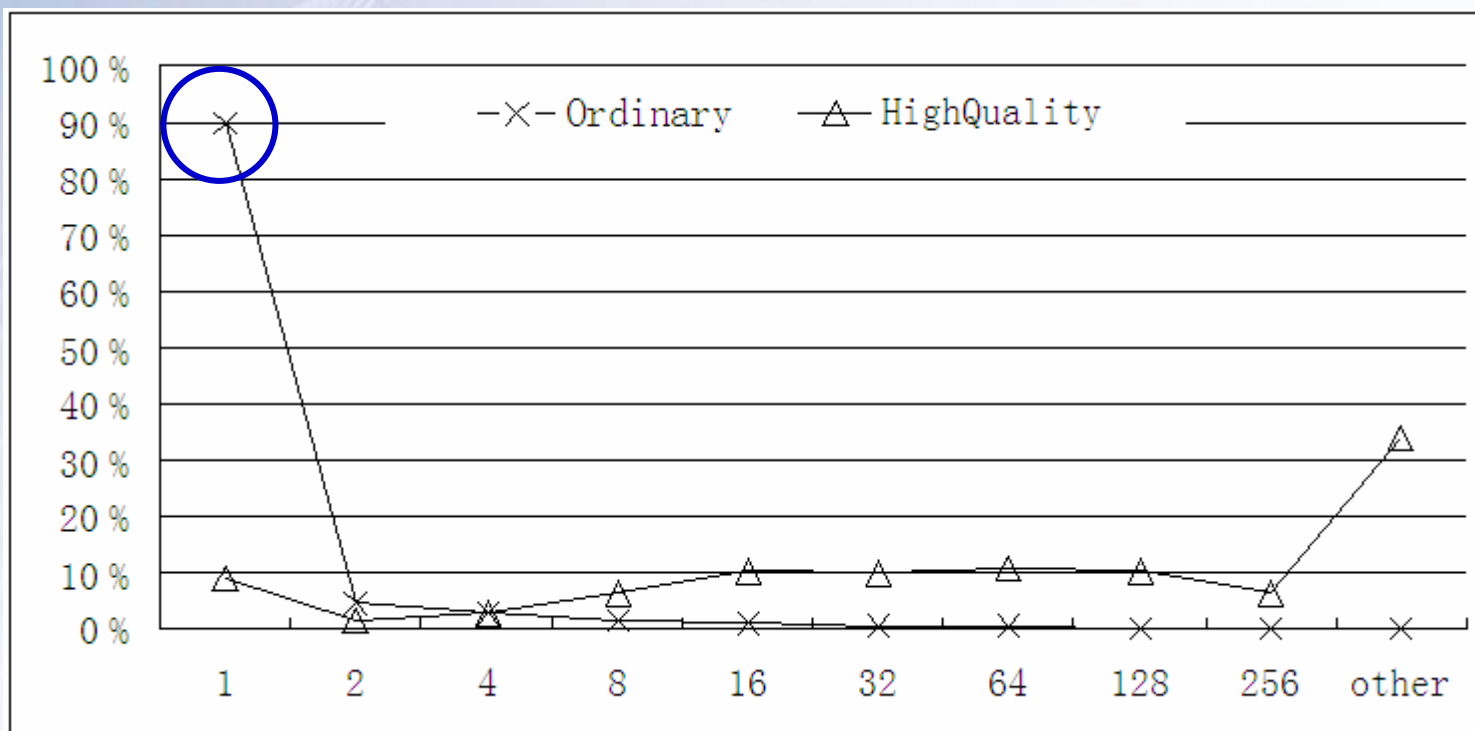
- PageRank





高质量页面的查询无关特征分析

- 入链接个数





高质量页面的查询无关特征分析

- 其他部分特征

	普通页面	查询目标页面
动态页面	13.06%	1.87%
编码为GBK	14.04%	1.39%
Hub类型	3.78%	24.77%

- 查询无关特征能够有效地区分目标页面与普通页面，亦即查询目标页面具有查询无关特征
- 利用这种特征进行页面质量评估是可行的



基于学习的页面质量评估算法

- 基于机器学习方法进行质量评估
 - 具有完备的理论基础和实践经验
 - 高质量页面与低质量页面有特征差别
 - 特征数量不多，且相对独立（信息来源不同）
- 主要问题
 - 缺乏有代表性的反例样例集合
 - 对Web页面进行均一采样是国际性的难题
- 利用基于正例的学习方法进行质量评估



已有工作

- 基于正例的学习方法
 - 证明利用正例样本和未标注样本进行粗分类的任务是PAC可学习的 (Denis, 1998)
需要对正例页面的比例进行预估
 - PEBL: 基于正例的学习框架 (Yu et.al., 2004)
迭代算法, 效率低下
- 我们的解决方案
 - 基于Bayes学习的方法: 估计页面成为用户需求页面 (某个查询的目标页面) 的概率



基于学习的页面质量评估算法

- 根据查询无关特征计算页面满足用户群体需求的可能性，用这种可能性表示页面质量的高低
- 形式化的表述为：

具有查询无关特征 $A_1, A_2, A_3, \dots, A_n$ 的页面 P 成为查询目标页面的可能性

$$P(p \in \text{Target page} \mid p \text{ has feature } A_1 A_2, \dots, A_n)$$



基于学习的页面质量评估算法

- 算法描述

[1] 单特征分析

$$\begin{aligned} & P(p \in \text{Target page} \mid p \text{ has feature } A) \\ &= \frac{P(p \text{ has feature } A \mid p \in \text{Target page})}{P(p \text{ has feature } A)} \times \boxed{P(p \in \text{Target page})} \end{aligned} \quad (\text{贝叶斯公式})$$

$$\begin{aligned} & \frac{P(p \text{ has feature } A \mid p \in \text{Target page})}{P(p \text{ has feature } A)} \\ &= \frac{\#(p \text{ has feature } A \cap p \in \text{Target page})}{\#(\text{Target page})} \bigg/ \frac{\#(p \text{ has feature } A)}{\#(\text{CORPUS})} \end{aligned} \quad (\text{先验概率定义})$$



$$\begin{aligned} & P(p \in \text{Target page} \mid p \text{ has feature } A) \\ & \propto \frac{\#(p \text{ has feature } A \cap p \in \text{Target page sample set})}{\#(\text{Target page sample set})} \bigg/ \frac{\#(p \text{ has feature } A)}{\#(\text{CORPUS})} \end{aligned}$$



基于学习的页面质量评估算法

- 算法描述

- [2] 多特征分析

在合理选取的基础上，可以发现特征之间的近似独立性关系

	URL Format	Encode	PageRank	Cluster	DocLength	URL Length	Indegree
URL Format	1.00	0.05	0.05	0.01	0.04	0.10	0.00
Encode		1.00	0.20	0.00	0.06	0.30	0.00
PageRank			1.00	0.01	0.06	0.03	0.05
Cluster				1.00	0.01	0.10	0.00
DocLength					1.00	0.04	0.00
URL Length						1.00	0.02
Indegree							1.00



基于学习的页面质量评估算法

- 算法描述

[2] 多特征分析 (续)

$$P(p \text{ has feature } A_1, A_2, \dots, A_n \mid p \in \text{Target page}) = \prod_{i=1}^n P(p \text{ has feature } A_i \mid p \in \text{Target page})$$

(朴素贝叶斯假设)

$$P(p \text{ has feature } A_1, A_2, \dots, A_n) = \prod_{i=1}^n P(p \text{ has feature } A_i)$$

(特征近似独立)

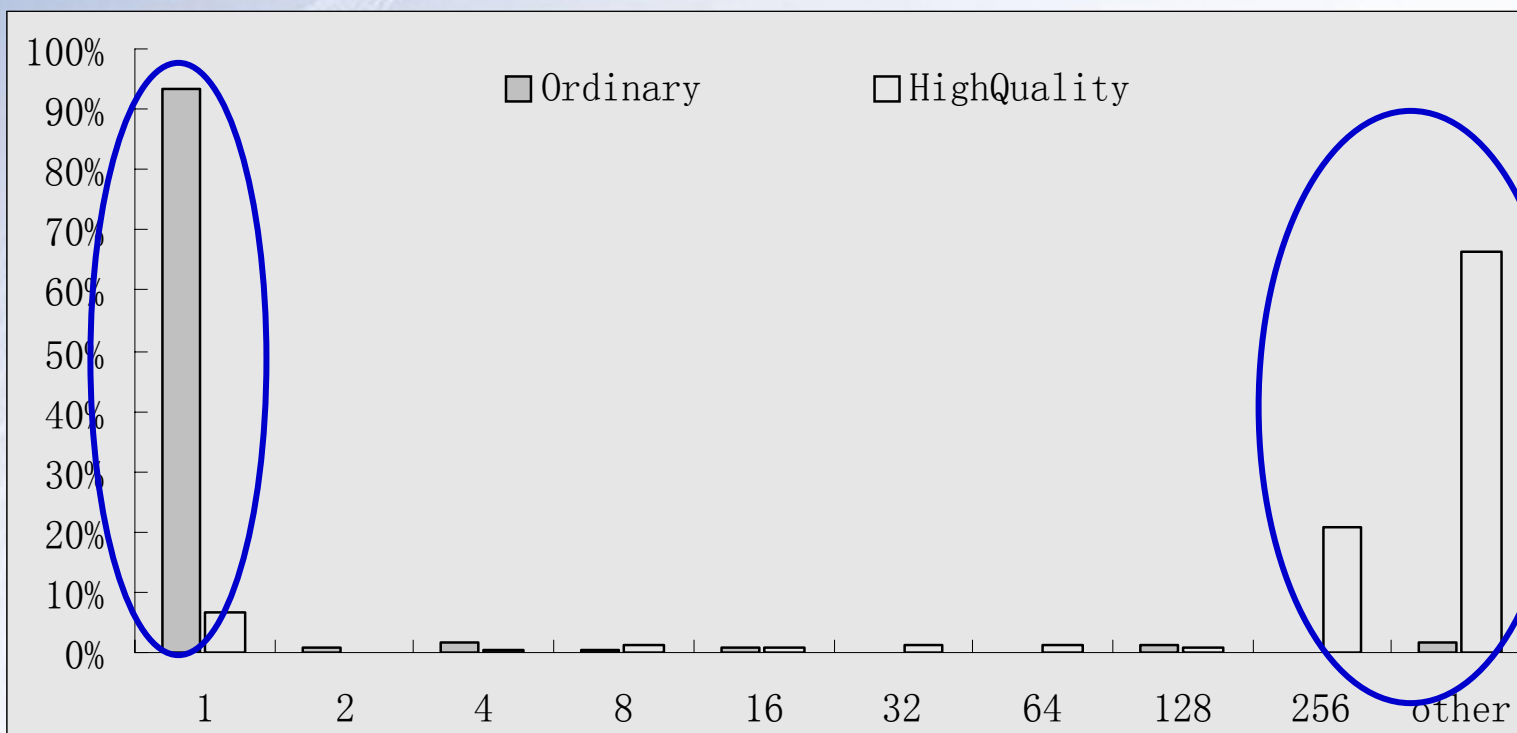


$$\begin{aligned} & P(p \in \text{Target page} \mid p \text{ has feature } A_1, A_2, \dots, A_n) \\ &= \frac{P(p \text{ has feature } A_1, A_2, \dots, A_n \mid p \in \text{Target page})P(p \in \text{Target page})}{P(p \text{ has feature } A_1, A_2, \dots, A_n)} \\ &\propto \prod_{i=1}^n \frac{P(p \text{ has feature } A_i \mid p \in \text{Target page})P(p \in \text{Target page})}{P(p \text{ has feature } A_i)} = \prod_{i=1}^n P(p \text{ has feature } A_i \mid p \in \text{Target page}) \end{aligned}$$



基于学习的页面质量评估算法

- 比较 $P(p \in \text{Target page} \mid p \text{ has feature } A_1 A_2, \dots, A_n)$ 的相对大小





基于学习的页面质量评估算法

- 测试效果

- 测试集合：人工标注，数目超过17000个（训练集规模的10倍）

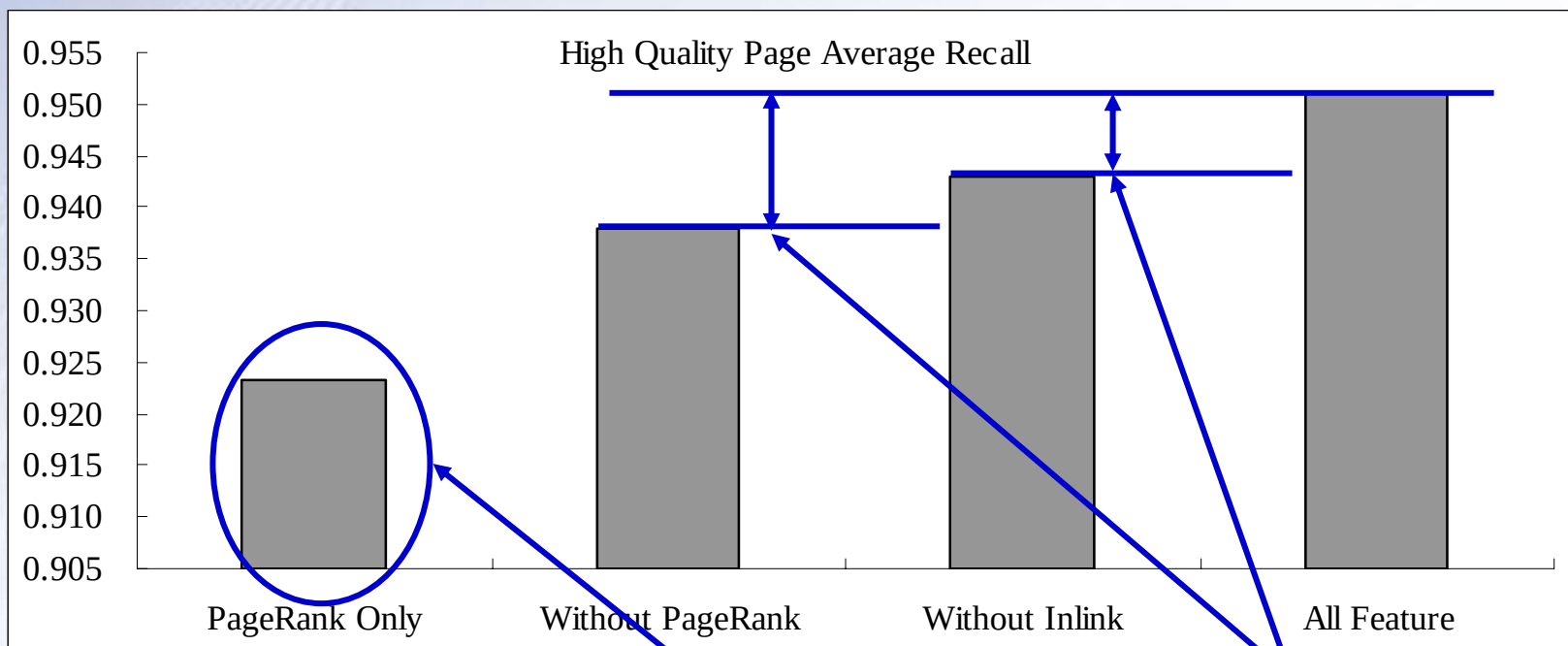
	普通页面	查询目标页面 训练集合	查询目标页面 测试集合
算法判定出的 低质量页面	95.04%	7.27%	7.63%
算法判定出的 高质量页面	4.96%	92.73%	92.37%

- 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



基于学习的页面质量评估算法

- 与传统的基于链接分析的评估方法比较
 - 使用“高质量页面平均召回率”(AR)作为评价指标



比仅使用PageRank
特征取得更好的效果

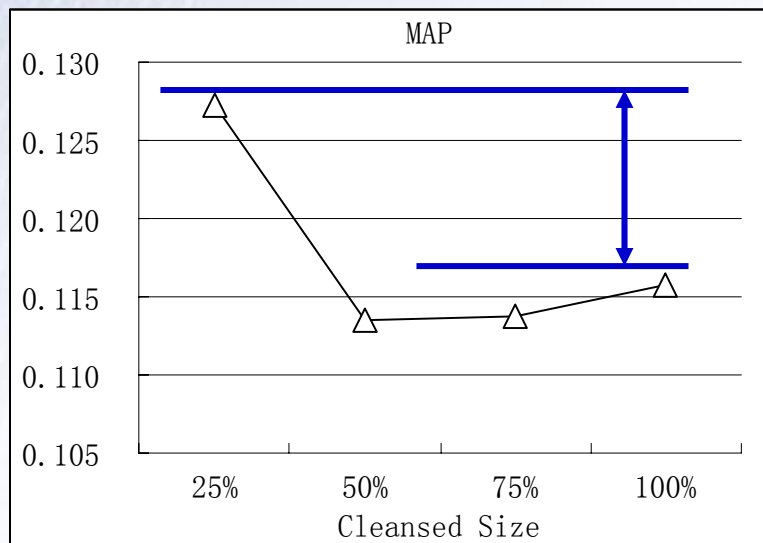
并不单独依靠某个特征
实现评估任务



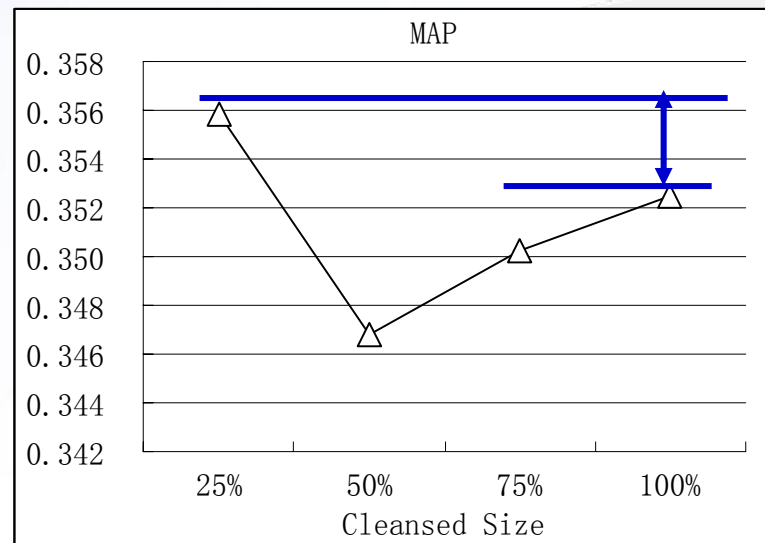
基于学习的页面质量评估算法

- 对检索效果的影响

- 在TREC 2002 – 2004的实验平台进行试验
- 针对高质量页面集合进行检索，内容相关的低质量结果被去除，因而有利于检索效果提高



TREC 2003

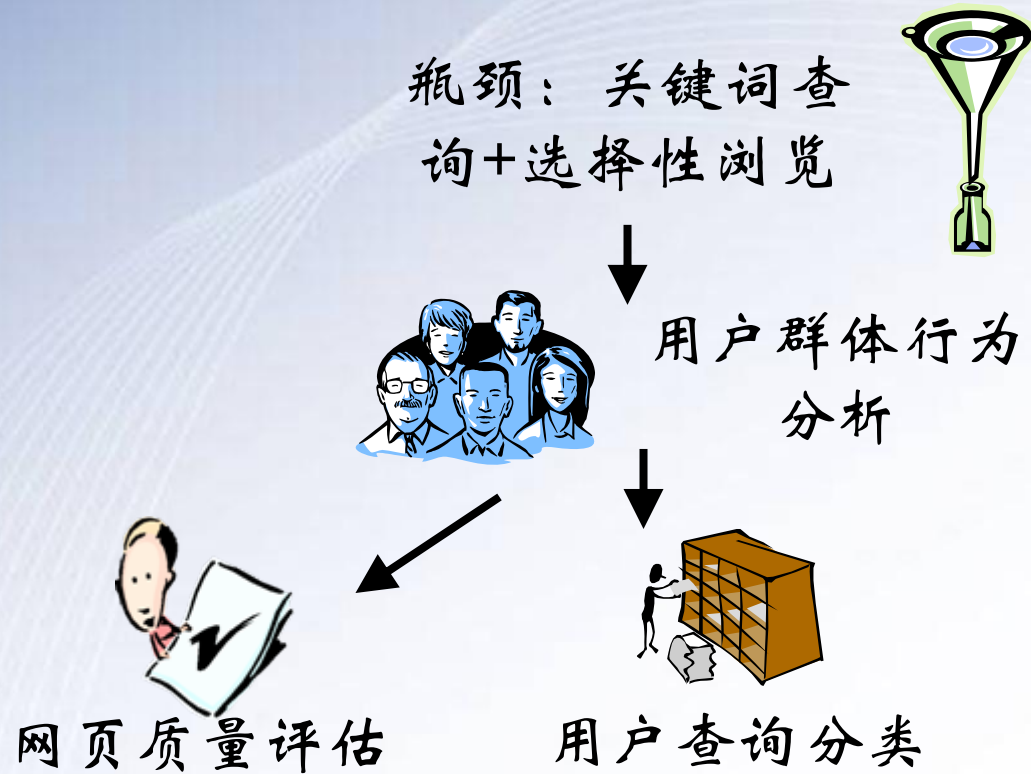


TREC 2004主题过滤任务



Information Systems © Tsinghua University

提纲





问题背景

- 用户查询需求分类

- 目的：依照信息需求的不同对查询进行不同处理

- 用户查询分类体系

- 由Broder(IBM)和Rose(Yahoo!)各自提出。

- 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





用户查询需求分类

- 根据用户信息需求的不同对查询进行分类
- 用户查询分类已有的工作
 - 利用查询自身特征和检索系统反馈特征
 - TREC 2004的相关工作，准确率较低，仅61%
 - 利用查询与网页链接文本的匹配程度
 - (Kang et al, 2004)基于此分类方法提高了检索系统性能
 - 利用查询所具有的历史信息
 - (Lee et al, 2005) 在UCLA校园网内进行的小规模实验，取得了90%以上的查询分类效果



相关工作

- 已有工作的问题

- 缺乏对分类方法所依据特征的全面考察

- 可用性：能否对大部分用户查询获取此类特征
 - 易用性：获取特征的资源消耗如何
 - 分类效果：特征分辨不同类型查询的能力如何





针对查询分类特征的考察

• 特征考察的对象

— 查询自身的特征（查询长度、查询词性）

- 可用性、易用性好

重点考察分类能力

— 与链接文本的匹配度特征

重点考察可用性

- 分类能力高：匹配程度高 => 导航类检索可能性大

— 查询所具有的历史行为特征

- 分类能力高：用户点击集中 => 导航类检索可能性大

— 检索系统反馈的特征

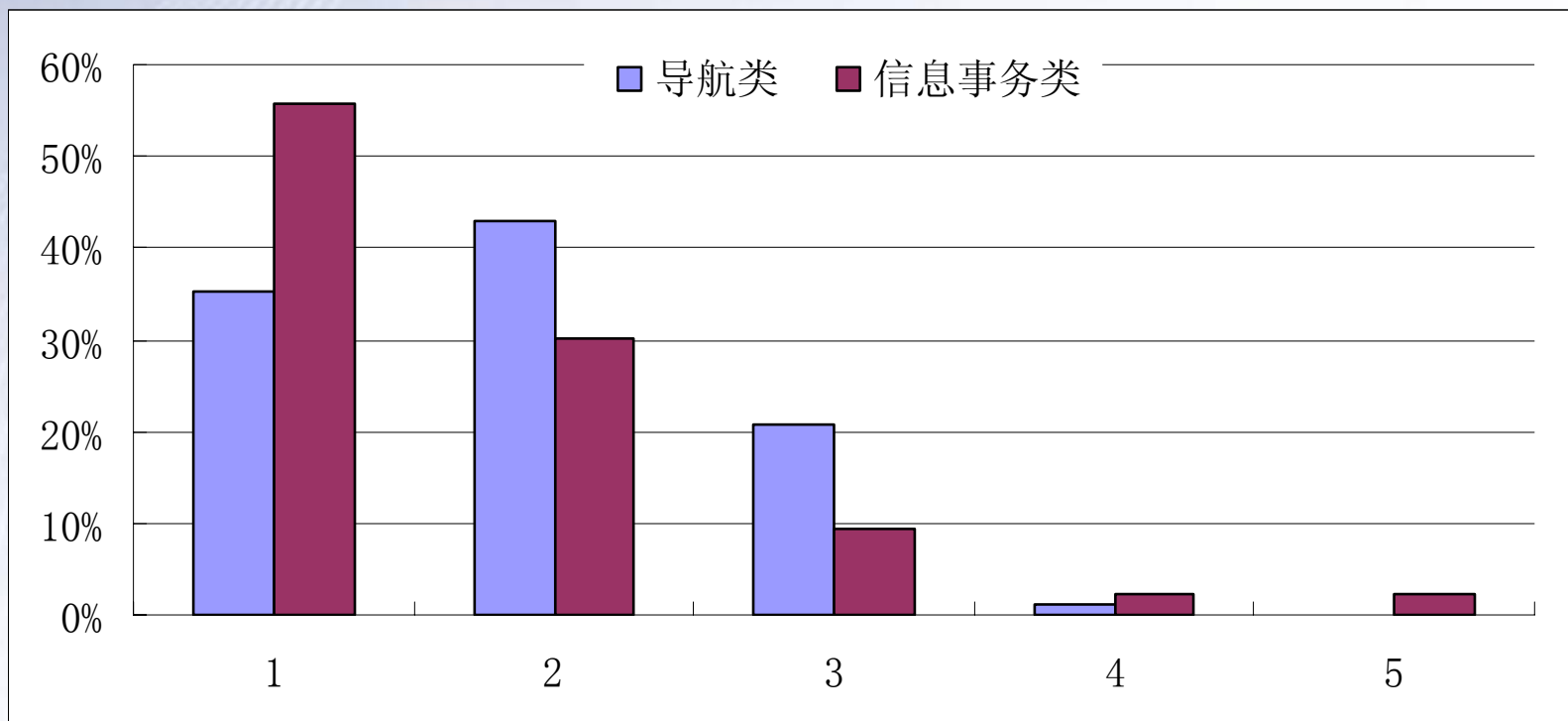
- 易用性差，分类能力一般

不作为考察重点



针对查询分类特征的考察

- 查询自身特征（分类能力考察）
 - 查询长度特征的分类能力较差

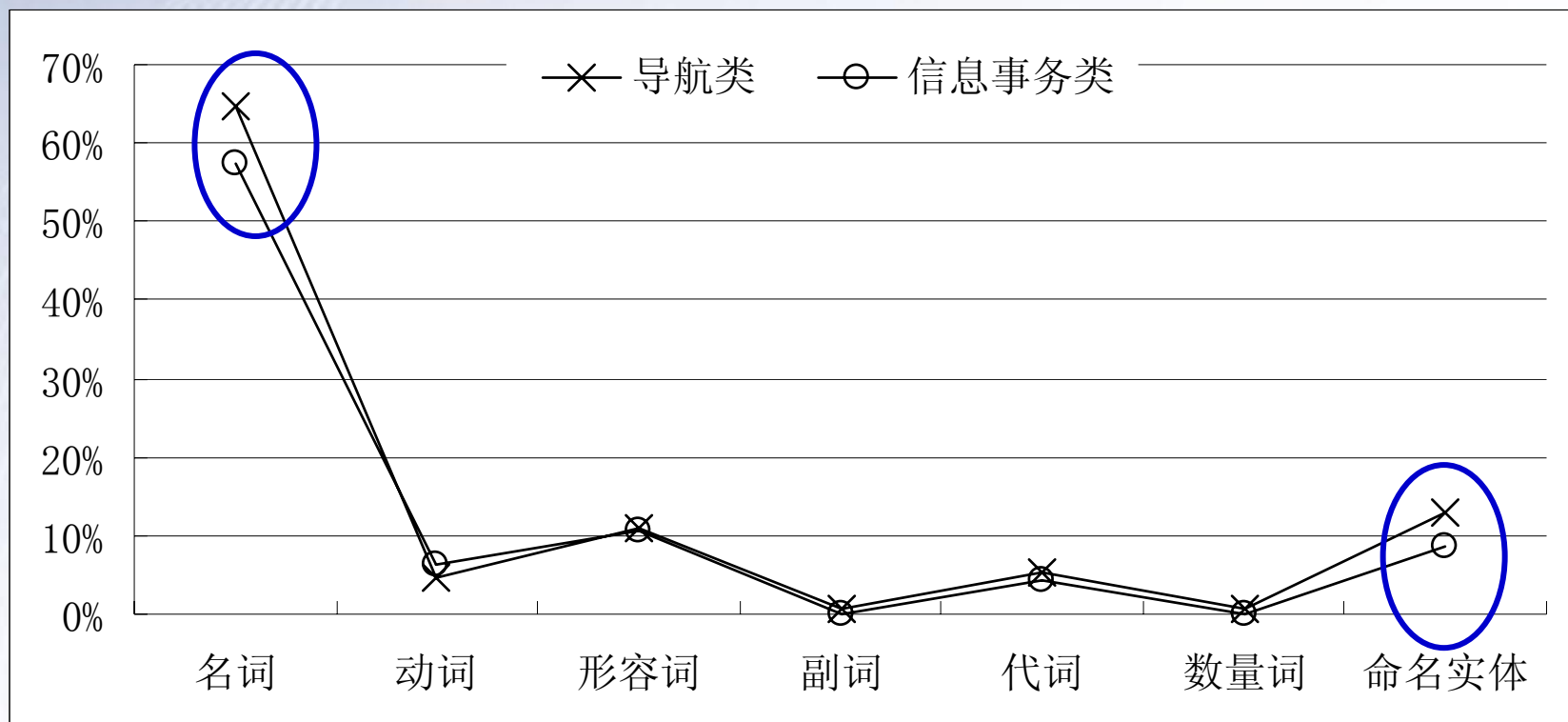




针对查询分类特征的考察

- 查询自身特征

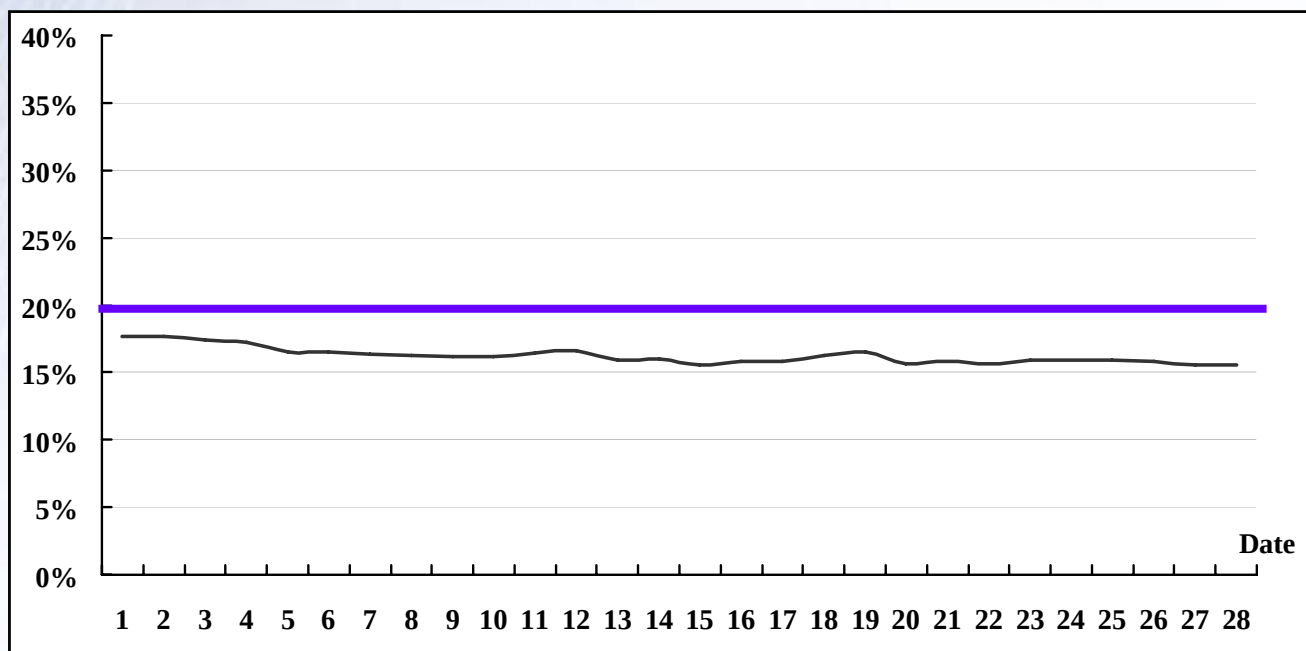
- 词性特征的分类能力较差





针对查询分类特征的考察

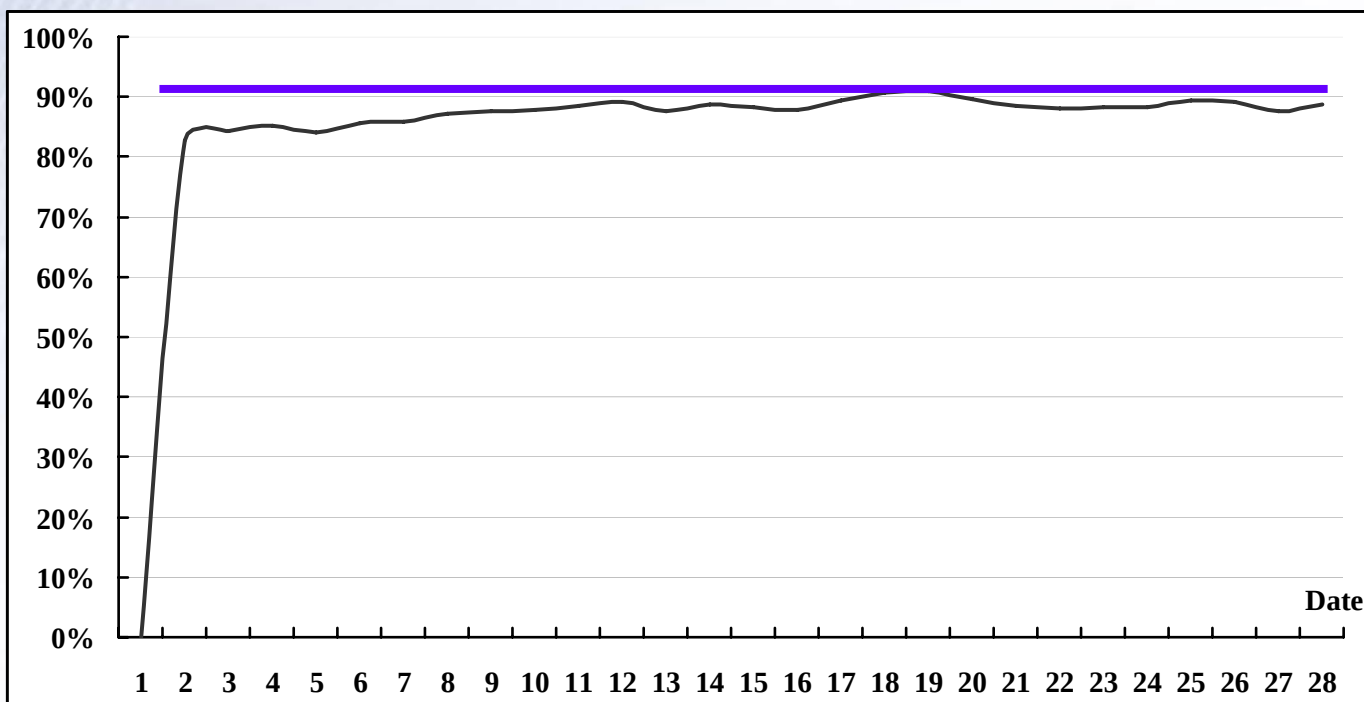
- 链接文本匹配特征（可用性考察）
 - 链接文本匹配特征可用的前提
 - 有多大比例的查询与某个页面的链接文本匹配





针对查询分类特征的考察

- 查询所具有的历史行为特征(可用性考察)
 - 查询所具有的历史行为特征可用的前提
 - 有多大比例的查询曾经在历史日志记录中出现过





Information Systems @ Tsinghua University

查询分类算法

- 针对查询历史行为信息的特征提取
 - Lee (2005) 提出了点击分布的概念

$$CD(\text{Query } q) = \frac{\#(\text{Session of } q \text{ that involves clicks on the most frequently clicked results})}{\#(\text{Session of } q)}$$

- 某个查询的用户群体点击集中在某个结果 => 这个查询有可能是导航类查询
- 问题
 - 缺乏对导航类查询特征的细致挖掘
 - 用户行为：点击顺序、点击结果的排名等



查询分类算法

- 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）

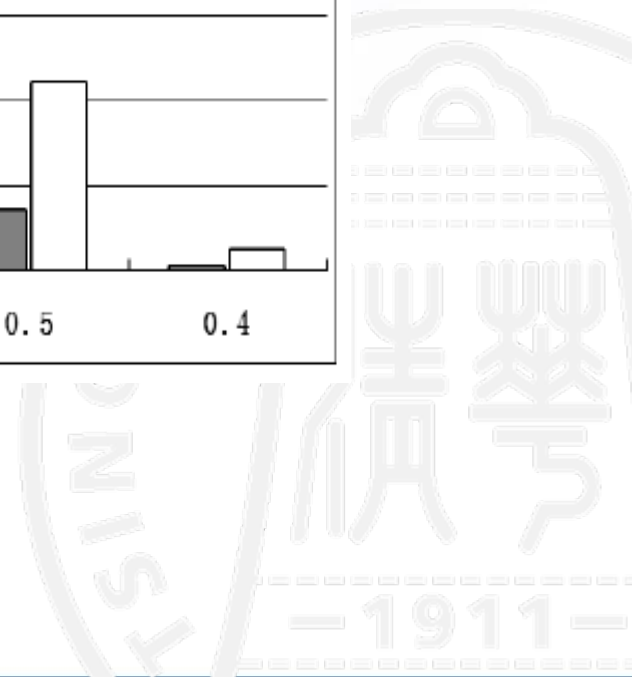
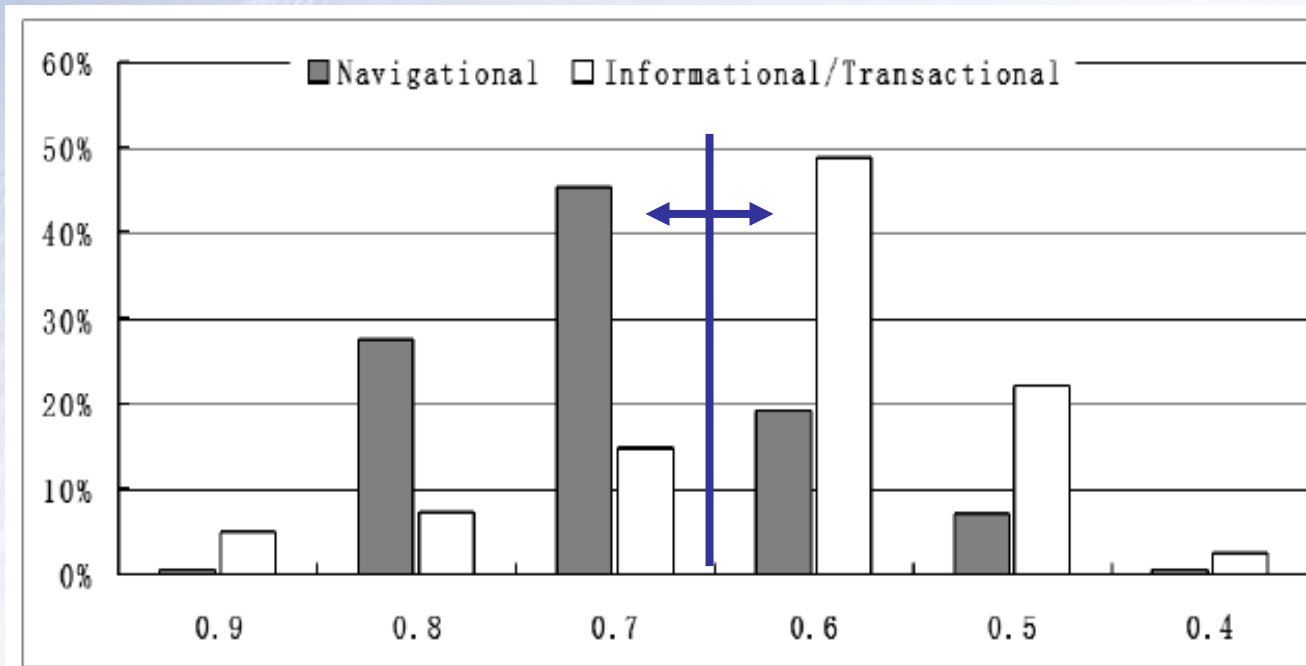
$$nCS(\text{Query } q) = \frac{\#(\text{Session of } q \text{ that involves less than } n \text{ clicks})}{\#(\text{Session of } q)}.$$



Information Engineering @ Tsinghua University

查询分类算法

- nCS 的分布情况

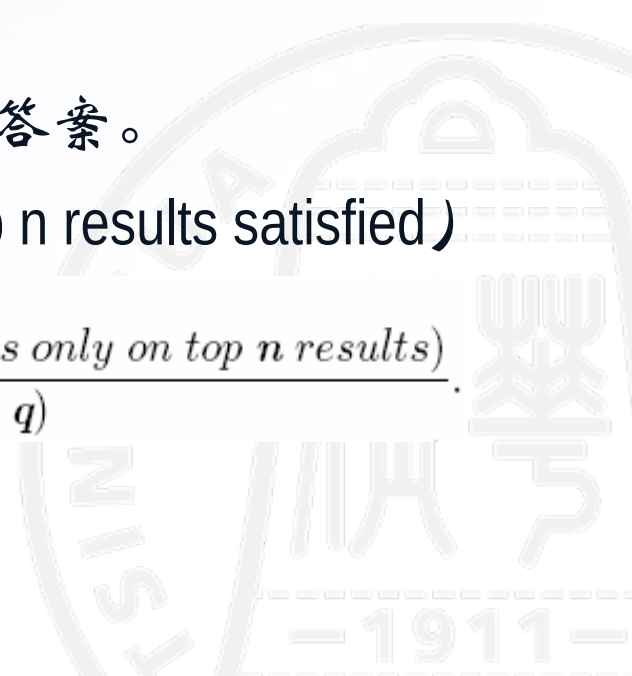




查询分类算法

- 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）

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

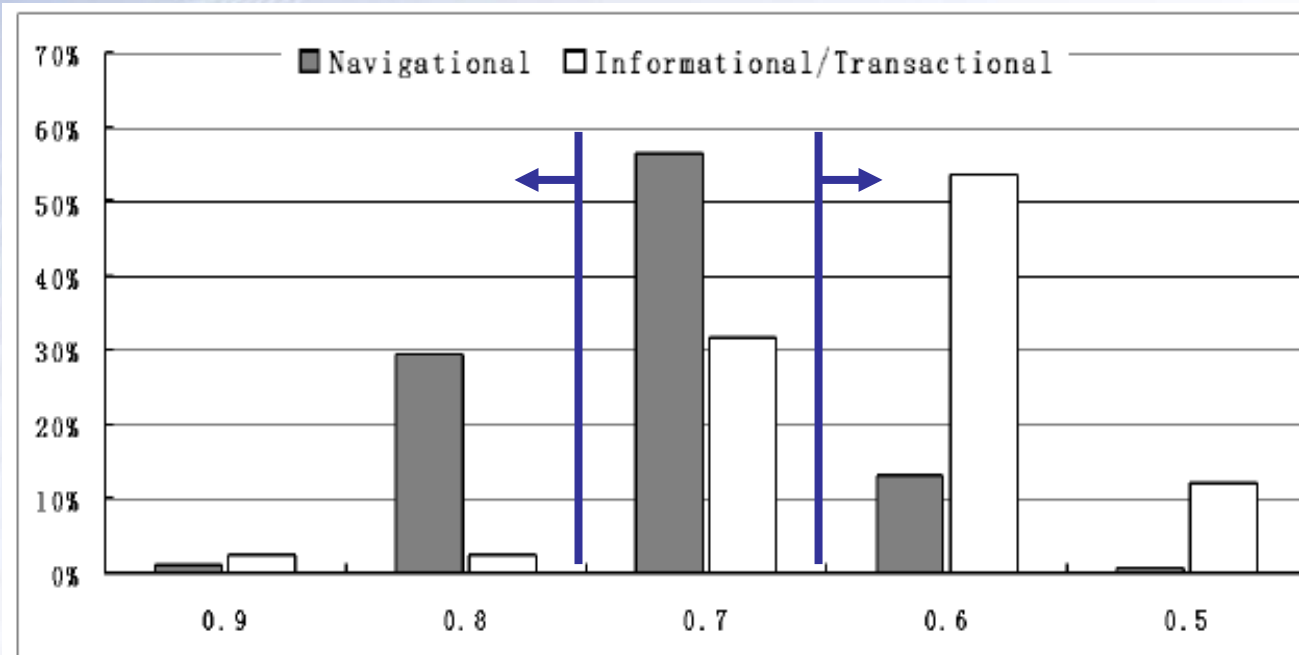




Information Systems @ Tsinghua University

查询分类算法

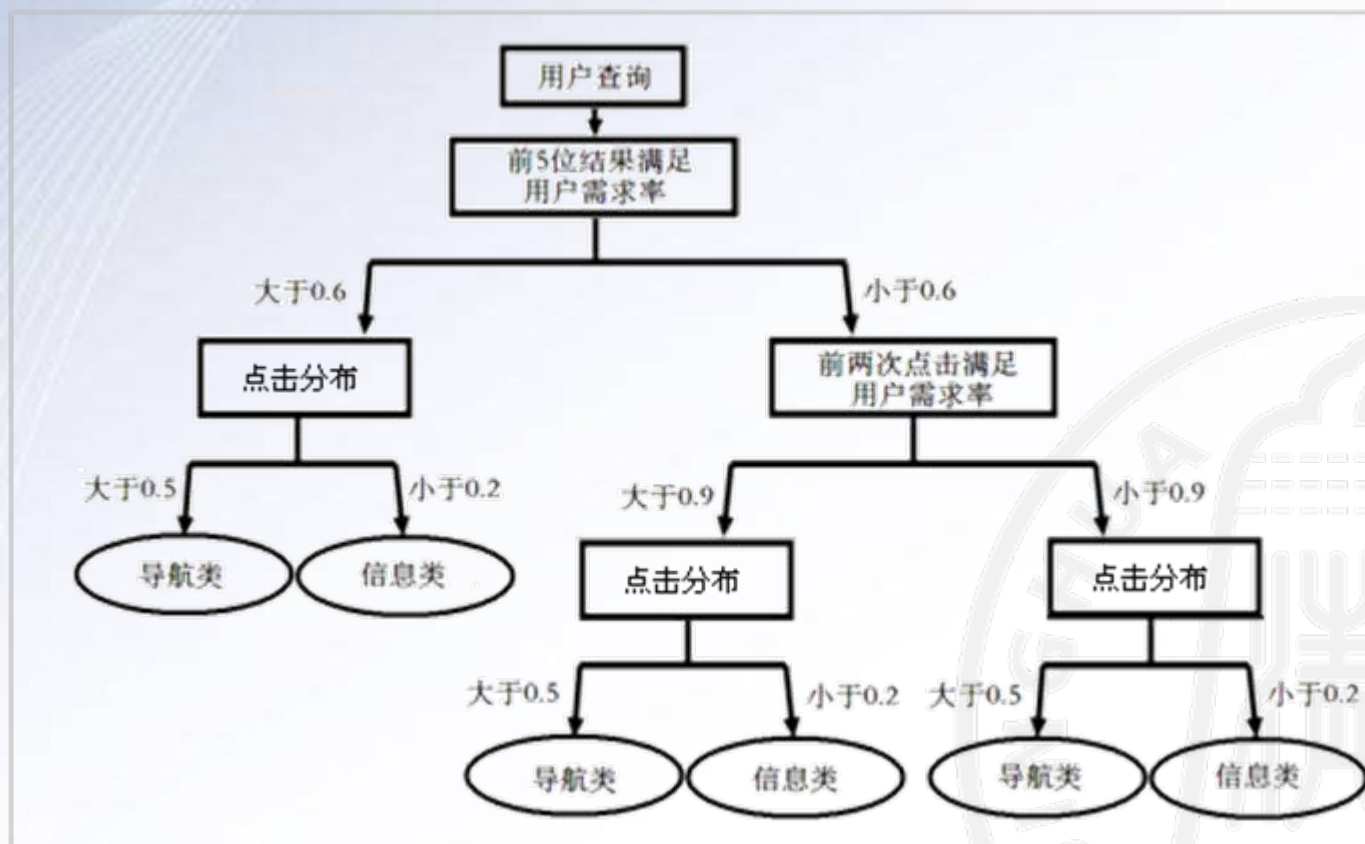
- nRS 的分布情况



清华大学
—1911—

查询分类算法

- 基于决策树学习的分类算法





实验设定

- 实验分析所使用的数据
 - Sogou 2006年2月全月查询和点击日志数据
 - 共86538613条点击，涉及26255952个用户session
- 训练集设定
 - 人工标注的高频度查询
 - 导航类型：153个；信息事务类型：45个
- 测试集设定
 - 信息事务类：天网搜索引擎评测，81个
 - 导航类：Hao123网站，152个



实验结果

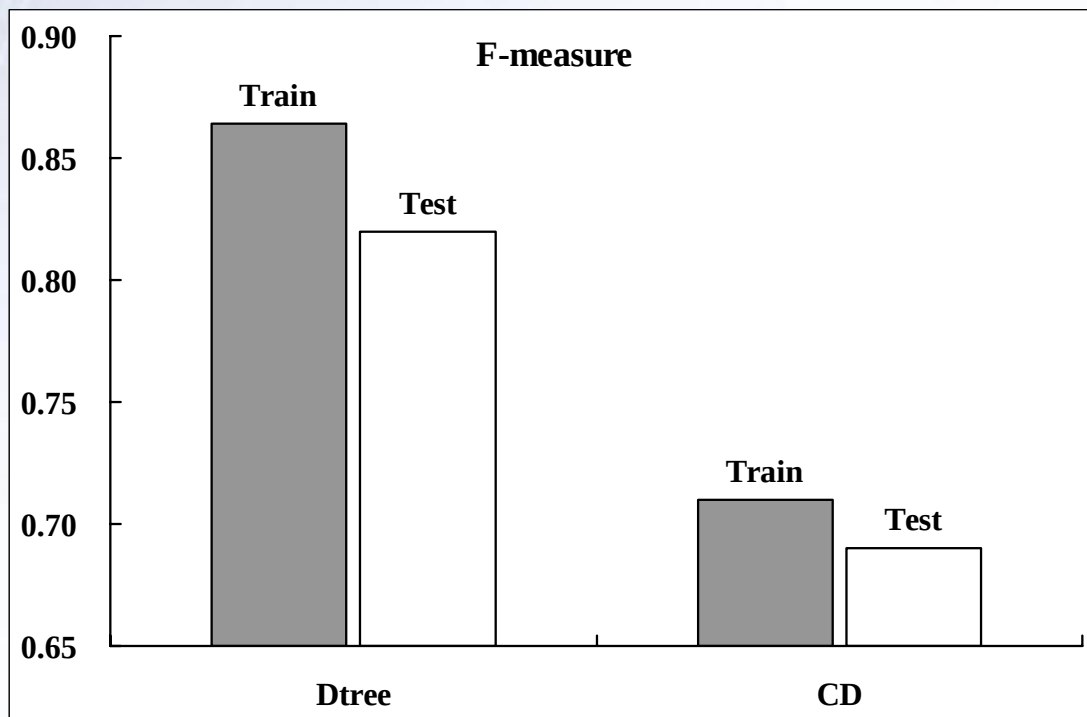
• 实验结果

	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

- 测试集和训练集上的分类准确率都在80%以上
- 基于用户群体行为分析方法，利用对历史行为特征的挖掘，可以实现用户查询对应的信息需求类别的划分。

- 实验结果

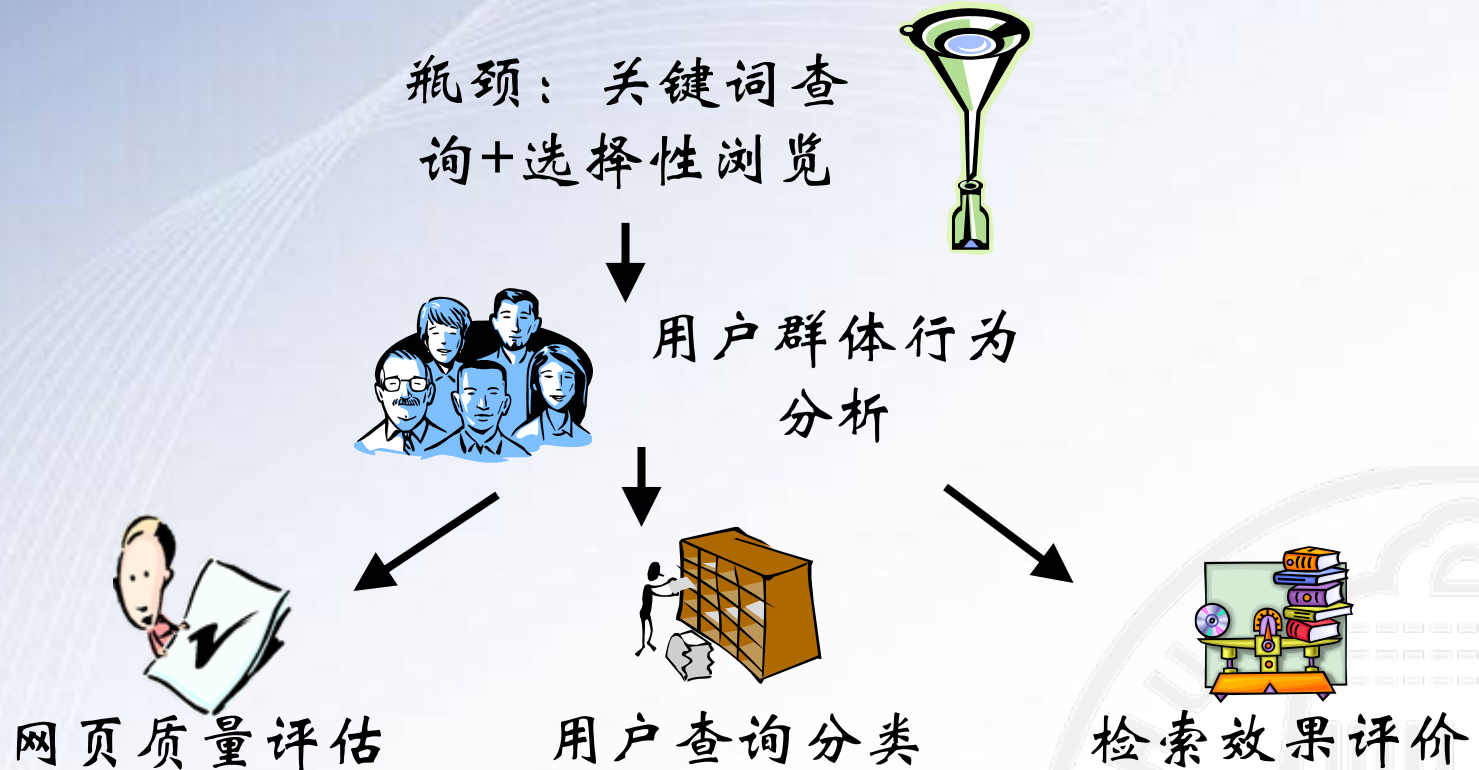
- 决策树学习算法的分类效果好于基于点击分布的分类方法(测试集提高 30%，训练集提高19%)



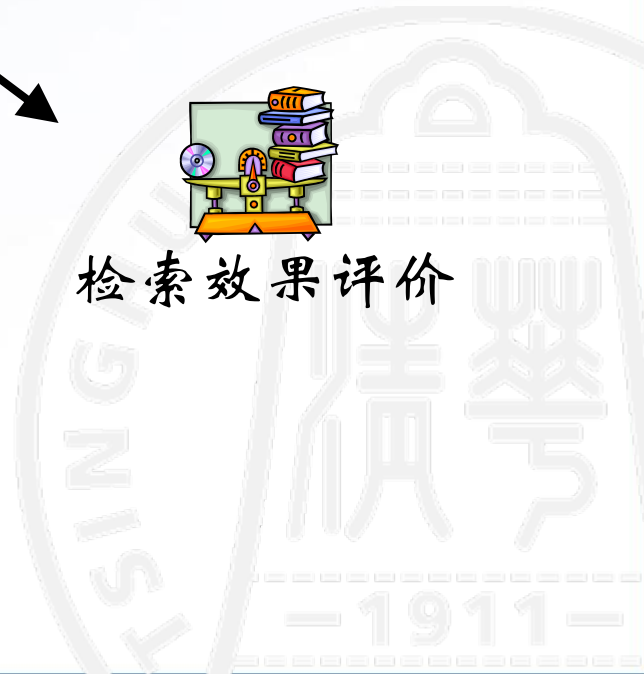


Information Systems @ Tsinghua University

提纲



检索效果评价





Information Systems © Tsinghua University

研究背景

- 效果评价是信息检索相关研究的基础内容
 - 评价在信息检索系统的研发中一直处于核心的地位，以致于算法与它们的效果评价方式是合二为一的。（Saracevic, 1995）
 - Cranfield评价方法
 - 被应用在包括TREC在内的几乎所有主流的信息检索研究工作中。
 - 组成：查询样例集、标准答案集
语料库





搜索引擎效果自动评价

- 网络信息检索系统效果评价的现状

- 数据对象繁杂，标注困难

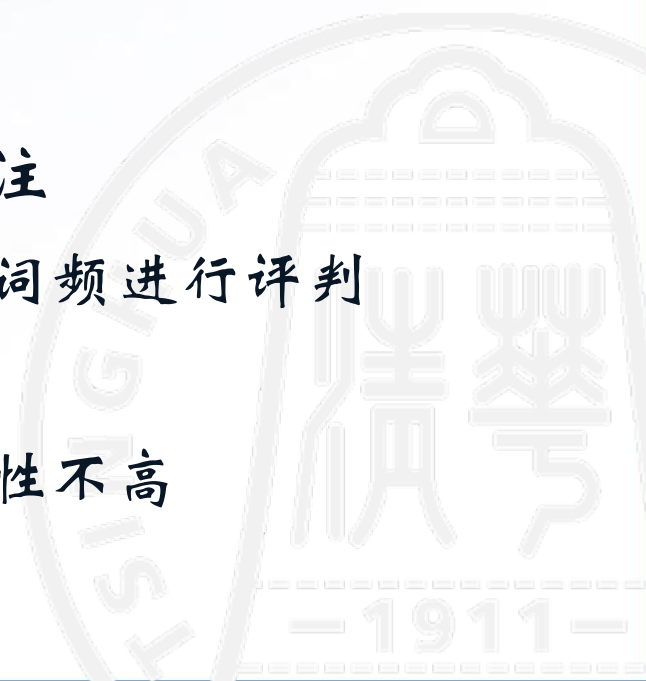
- Voorhees 估计，对一个规模为800万的文档集合进行针对1个查询主题的相关性评判需要耗费1名标注人员9个月的工作时间
 - TREC提出pooling方法，在保证评价结果可靠性的基础上大大减少了评判工作量
 - 缺点：处理的查询数目少，针对小规模查询集合，仍需要耗费十余名标注人员1-2个月的工作时间

- 提高搜索引擎评价方法的自动化程度势在必行



相关工作

- 核心问题: 如何提高结果标注的自动化程度
 - 利用伪相关反馈的方法进行结果自动标注
 - 伪相关反馈的结果被认为是答案
 - Soboroff et.al, 2001; Nuray et.al, 2003; Beitzel et.al., 2003
 - 标注的准确程度和可靠性不高
 - 对查询词的近义词集合进行标注
 - 根据结果网页的查询词及近义词词频进行评判
 - Amitay et.al, 2004
 - 初始标注工作量大, 评判的可靠性不高





相关工作

- 核心问题: 如何提高结果标注的自动化程度
 - 利用用户行为分析方法
 - Joachims et.al.(1997) 创建了一个元搜索接口, 将不同搜索引擎的结果加以混杂, 把用户点击不同结果的情况加以记录, 并以此作为评判依据
 - 用户行为分析的思路值得肯定
 - 缺点: 改变用户检索习惯、个体用户行为不可靠等
 - 我们的研究思路
 - 通过对搜索引擎记录下的用户行为日志进行分析, 不额外增加用户负担。
 - 使用用户群体, 而不是个体的点击行为作为依据



搜索引擎效果自动评价

- 对查询集合结果进行标注
 - 个体用户可能在点击过程中出现谬误
 - 用户群体的点击信息则是很大程度上可靠的
 - 统计用户群体的点击行为，进而对查询对应的结果进行标注
- 结果标注的可靠性分析
 - 结果标注是否准确
 - 结果标注是否全面





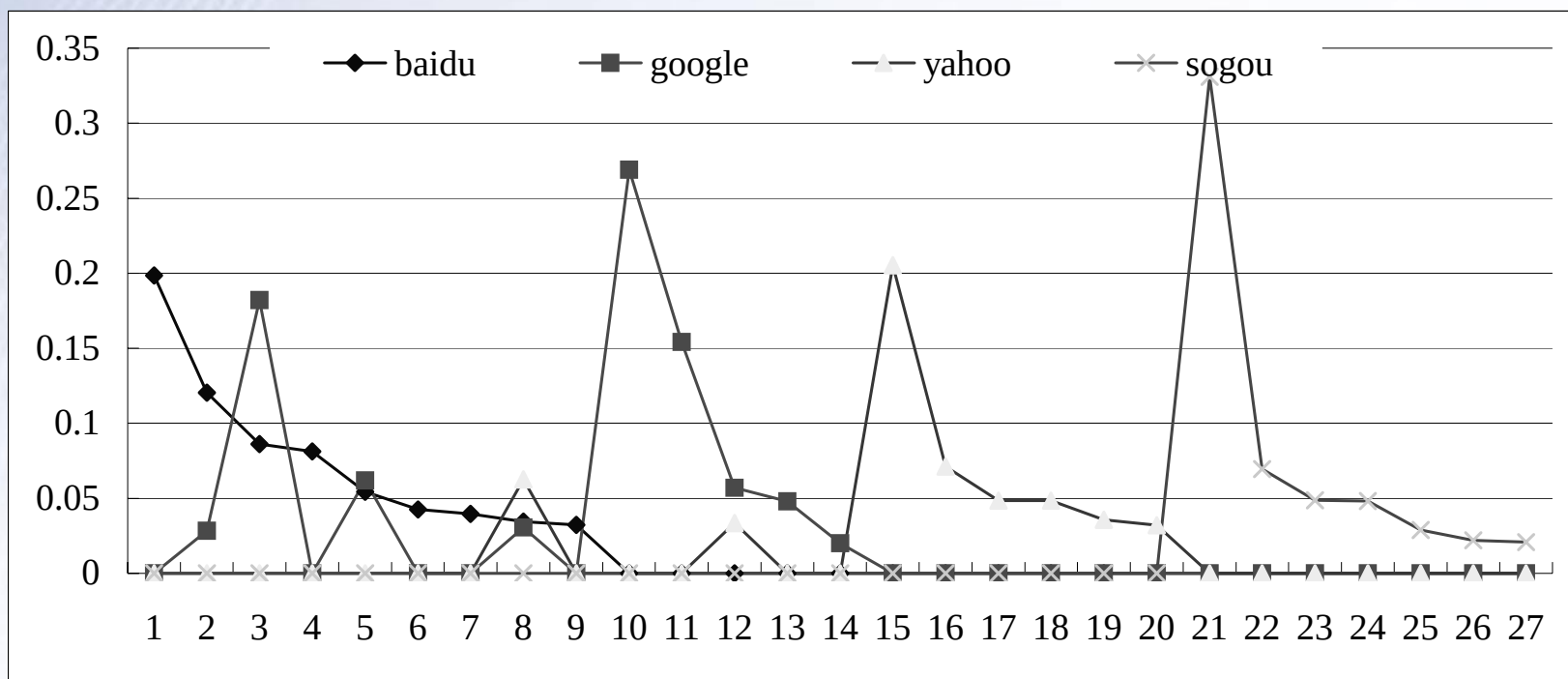
结果自动标注

- 用户行为的可靠性分析
 - 对导航类查询
 - 结果唯一，不存在“找不全”的问题
 - 搜索引擎检索性能高，点击准确度较可靠
 - 能够发现用户的关注热点，如
 - “163” => mail.163.com
 - “搜狗” => d.sogou.com
 - 适合作为搜索性能自动评测的对象



结果自动标注

- 用户行为的可靠性分析
 - 对信息事务类查询（以查询词“电影”为例）
 - 不同搜索引擎的点击分布差异大





Information Systems © Tsinghua University

结果自动标注

- 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

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

- q 的点击集中度最高的 r 即为其检索目标页面



Information Systems © Tsinghua University

结果自动标注

- 针对导航类查询的结果自动标注
 - 标注算法：寻找针对 q 的ClickFocus值最大的 r
 - 标注样例

查询	自动标注的URL
读写网	http://www.duxie.net/
北京现代	http://www.beijing-hyundai.com.cn/
NBA	http://china.nba.com/
南方都市报	http:// www.nanfangdaily.com.cn/
卓越网	http:// www.joyo.com/



结果自动标注实验

- 实验数据

- Sogou搜索8个月查询日志(超过7亿条日志信息)
- 针对其中高频导航类查询进行自动标注
- 对每个时间段抽取约5%的数据进行手工检查

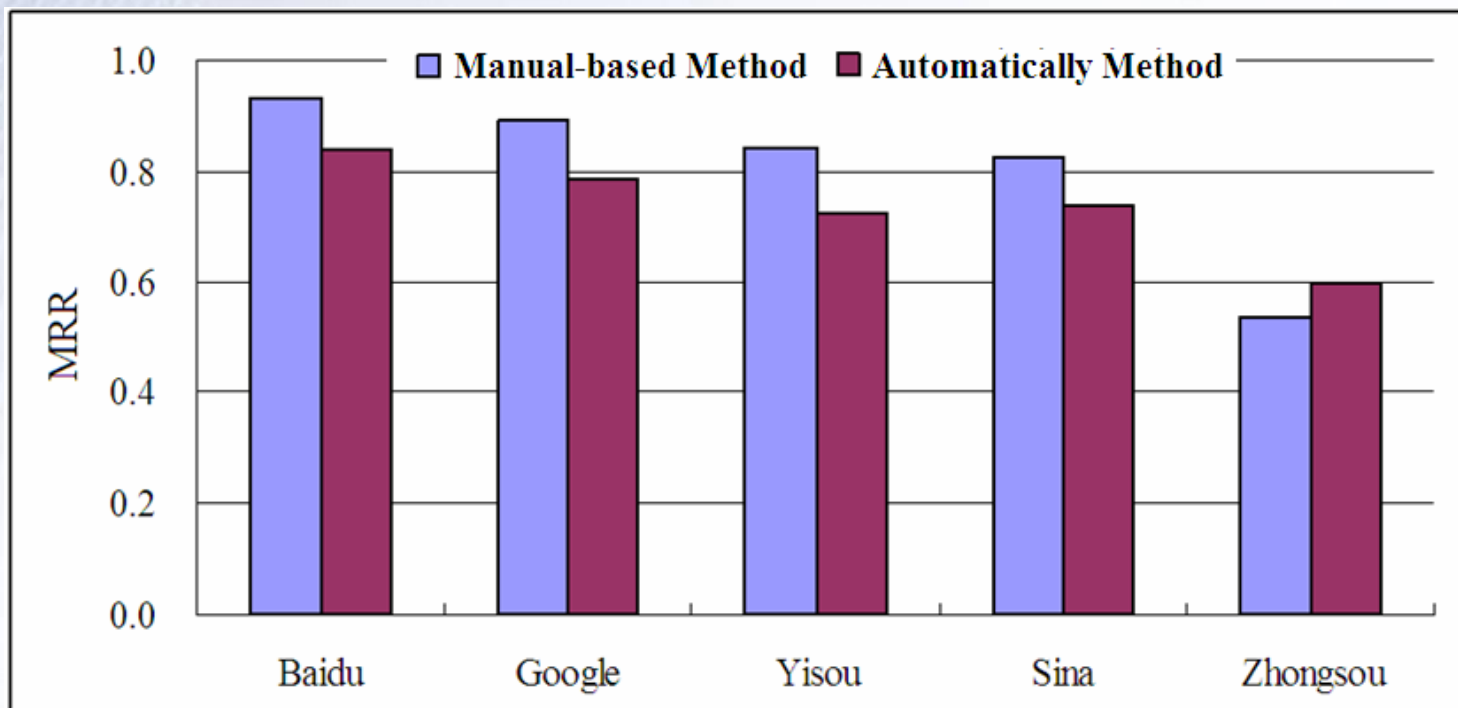
时间段	标注查询数	检验样例数	样例集合 准确度
Jun. 06 - Aug. 06	13,902	695	98.13%
Sept.06 - Nov. 06	13,884	694	97.41%
Dec. 06 - Jan. 07	11,296	565	96.64%



Information Systems © Tsinghua University

自动性能评价实验

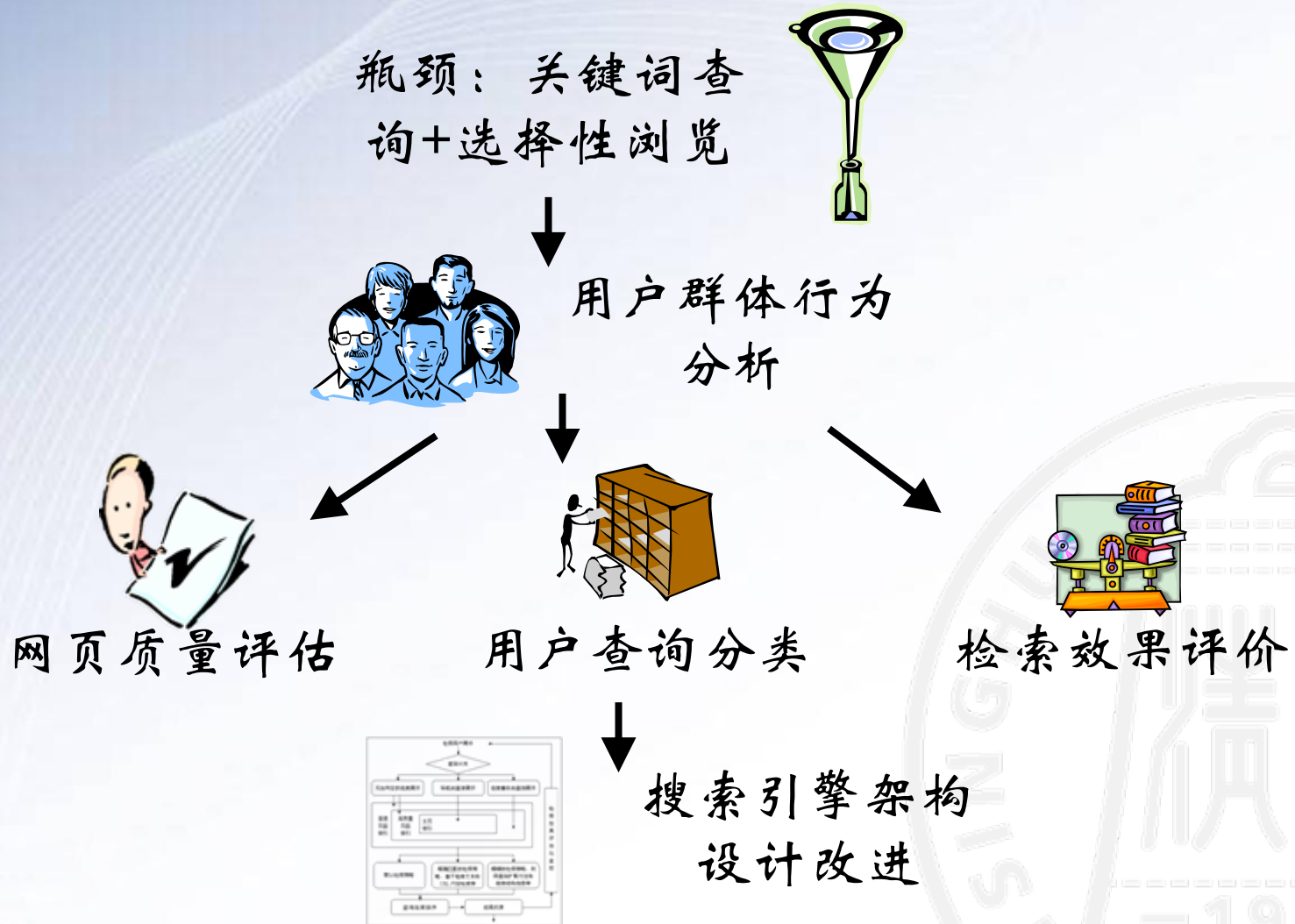
- 与手工评价结果类似
 - 利用用户群体行为分析方法进行搜索引擎自动性能评价是可行的





Information Institute of Tsinghua University

提纲





网络信息检索系统架构设计

- 改进方式

- 根据网络数据质量的不同，构建层次索引结构
 - 网页质量评估
- 根据用户信息需求的不同，采用不同的查询处理策略，对不同级别的索引进行检索
 - 用户信息需求分类
- 对信息检索系统的检索效果进行实时监控
 - 搜索引擎效果自动评价



层次索引结构设计

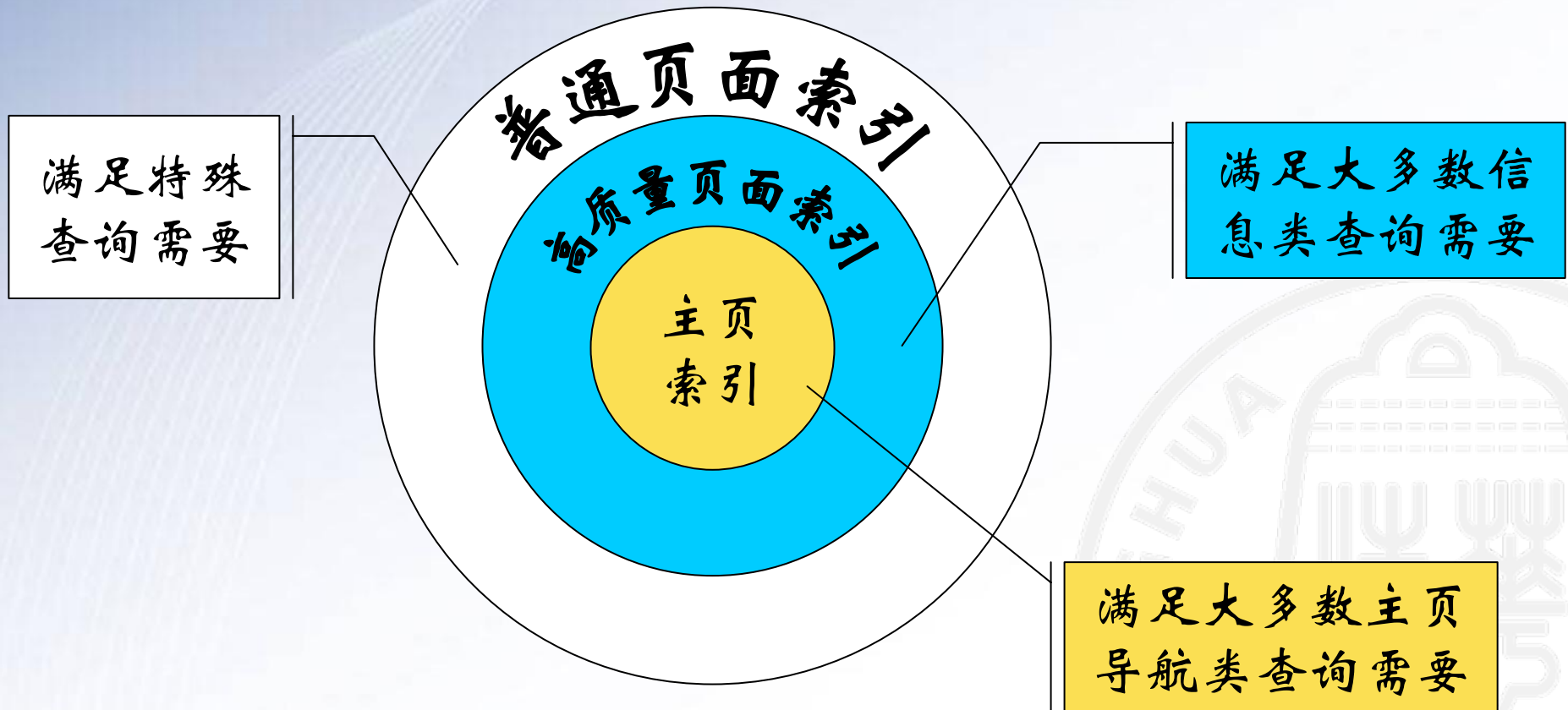
- 不同页面在检索系统中的需求量差异
 - 页面质量评估：质量最高的5%页面满足了测试集中超过92%的用户查询需求
 - 分级索引结构
 - 主页索引：约百万量级，涵盖所有主要站点首页（来源于网页目录）
 - 高质量页面索引：约占索引总量的5%，基本包含主页索引的所有页面
 - 全部页面索引：搜索引擎索引到的所有页面



Information Engineering @ Tsinghua University

层次索引结构设计

- 索引结构的层次存储结构





Information Engineering Department, Tsinghua University

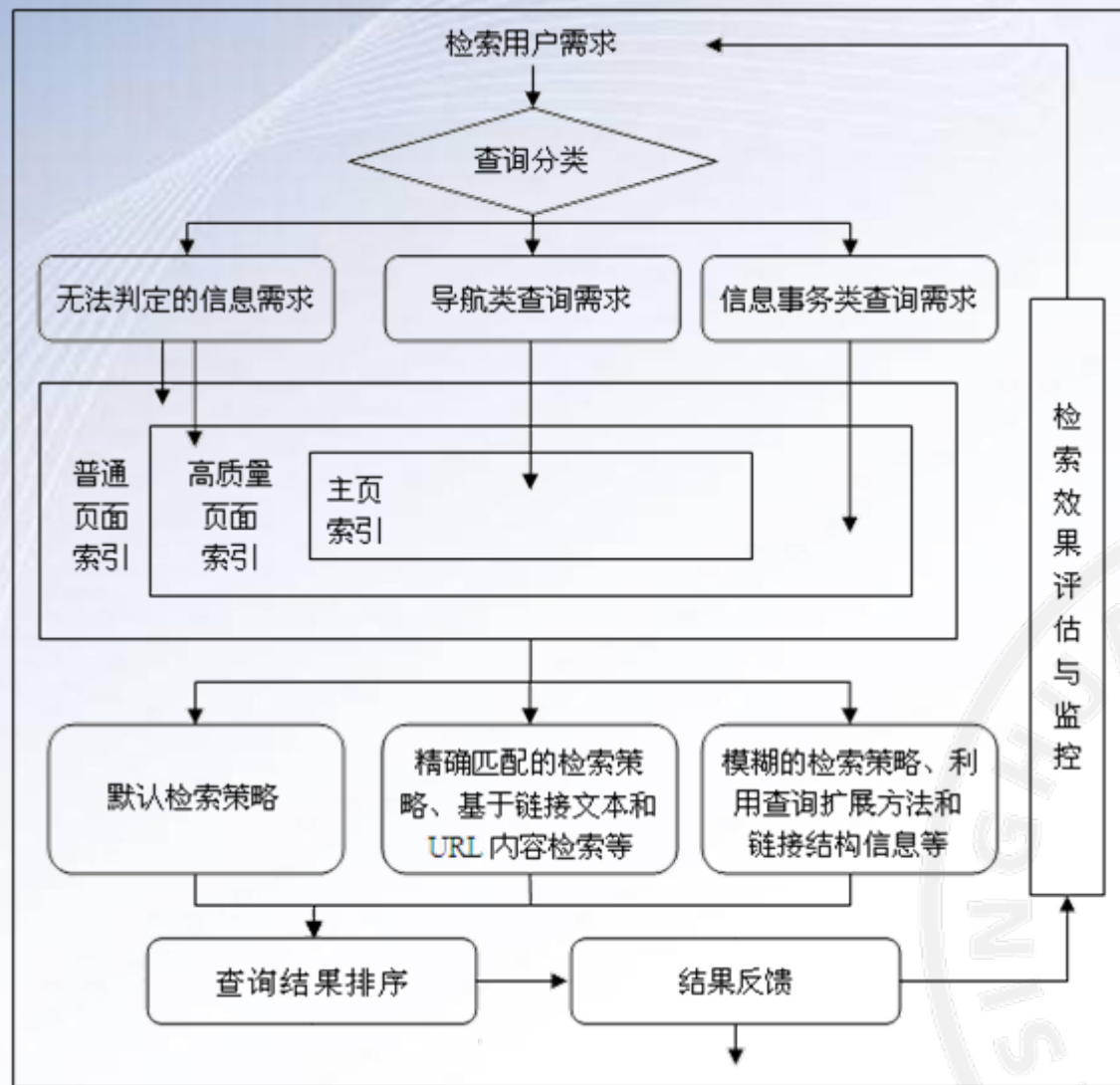
查询处理策略与效果监控

- 以查询信息需求指导检索模型设计
 - 针对不同的查询需求类别
 - 指导相似度计算方式
 - 指导页面质量评估指标使用
 - 指导文本结构化信息的利用
 - 指导索引选择策略
- 检索效果的实时监控
 - 检索系统自身运营维护的需要
 - 及时发现并预警系统问题



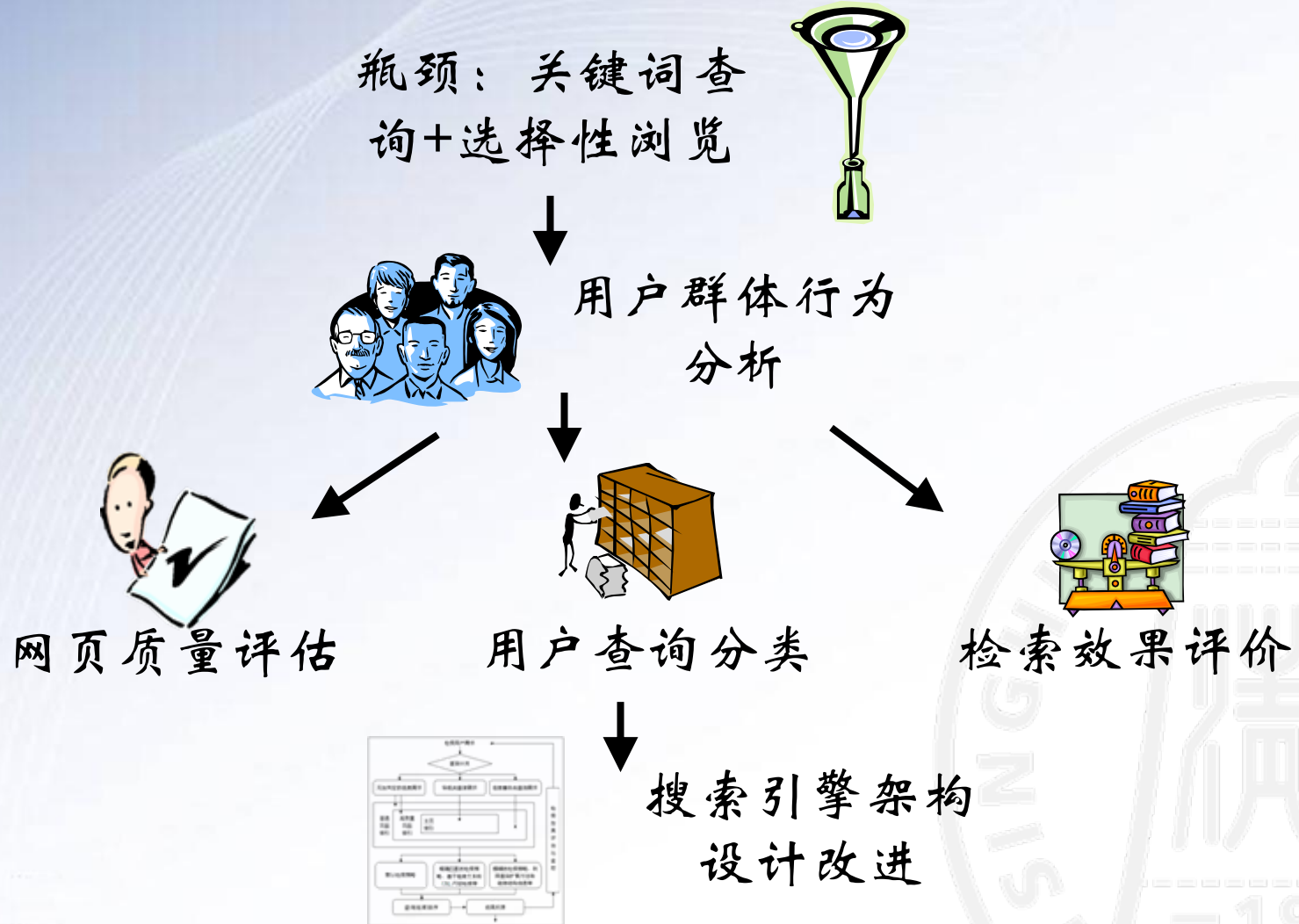


系统结构设计





结论





结论

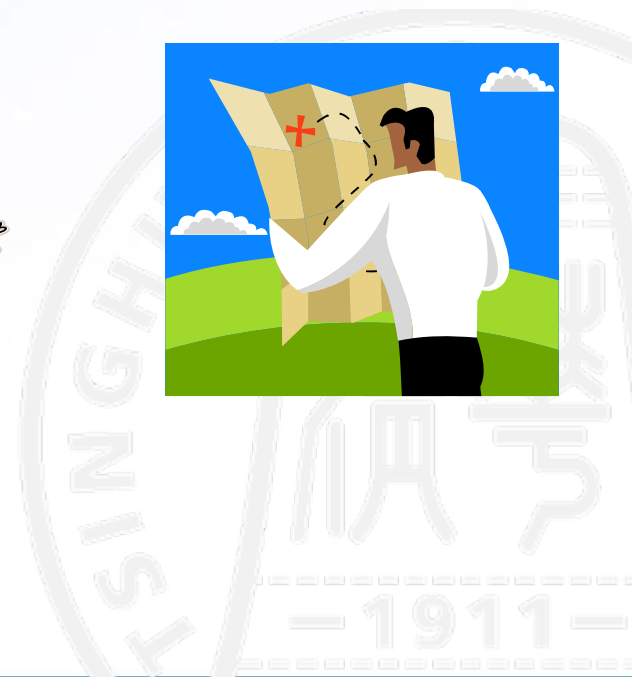
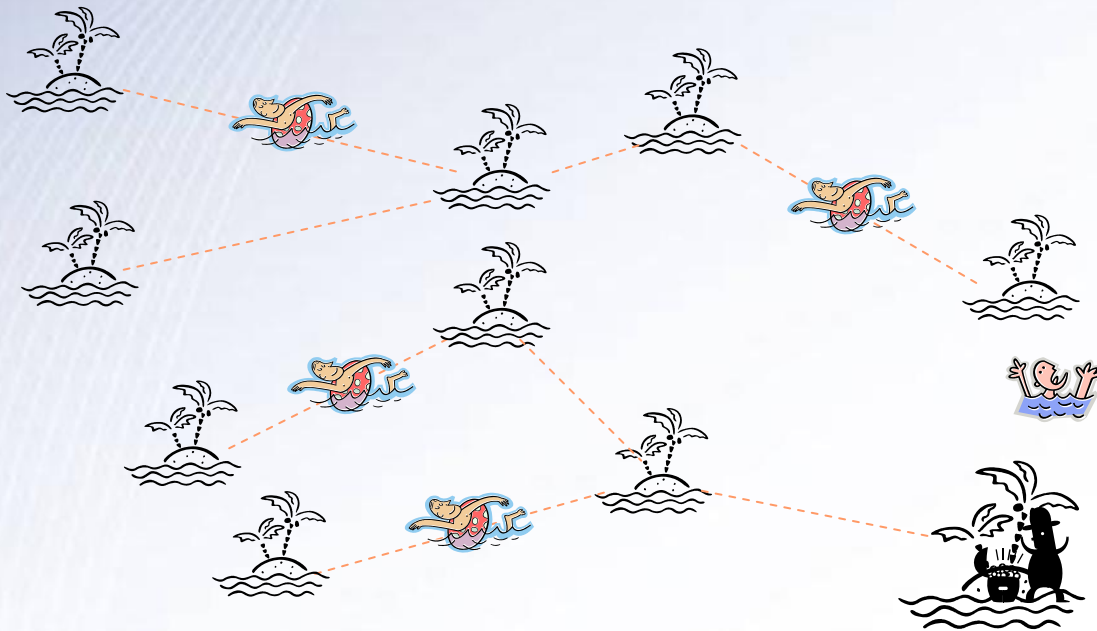
- 用户群体行为分析的方法一直是信息检索相关研究向前发展的推动力
 - 实质：对人类信息获取规律的研究
- 论文工作特点
 - 从用户空间而不是网络数据空间着手
 - 提炼出用户信息需求传递的瓶颈问题
 - 针对搜索引擎发展面临的三个重要的基础性问题进行研究
 - 整合研究成果，改进系统结构设计



Information Engineering @ Tsinghua University

结论

- 海量规模的互联网数据资源，只有依靠海量规模用户行为规律的研究才能更好的解读。





Information Systems © Tsinghua University

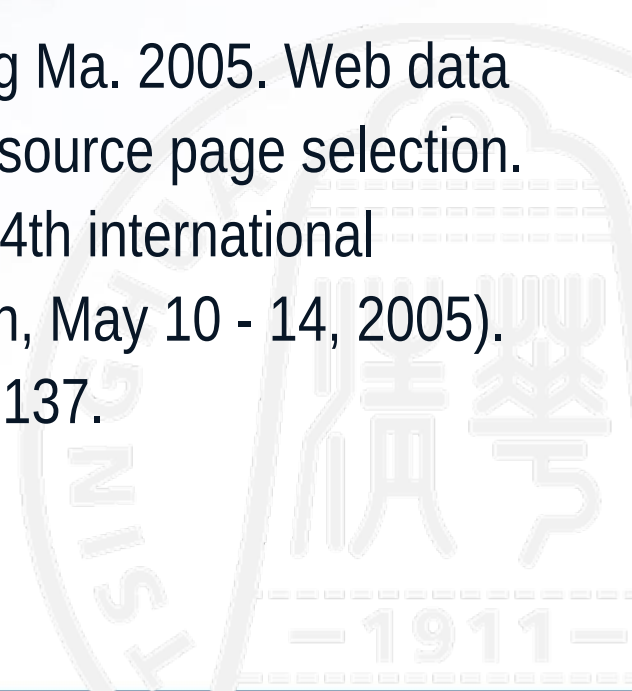
论文工作成果

- [1] 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). (已录用, SCI源刊, 影响因子1.583) .
- [2] Yiqun Liu, Canhui Wang, Min Zhang, Shaoping Ma, Learning-based Web Data Cleansing for Information Retrieval, Journal of Computational Information Systems, Vol 1(n4): pp. 709-716, 2005. (EI 收录, 检索号063610101006)
- [3] Yiqun Liu, Min Zhang, Shaoping Ma, Effective Topic Distillation with Key Resource Pre-selection, Lecture Notes in Computer Science, Volume 3411, 129 – 140, 2004. SCI indexed. (SCI收录, 检索号BCF37, EI收录, 检索号05379358906)



论文工作成果

- [4] Yiqun Liu, Min Zhang, Liyun Ru, Shaoping Ma, Automatic Query Type Identification Based on Click Through Information, Asia Information Retrieval Symposium, AIRS 2006, in Lecture Notes in Computer Science Vol. 4182: pp. 593-600, 2006. SCI indexed.
(SCI收录, 检索号BFG22, EI收录, 检索号064810275745)
- [5] Yiqun Liu, Canhui Wang, Min Zhang, Shaoping Ma. 2005. Web data cleansing for information retrieval using key resource page selection. In Special interest Tracks and Posters of the 14th international Conference on World Wide Web (Chiba, Japan, May 10 - 14, 2005). WWW '05. ACM Press, New York, NY, 1136-1137.





Information Engineering @ Tsinghua University

论文工作成果

- [6] Yiqun Liu, Yupeng Fu, Min Zhang, Shaoping Ma, Liyun Ru. Automatic Search Engine Performance Evaluation with Click-through Data Analysis. In the Special interest Tracks and Posters of the 16th international Conference on World Wide Web (Banff, Canada, May 8 - 12, 2007). WWW '07. ACM Press, New York, NY, 1133-1134.
- [7] 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. (EI收录, 检索号: 06129772414)
- [8] Yiqun Liu, Canhui Wang, Min Zhang, Shaoping Ma, Finding "Abstract Fields" of Web Pages and Query Specific Retrieval--THUIR at TREC 2004 Web Track, NIST Special Publication: SP 500-261, Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004).



论文工作成果

- [9] 刘奕群,张敏,马少平.利用虚拟站点定位技术的网络信息检索研究. 中文信息学报, Vol. 19(2): pp. 44-50, 2004. (核心期刊)
- [10] 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. (核心期刊)
- [11] 刘奕群,张敏,马少平.用户行为分析在网络信息检索中的应用概述. 全国第七届计算语言学联合学术会议, 南京, pp 488-494, 2005.
- [12] Yiqun Liu, Min Zhang, Liyun Ru and Shaoping Ma, Data Cleansing for Web Information Retrieval using Query Independent Features. Proceedings of the 1st China-Kyoto International Student Workshop on Digital Content and Web Computing. 2006.3.



Information Partner of Tsinghua University

论文工作成果

- [13] 刘奕群,张敏,马少平.基于非内容信息的网络关键资源有效定位.智能系统学报,2007年1月。
- [14] 刘奕群,岑荣伟,张敏,马少平,茹立云.基于用户行为分析的搜索引擎自动性能评价,软件学报,已通过初审。
- [15] 刘奕群,张敏,金奕江,马少平.一种搜索引擎性能评价的自动化处理方法.专利申请号:200610144289.X,实质审查阶段。





Information Partner © Tsinghua University



Thank you!

Questions or comments?

