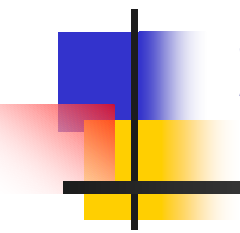
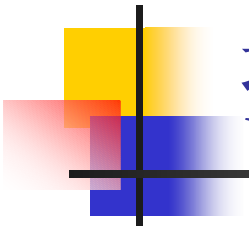


中文搜索引擎用户日志分析

彭波

2004-11-15



提纲

- 相关背景
- 实验设置
- 研究结果
- 进一步工作

目标与内容

- 对搜索引擎用户日志进行分析和挖掘，试图从中发现用户搜索的行为规律，可用于改善和提高系统性能。
- 用户日志有两类：
  - 用户查询日志
  - 用户点击日志

相关研究

- 用户输入查询串平均包含2.2到2.4个英文单词，多数为两个英文单词；
- 查询串所包含的英文单词的数量遵从Poisson 分布；
- 多数用户并不基于返回结果修正查询词；
- 重复查询词的数量遵从Pareto分布；
- 查询串的分布具有明显的局部性，查询串的出现过程具有自相似性特征；
- ……

本文研究的问题

- 中文与英文用户的搜索情况有差异吗？
- 中文用户输入查询串中包含多少个词项？
- 有多大比例的查询串中包含中文字符？
- 用户查看结果页面的时间大概有多长？
- 用户访问系统的时间有什么特点?
- 用户访问量与不同查询串、不同用户量和点击不同url的数量间有什么关系？
- ……

提纲

- 相关背景
- 实验设置
- 研究结果
- 进一步工作

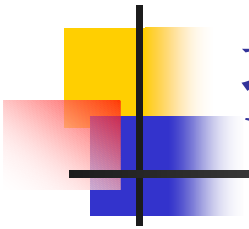
实验设计

- 天网日志
- 选取2003年11月18日0时至24时的用户查询与点击日志 （北大燕穹提供的数据产品中的编号分别为 YQ-QUERYLOG.0311和YQ-CLICKLOG.0311 ）
- 天网用户查询日志
  - 查询时间，用户IP，是否Cache命中，查询串和结果页面编号
- 天网点击日志
  - 点击时间，用户IP，查询串，点击的URL，点击页面的编号，点击URL的序号

数据准备

- 数据清理
  - 非用户行为的查询纪录→删除同一IP且查询次数多于400次以上的记录
  - 错误操作→空查询串剔除。
- 基本统计结果

	总记录数	不同查询/URL	不同IP数
查询日志	125636	43064	21613
点击日志	118008	90184	14795



提纲

- 相关背景
- 实验设置
- 研究结果
- 进一步工作

用户的查询类型与数量

查询类型	数量	百分比
第一次查询	21613	17.2%
修正查询	36719	29.2%
相同查询（包括翻页、重新检索与重新输入）	67304	53.6%
整体	125636	

- 少数用户进行了较多的查询，查询次数的差异比较大 。
- 平均单个用户输入的不同查询串为 2.7，差异要小的多。

查询串中包含的字符类型

查询串的类型	数量	百分比
纯中文	91958	73.19%
纯英文	14593	11.62%
中英混合	8987	7.15%
纯数字	667	0.53%

- 显示了天网用户主要以中文查询为主，同时搜索引擎系统要加强对于中英数字混杂的新词汇的理解和词表建立

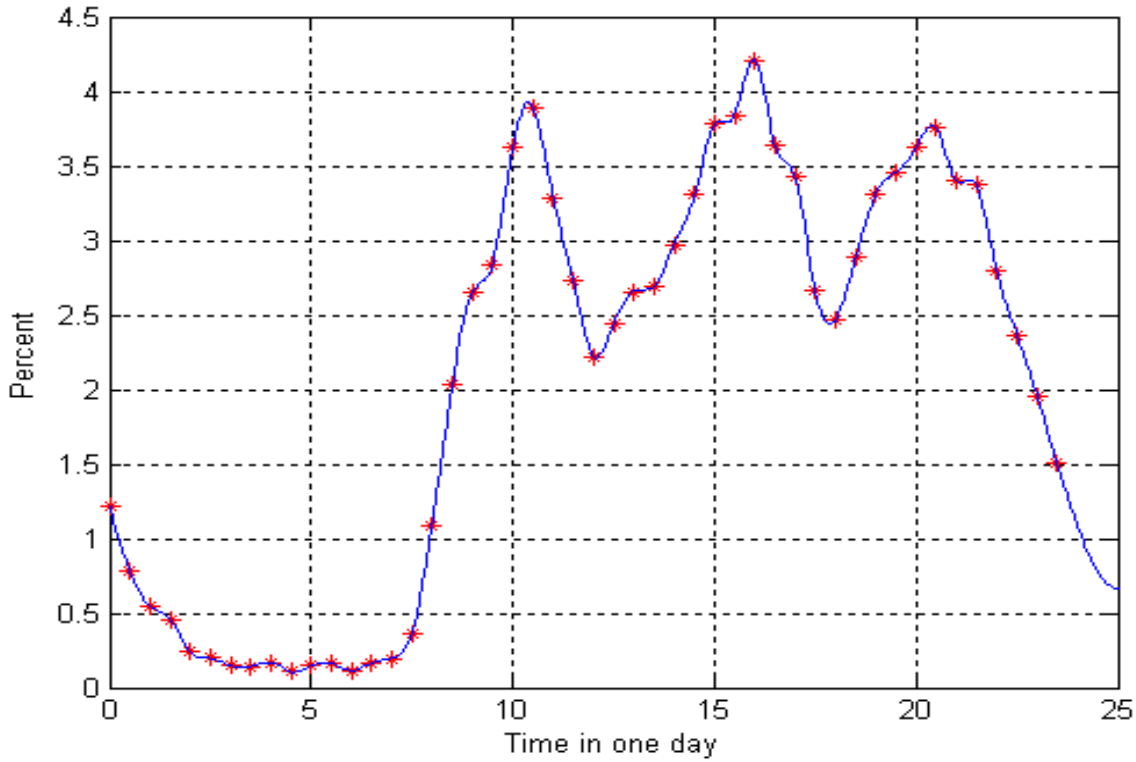
查询串中含有的词项个数

- 87．04%的查询不含空格，包含1个空格的查询占9.16%，包含两个空格的查询占2. 81%，超过两个以上空格的查询不到 1%。这表明多数中文用户只输入一个词项。
- 前100个热点查询串中只有一个查询串包含一个空格，前80个热点查询串没有一个包含空格。

结果页面的查看与时间间隔

- 约有一半（49.8%）的用户只查看了第一个结果页面，80.8%的用户查看了前三个结果页面，只有不到 0.1%的用户查看了10个以上的结果页面。
- 用户翻页的时间间隔大约是2到3分钟 。

用户到达时间的分布

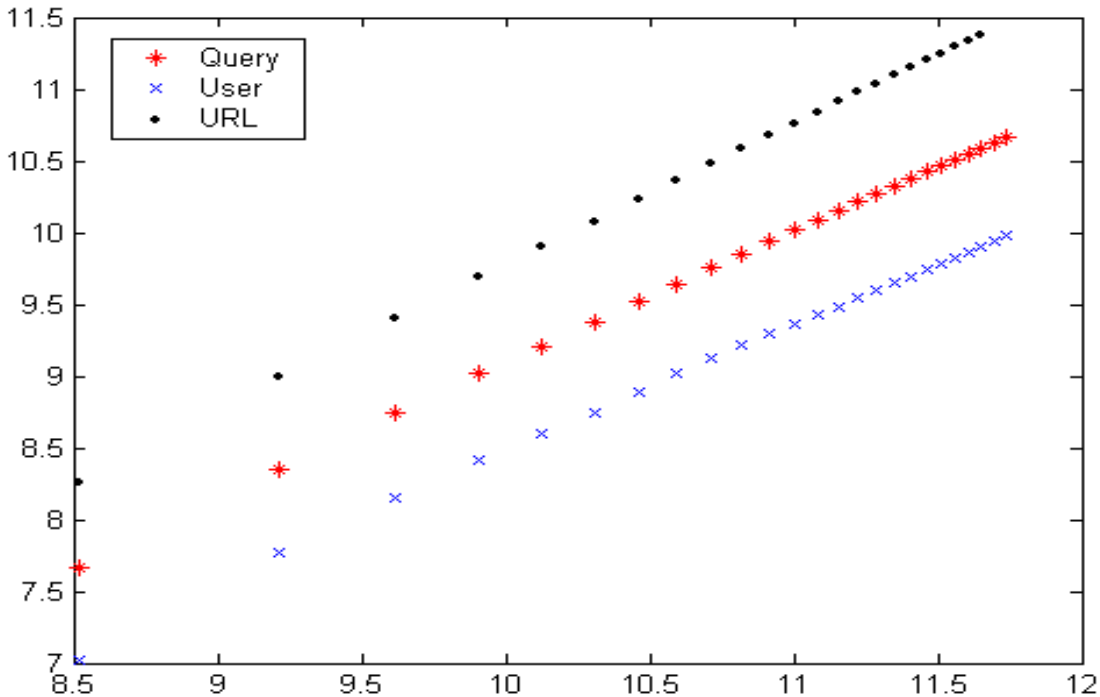


- 三个波峰，早晨10:30，下午 4:30和晚上8:30，最少访问量发生在凌晨 3:00—7:00
- →短期内用户每个工作日的访问量基本相同，时间分布类似，周六周日的到达时间略有差异,整体呈现周期性波动

用户点击URL与历史网页

- 近70%的用户点击url的次数不大于6，约有20%的用户其点击url的次数大于10 。
- 所有查询用户中有68. 5%（即多于2/3）的用户点击了系统反馈的某一查询结果。
- 4213个历史网页被点击，只占总点击量的3.5%，这些历史网页来自于1112个不同的IP，占点击用户总量的7. 5% 。

不同查询串、用户量和url数量



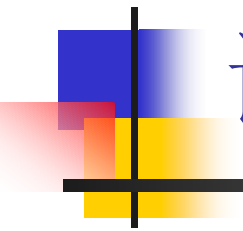
- $M=C*N^a$，满足Heaps定律

小结

- 用户输入的查询串一半以上是重复的。
- 87．04%的查询不含空格。
- 查询主要是中文查询，而中英文混合词也占一定比例
- 约有一半的用户只查看了第一个结果页面。
- 查看历史网页(或称网页快照)的用户占的比例比较小。
- 用户的访问时间并不均等，一天中出现三个波峰；其分布整体呈周期性波动。
- 用户日志中不同查询串、不同用户量和点击不同url的数量满足Heaps定律。

进一步的工作

- 用户查询串的分类，这可以揭示用户对查询内容的关注程度。按时间顺序可揭示用户关注主题信息（兴趣）迁移情况
- 点击URL对应页面的分类，分析用户关注页面的类别特征
- 通过对用户的点击或查询日志进行分析，聚类产生相近查询串、相近页面（URL）或兴趣相近的用户群（提供个性化服务）。



谢谢