一个有关评价的评价

——信息检索系统评价技术的历史与现状

智能技术与系统国家重点实验室 智能检索组

刘奕群 2006年10月

信息检索系统评价技术

清华大学
Tsinghua University

- 信息检索的重新认识
- 信息检索系统评价的重新认识
- 信息检索评价方法的回顾
- 基于用户行为分析的自动评价系统

信息检索系统评价技术

清华大学
Tsinghua University

- 信息检索的重新认识
  - 信息检索的起源
  - 信息检索的本质
  - 多个角度看信息检索
- 信息检索系统评价的重新认识
- 信息检索评价方法的回顾
- 基于用户行为分析的自动评价系统

信息检索的再认识

- 信息检索的历史有多久？
  - 从WEB诞生开始？
  - 从个人计算机诞生开始？
  - 从计算机诞生开始！
- 信息检索的起源
  - 二次大战中，美国科学技术领头人之一的Vannevar Bush提出了一个当时看来异常复杂的任务
  - “the massive task of making more accessible the bewildering array of knowledge”（Bush, 1945, As We May Think）
  - 1950年美国政府开始了对信息检索研究的资助

信息检索的再认识

- 信息检索的本质使命
  - 解决各个应用领域信息爆炸的问题
  - 帮助人们进行在面向海量信息时的信息寻找、信息发现、信息使用与信息交互工作。
  - Calvin Mooers 1951年第一次提出了IR的概念
    - 信息检索是用户将其信息需求转换为有存储介质的引用的列表的过程；它的范畴包括了信息的描述及其为检索进行特化的过程。（Mooers, 1951）

信息检索的再认识

- 多个角度认识信息检索（Harter, 1997）
  - 黑盒模型：输入信息需求，输出检索答案列表
  - 交互模型：一个交互通讯过程
  - 一个处理以物理形式存在的知识记录为主，而不是处理人们头脑中知识的专家系统
  - 一个依据网络信息组织结构进行导航、浏览的过程
  - 一个挖掘、发现与创新的过程
  - 一个逻辑推理的过程

信息检索的再认识

- 信息检索系统的评价与对信息检索系统的认识程度紧密相关（Agosti, 1999）
  - 黑盒模型：基于对output的评价
  - 交互模型：信息传递有效性的评价
  - 网络导航模型：能否满足导航浏览需求
  - 专家系统模型：解决问题上的有效程度
  - …

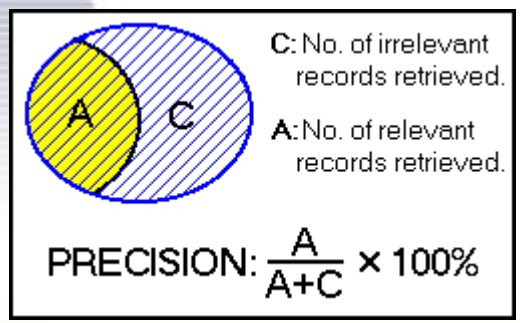
信息检索系统评价技术

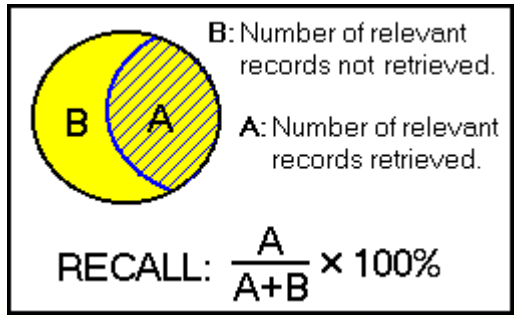
清华大学
Tsinghua University

- 信息检索的重新认识
- 信息检索系统评价的重新认识
  - 信息检索评价的起源
  - 信息检索评价的不同层次
  - 信息检索评价系统的组成
- 信息检索评价方法的回顾
- 基于用户行为分析的自动评价系统

信息检索系统评价的重新认识

- 信息检索系统评价的起源
  - 与信息检索系统一起被人重视
    - Evaluation became central to R&D in IR to such an extent that new designs and proposals and their evaluation became one. （Saracevic, 1995）
  - Kent等人第一次提出了关于Precision和Recall（开始被称为relevance）的概念（Kent, 1955）





信息检索系统评价的重新认识

清华大学 Tsinghua University

- 信息检索系统评价的起源（续）
  - Cranfield-like evaluation methodology
    - Cranfield在上世纪伍十年代末到六十年代初提出了基于查询样例集、标准答案集和语料库的评测方案，被称为IR评价的“grand-daddy”
    - 确立了评价在信息检索研究中的核心地位
  - Gerard Salton & SMART system
  - Sparck-Jones & “Information Retrieval Experiment”
    - Available both online and in-library (G354 FJ77)

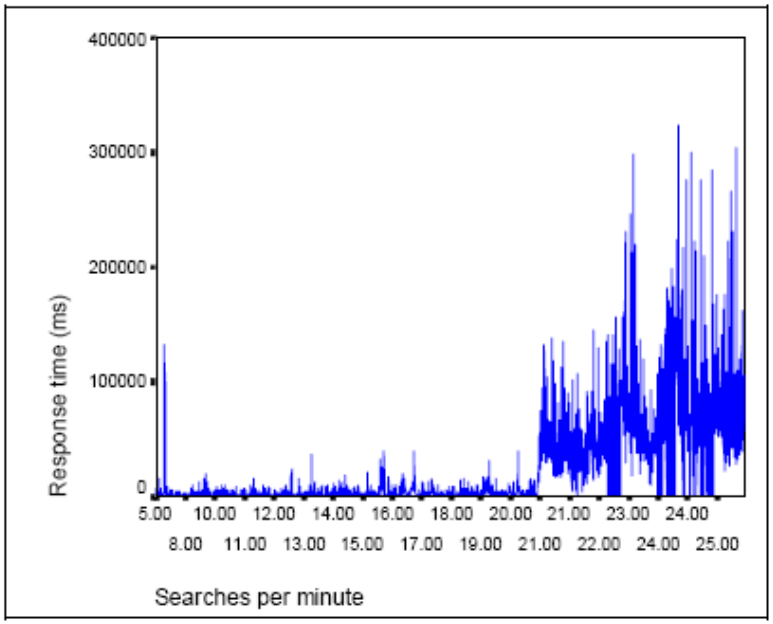
信息检索评价的不同层次

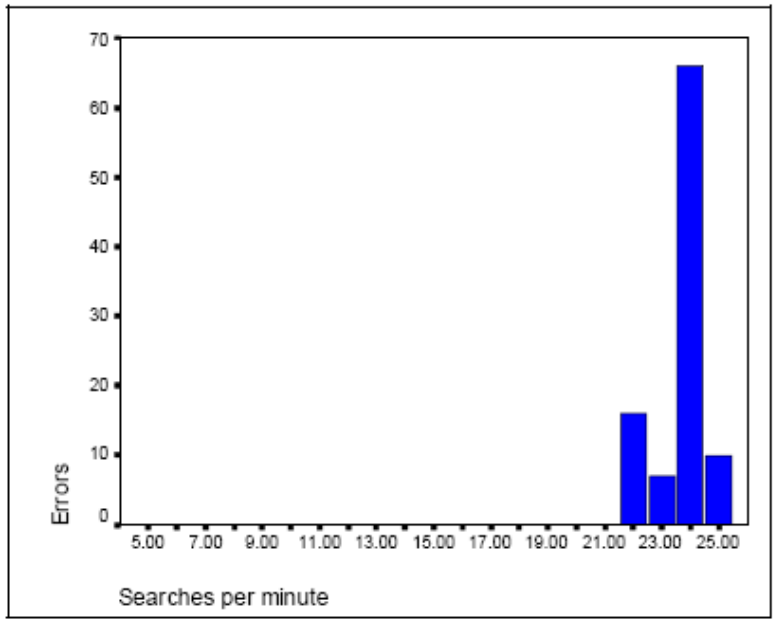
清华大学
Tsinghua University

- Engineering level
  - Hardware/software performance
  - Computational effectiveness and efficiency of given retrieval methods and algorithms
  - Example: USim system
    - Cacheda et al, 2004
    - A simulation tool for the performance evaluation of Web IR systems based on the simulation of users' behavior.

信息检索评价的不同层次

- 性能评价样例





信息检索评价的不同层次

- 性能评价样例

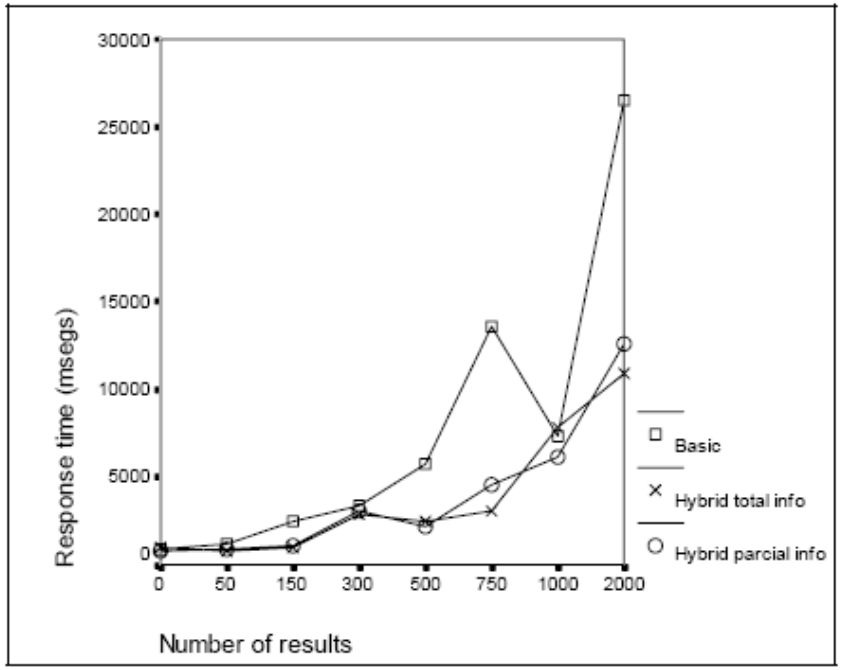


Figure 7: Response (medium workload)

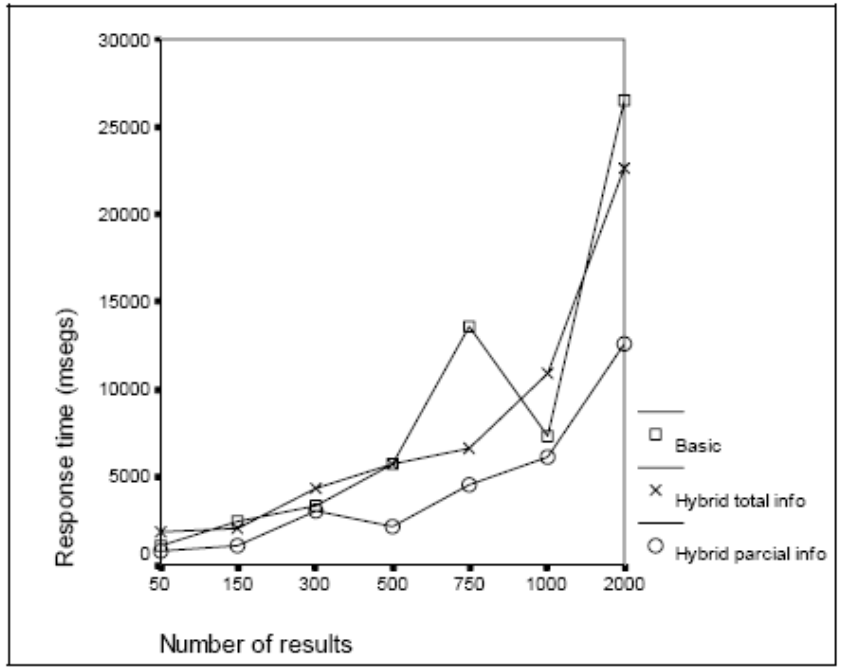


Figure 8: Response time (high workload)

信息检索评价的不同层次

清华大学
Tsinghua University

- Input level
  - Coverage in the designated area
  - 索引数据的规模和代表性如何？
  - 评估用的查询如何按用户查询需求类别构造才具有代表性？
  - Example1: 天网：搜索引擎检索系统质量评估
  - Example2: 天网：一种评价搜索引擎信息覆盖率的模型及其验证
    - 评价Web Infomall的数据代表性与有效性

信息检索评价的不同层次

清华大学 Tsinghua University

- 性能评价样例1
  - 按用户查询需求类别构造查询评估集合
    - 手工判断，使用普遍情况下最大可能的用户需求作为查询类别的判断准则
    - 随机挑选一定量的查询，定义一个用户提交查询,翻页浏览结果的一系列动作为一次查询

Category	Count	Ratio(%)
Navigational	106	12.3
Informational	535	62.0
Transactional	222	25.7
Others	37	

信息检索评价的不同层次

清华大学 Tsinghua University

- 性能评价样例2
  - 数量覆盖率
    - 随机IP地址作为种子进行扩展
    - 广度优先（类似crawler）
    - 提取出的URL的覆盖率的均值认为是总的覆盖率

样本编组	1	2	3	4
种子 URL	www.21cn.com	www.etang.com	net.cs.pku.edu.cn	www.pku.edu.cn
扩展 URL 数	66891	211629	154314	723866
覆盖数量	26732	87562	67178	273066
数量覆盖率	40.0 %	41.4 %	43.5 %	37.7 %

信息检索评价的不同层次

清华大学 Tsinghua University

- 性能评价样例2
  - 质量覆盖率
    - 搜集系统搜到的重要网页占WWW 中该类网页总量的比例
    - 根据页面的PageRank值判断页面的重要性
    - 高PageRank页面的覆盖率即为质量覆盖率

样本编组	1	2	3	4
种子 URL	www.21cn.com	www.etang.com	net.cs.pku.edu.cn	www.pku.edu.cn
初始扩展数量	66891	211629	154314	723866
PageRank 取数	3523	8497	8702	33436
所占比例	5.3 %	4.0 %	5.6 %	4.6 %
覆盖 URL 数	1542	4032	4754	17717
质量覆盖率	43.8 %	47.5 %	54.6 %	53.0 %

信息检索评价的不同层次

- Processing level
  - Assessment of performance of algorithms, techniques, approaches, and the like
  - Cranfield (Cleverdon, Mills & Keen, 1966)
  - SMART (Salton,1971, 1989)
  - TREC (Harman, 1995).
  - 对搜索引擎而言，如何消除不同搜索引擎索引数据集差异对评估的影响？需要消除这个影响么？（Example: 天网）

信息检索评价的不同层次

- 性能评价样例
  - 消除不同搜索引擎索引数据集差异对评估的影响
    - 把查询结果限定在一个固定网页集合内
    - 只承认在某个集合中的检索结果
  - 事实上评价的结果同样受到搜索引擎索引数据集与固定网页集合交集大小的影响
  - 搜索引擎索引数据的有效性本身，是比搜索引擎排序算法有效性更有价值的评价内容

信息检索评价的不同层次

清华大学
Tsinghua University

- Output level
  - Assessment of searching, interactions, feedback
  - 从用户角度对搜索引擎的易用性进行评价（Byrne，1999）
    - 界面友好程度
    - 可读性，易用性
    - 帮助文档、提示文档等是否全面清晰
    - 高级检索功能、检索语法是否易用
    - 个人化功能的多少
    - 查询纠错、查询提示
    - 网页快照

信息检索评价的不同层次

清华大学
Tsinghua University

- Social level
  - Impact on the environment, productivity, decision-making in a given area
  - Example1: studies of impact of MEDLINE on clinical decision making (Lindberg et al., 1993)
  - Example2: 各类搜索引擎市场调查通常涉及的内容：搜索引擎营销应用状况及效果评价方式
  - Example3: 根据Toolbar日志进行的搜索引擎流量引导情况调查

信息检索评价的不同层次

- 性能评价样例
  - 搜索引擎流量引导情况调查

	总访问量	文学	下载	音乐	新闻	购物	人才	邮箱
百度	20463	3599	3276	894	6326	3812	701	1855
Google	5558	654	739	174	2432	1203	157	199
雅虎中国	5747	854	436	184	1964	1717	122	470
搜狗	4466	954	60	131	2327	343	103	548
中搜	485	73	76	59	166	71	10	30
msn中国	25	8	4	0	10	0	0	3
tom	5	0	0	0	5	0	0	0

信息检索评价系统的构成

清华大学
Tsinghua University

- System
  - Smart, TREC
- Criteria
  - relevance, utility, success, completeness, satisfaction, worth, value, time, cost
- Measure based on the criteria
  - precision and recall (many proposals for a unified measure)

信息检索评价系统的构成

清华大学
Tsinghua University

- Measuring instrument
  - Assessors
    - based on an exhaustive examination of the documents
    - Limited applications
  - Pooling
    - Proposed by Karen Sparck Jones and Keith van Rijsbergen
    - Fit for TREC-sized collections

信息检索评价系统的构成

清华大学
Tsinghua University

- Methodology
  - A long tradition in examining, and challenging the methodologies used in evaluation.
    - Blair & Maron proposed in 1985 a large evaluation of full text retrieval
    - Salton (1986) raised a number of methodological issues bringing the whole project into question
    - Blair & Maron (1990) retorted explaining and justifying their methodology for evaluation

信息检索系统评价技术

清华大学
Tsinghua University

- 信息检索的重新认识
- 信息检索系统评价的重新认识
- 信息检索评价方法的回顾
  - 基于标准测试语料库的评价
  - 基于用户日志的评价
  - 其他评价方式
- 基于用户行为分析的自动评价系统

信息检索评价方法的回顾

清华大学
Tsinghua University

- 针对整体，各有侧重
  - Test collection approaches
  - Search log analysis
  - Evaluating IR by human experimentation in the lab
  - Naturalistic observation

信息检索评价方法的回顾

清华大学
Tsinghua University

- Test collection approaches (Black-box)
  - Cranfield-like evaluation methodology
  - The style of experiment taken up by the TREC, CLEF, INEX, and NTCIR conferences
  - Composition
    - a standard corpus of documents
    - a large set of information needs which may be satisfied by documents in the corpus
    - “complete” lists of relevant documents corresponding to each information need

信息检索评价方法的回顾

清华大学
Tsinghua University

- Test collection approaches: Advantages
  - The low cost of evaluating a system once the collection is in place;
  - Reproducibility of experiments;
  - Reusability of the collection;
  - Possibility of creating test collections including a sufficient number of information needs to permit robust, reliable comparisons.

信息检索评价方法的回顾

清华大学
Tsinghua University

- Test collection approaches: Disadvantage
  - personal information corpora almost always contain private data
  - some corpora will be rapidly evolving and may even change from use to use
  - most search will typically cover tens of billions of documents as most searches will include the Web
  - information needs are likely to be diverse and unarticulated
  - judgements seem likely to be set-based and contextual

信息检索评价方法的回顾

清华大学
Tsinghua University

- Test collection approaches
  - The possibility of using TREC-style evaluation for Web search engines (Craswell, 2001)
    - TREC VLC show that TREC-style systems outperforms commercial search engines
  - Differences between TREC and real web search
    - The importance of hyper-links
    - Different topics
    - Document quality
    - Duplicates

信息检索评价方法的回顾

清华大学
Tsinghua University

- Test collection approaches
  - SIGIR 99 workshop (Evaluation of Web Document Retrieval) conclusion:
    - The Cranfield model may be the most adequate one if user studies are hard and expensive to do.
    - More work should be done to investigate what ``Web user's information need'' means.
    - Web queries are rather different from th e TREC topics.

信息检索评价方法的回顾

清华大学
Tsinghua University

- Search log analysis
  - Major problems
    - "Trust bias" leads to more clicks on high-ranked documents, regardless of the documents' utility, as a result of users' faith in IR systems.
    - "Quality bias" is the result of users being given a set of documents to choose amongst, not a single document at a time.
  - Example: Unbiased Clickthrough Data for Comparing Search Engines (Joachims, 2002)

信息检索评价方法的回顾

清华大学 Tsinghua University

- Search log analysis

Google Results:

1. Kernel Machines
   http://svm.first.gmd.de/
2. SVM-Light Support Vector Machine
   http://ais.gmd.de/.../svm_light/
3. Support Vector Machine ... References
   http://svm.....com/SVMrefs.html
4. Lucent Technologies: SVM demo applet
   http://svm.....com/SVT/SVMsvt.html
5. Royal Holloway Support Vector Machine
   http://svm.dcs.rhbnc.ac.uk/
6. Support Vector Machine - The Software
   http://www.support-vector.net/software.html
7. Support Vector Machine - Tutorial
   http://www.support-vector.net/tutorial.html
8. Support Vector Machine
   http://jbolivar.freeservers.com/

MSNSearch Results:

1. Kernel Machines
   http://svm.first.gmd.de/
2. Support Vector Machine
   http://jbolivar.freeservers.com/
3. An Introduction to Support Vector Machines
   http://www.support-vector.net/
4. Archives of SUPPORT-VECTOR- ...
   http://www.jiscmail.ac.uk/lists/...
5. SVM-Light Support Vector Machine
   http://ais.gmd.de/.../svm_light/
6. Support Vector Machine - The Software
   http://www.support-vector.net/software.html
7. Lagrangian Support Vector Machine Home Page
   http://www.cs.wisc.edu/dmi/lsvm
8. A Support ... - Bennett, Blue (ResearchIndex)
   http://citeseer.../bennett97support.html

Combined Results:

1. **Kernel Machines**
   http://svm.first.gmd.de/
2. Support Vector Machine
   http://jbolivar.freeservers.com/
3. **SVM-Light Support Vector Machine**
   http://ais.gmd.de/~thorsten/svm_light/
4. An Introduction to Support Vector Machines
   http://www.support-vector.net/
5. Support Vector Machine and Kernel Methods References
   http://svm.research.bell-labs.com/SVMrefs.html
6. Archives of SUPPORT-VECTOR-MACHINES@JISCMAIL.AC.UK
   http://www.jiscmail.ac.uk/lists/SUPPORT-VECTOR-MACHINES.html
7. **Lucent Technologies: SVM demo applet**
   http://svm.research.bell-labs.com/SVT/SVMsvt.html
8. Royal Holloway Support Vector Machine
   http://svm.dcs.rhbnc.ac.uk/
9. Support Vector Machine - The Software
   http://www.support-vector.net/software.html
10. Lagrangian Support Vector Machine Home Page
    http://www.cs.wisc.edu/dmi/lsvm

信息检索评价方法的回顾

清华大学 Tsinghua University

- Search log analysis
  - The data was gathered from three users during the 25th of September and the 18th of October, 2001, using a simple proxy system.
  - 180 queries and 211 clicks were recorded

A	B	$c_a > c_b$ (A better)	$c_a < c_b$ (B better)	$c_a = c_b > 0$ (tie)	$c_a = c_b = 0$	total
Google	MSNSearch	34	20	46	23	123
Google	Default	18	1	3	12	34
MSNSearch	Default	17	2	1	4	24

A	B	$r_a > r_b$ (A better)	$r_a < r_b$ (B better)	$r_a = r_b > 0$ (tie)	$r_a = r_b = 0$	total
Google	MSNSearch	26	17	51	29	123
Google	Default	19	1	1	13	34
MSNSearch	Default	15	1	0	8	24

信息检索评价方法的回顾

清华大学
Tsinghua University

- Evaluating IR by human experimentation in the lab
  - TREC Interactive Track
  - IIR (interactive IR) evaluation model. (Borlund and Ingwersen, 2003)
  - Problems
    - artificial information needs
    - complex experimental design
    - …

信息检索评价方法的回顾

清华大学
Tsinghua University

- Naturalistic observation
  - Beaulieu observed library users as they used catalogue services and continued to browse the shelves.
  - Nordli, Hansen and Jarvelin carried out studies with library users and staff of the Swedish Patent Office.
  - Problems
    - too expensive to get conclusions over enough time
    - risks of altering the behaviour

信息检索系统评价技术

清华大学
Tsinghua University

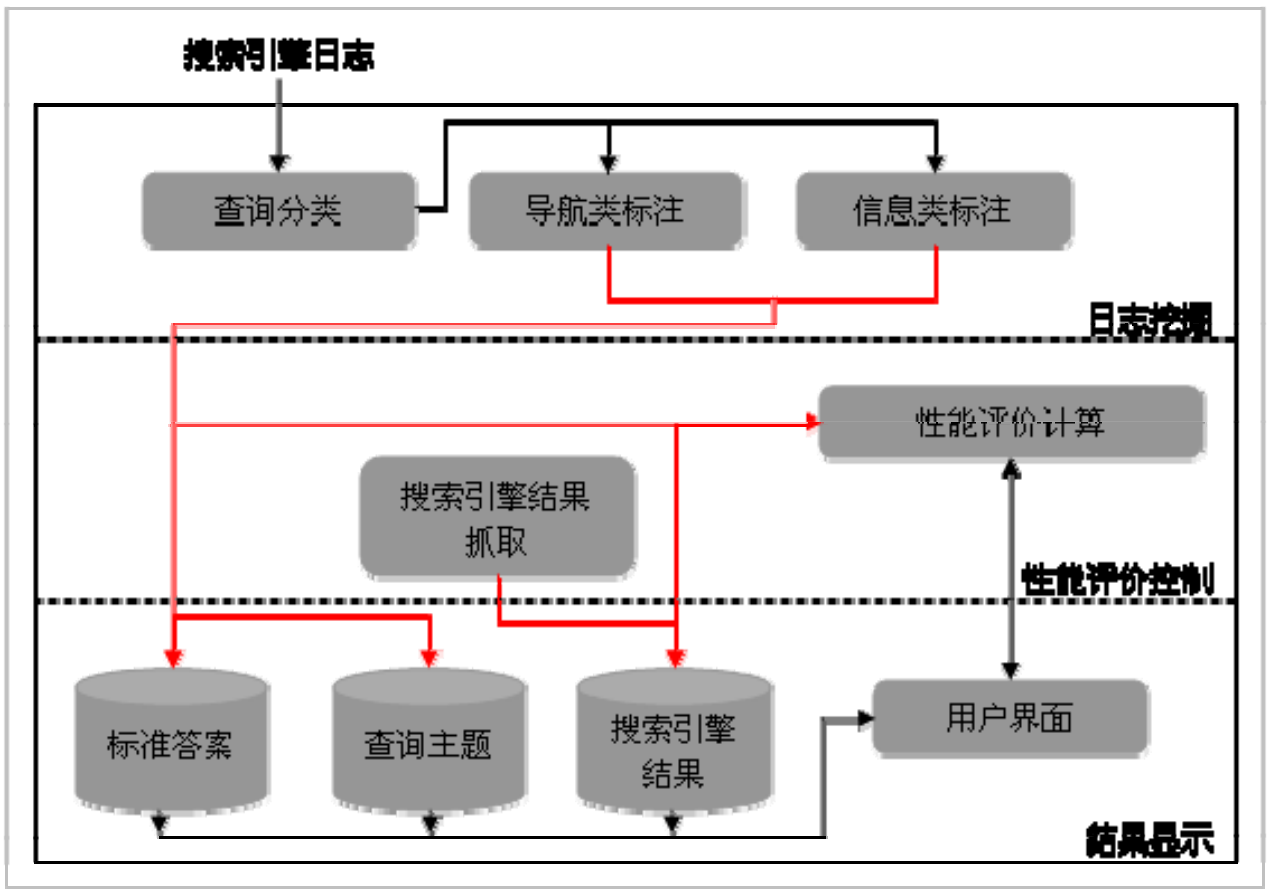
- 信息检索的重新认识
- 信息检索系统评价的重新认识
- 信息检索评价方法的回顾
- 基于用户行为分析的自动评价系统
  - 设计原理
  - 系统架构与运行
  - 评价实验

基于用户行为分析的自动评价系统

- 评价系统的设计定位
  - 涉及到Input level和Processing level的评价
  - Test collection方法和Search log analysis方法的综合
  - 从宏观用户行为的角度客观反映搜索引擎检索质量的变化。
  - 提供一个快速、半即时的评价各大搜索引擎检索效果的平台。

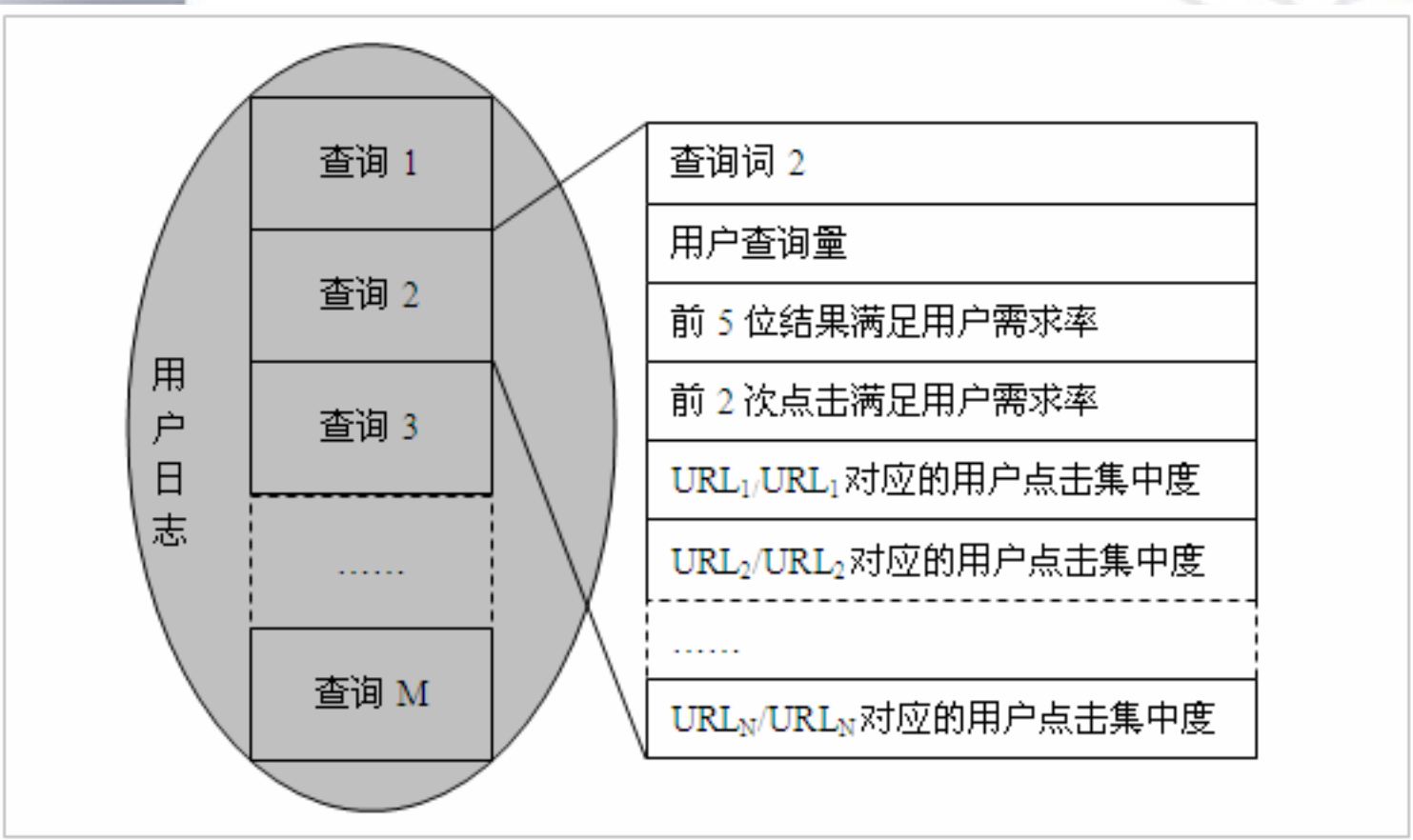
基于用户行为分析的自动评价系统

- 总体设计方案



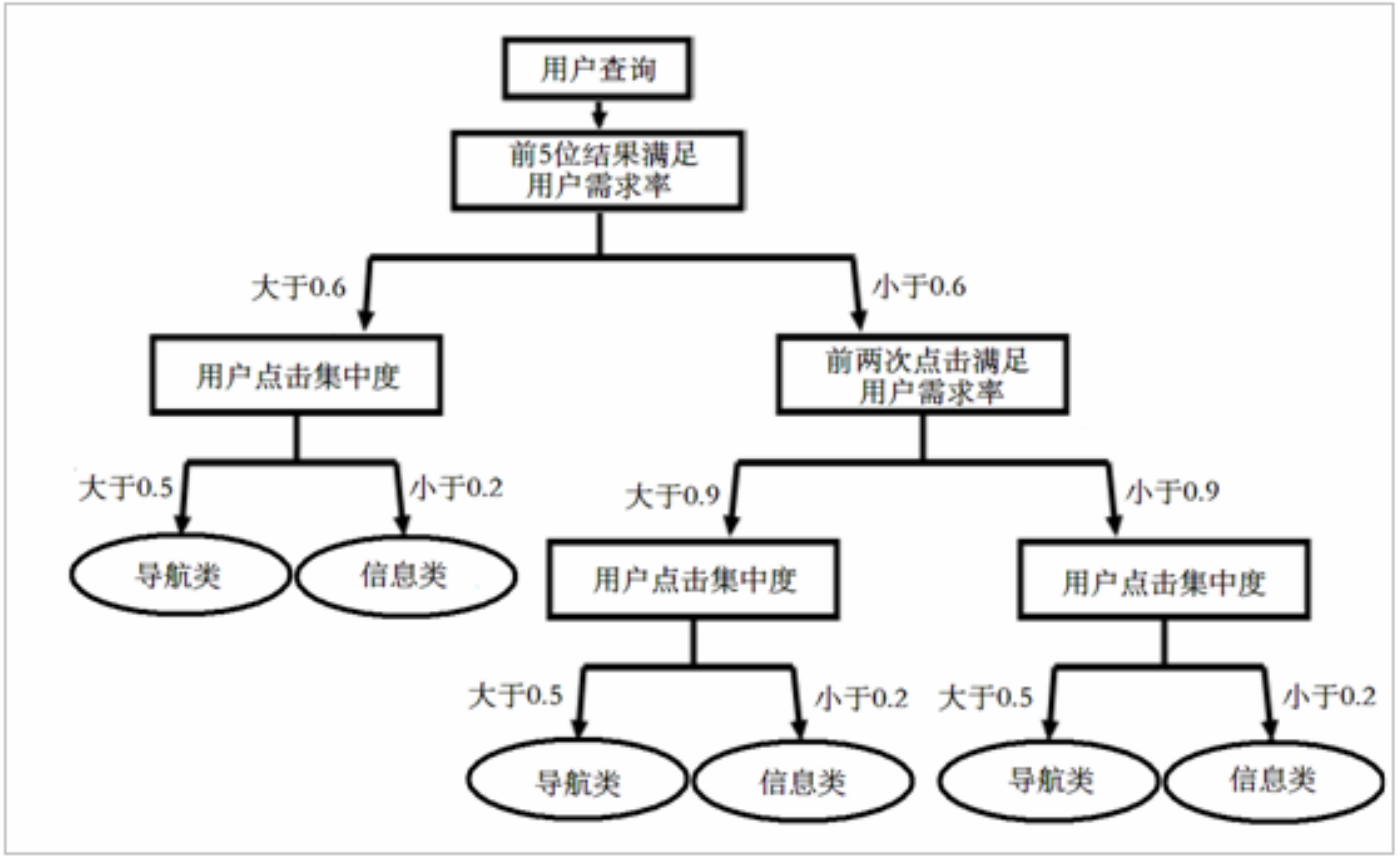
基于用户行为分析的自动评价系统

- 输入：用户查询日志



基于用户行为分析的自动评价系统

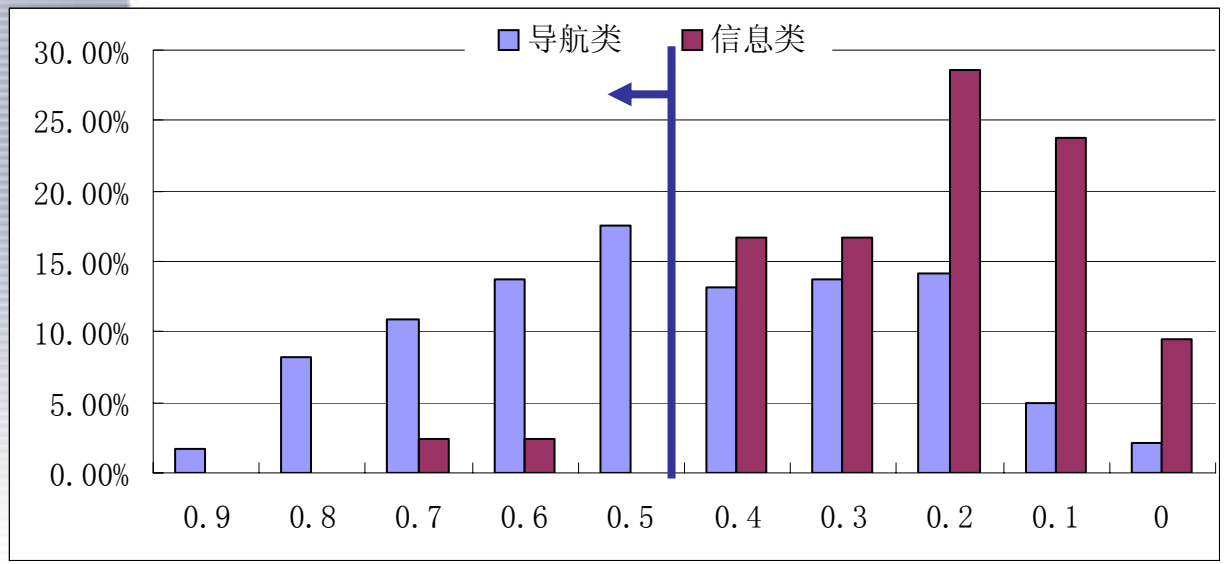
- 查询分类



基于用户行为分析的自动评价系统

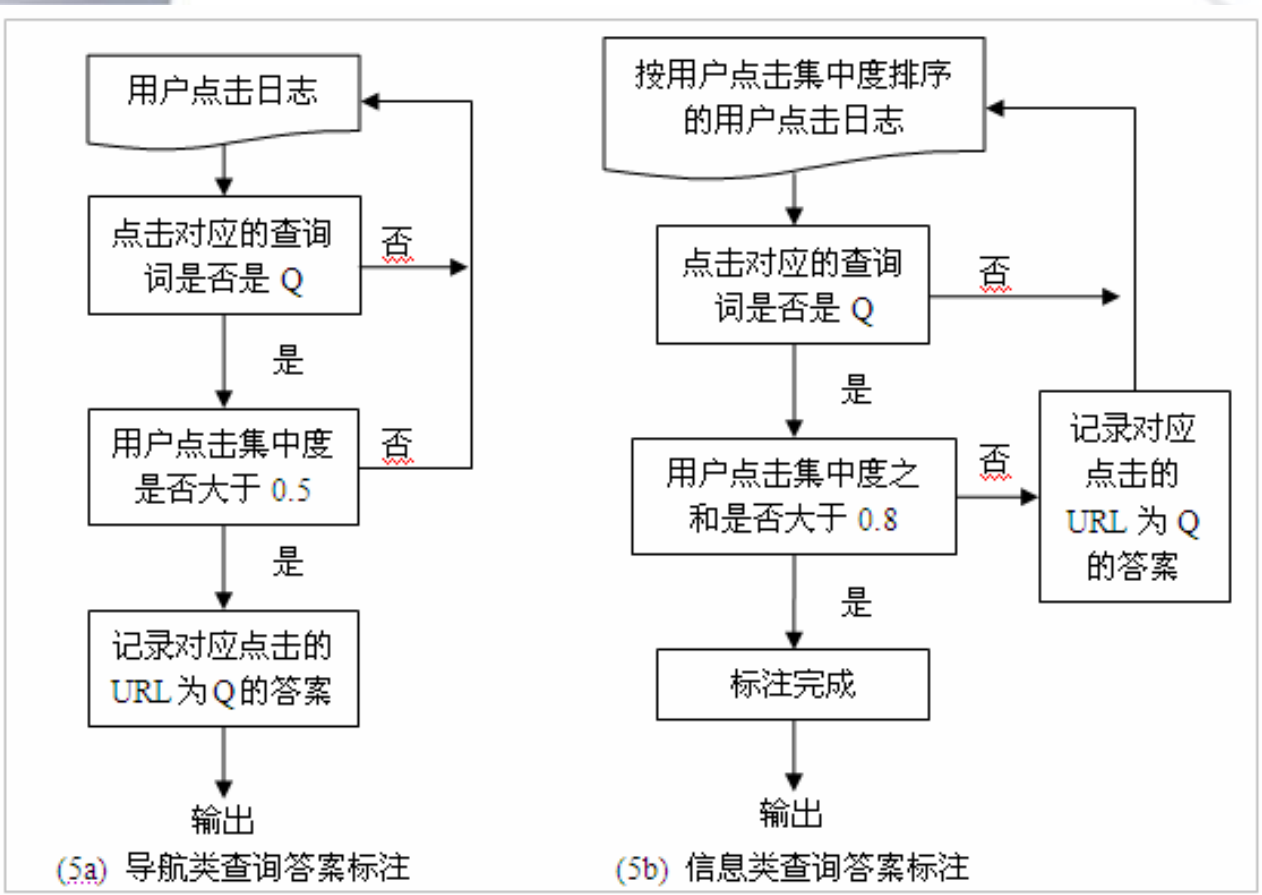
- 查询对应的答案标注
  - 点击的集中程度

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



基于用户行为分析的自动评价系统

- 查询对应的答案标注



基于用户行为分析的自动评价系统

- 查询对应的答案标注
  - 信息类答案标注样例
  - 导航类答案标注样例

基于用户行为分析的自动评价系统

- 评价实验结果（导航类）

ID	Search Engine	自动评价 MAP	自动评价MRR	手动评价MRR
1	Baidu	0.839416	0.839416	0.924
2	Google	0.773198	0.773198	0.888
3	Yisou	0.744804	0.744804	0.827
4	Sina	0.719278	0.719278	0.601
5	Zhongsou	0.595894	0.595894	0.726

基于用户行为分析的自动评价系统

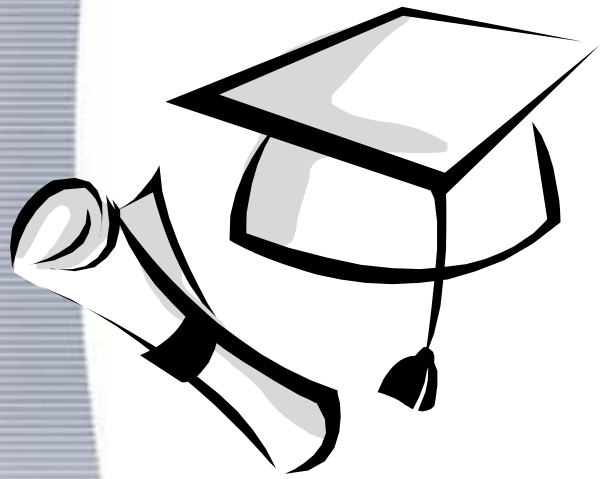
- 评价实验结果（信息类）

ID	SearchEngine	MAP	S@10	手动P@10
1	Baidu	0.377376	0.845666	0.630
2	Google	0.374482	0.866431	0.622
3	Yisou	0.334807	0.818182	0.657
5	Sina	0.375051	0.833805	0.573
6	Zhongsou	0.281622	0.737139	0.511

CNNIC中国搜索引擎市场调查报告给出的“新用户首选的搜索引擎”排序
百度
Google
新浪
3721

基于用户行为分析的自动评价系统

- 主要问题
  - 对Sogou自身的评价问题
  - 答案标注的客观性问题
    - 索引规模、相关性算法的差异
    - 正确性与完整性
- 可能的解决途径
  - 多个搜索引擎日志的结合
  - Toolbar日志的利用
    - 修正答案、其他层次的评价（social level等）



Thank you!

Questions or comments?