# 第 3 章　搜索引擎性能评价

有评测才会有鉴别。
评判一种方法优劣的唯一标准是相互可比的评测，
而不是设计人员自己设计的“自评”，
更不是人们的直觉或某个人的“远见”。
——黄昌宁，清华大学教授，微软亚洲研究院主任研究员

民以食为天，中国的饮食文化可谓源远流长。国家的历史有长有短，疆域有大有小，实力有强有弱，人口有多有少，民族构成、宗教信仰、政权性质和经济结构也有差异，故而各国的饮食文化也有较大差异。直接反映饮食习惯的菜谱，可以说是文化的缩影。下面列出的是作为中国菜代表的啤酒鸭，以及作为西餐代表的苹果派的菜谱，请读者首先过过“眼瘾”。

**啤酒鸭菜谱**

(1) 鸭子洗净，剁成小块后沥干水，豆角择成段后洗净备用；

(2) 健康锅小火预热，将鸭子带皮部分朝下，盖盖；

(3) 鸭油成液体状时，将姜、蒜、干红辣椒、八角放入炒香，再加入啤酒和老抽，盖上锅盖；

(4) 待汁水收到一半左右时，放入豆角，撒上适量盐，继续盖上锅盖；

(5) 豆角软熟时，淋上少许生抽，即可。

**苹果派菜谱**

(1) 将黄油用木棍砸软，和砂糖(100 克)，鸡蛋 50 克，香草粉，发酵粉调匀，放入面粉和成油面，用 2/3 的油面擀成一片，放入排盘，入炉烤至九成熟取出，作为排底；

(2) 将苹果去皮，去核，切成薄片，放入砂糖 100 克拌匀，稍浸去其水分，再加入桂皮粉拌好，摆在排底摊平；

(3) 将其余的油面擀成圆片，盖在苹果上面，刷上鸡蛋用花推子划上花纹，入炉烤熟，凉后切成三角形块 12 块。

百余字的两个菜谱，可以引发我们对于中西饮食乃至中西文化差异的不少思考：苹果派的菜谱中，前前后后用了 5 个数字来准确描述整个制作过程，以保证成品后味道乃至形状能够与菜谱作者预料的一致；但啤酒鸭的菜谱中却是用“左右”、“适量”、“少许”等形容词来介绍制作菜肴中非常关键的收汁、调味等过程。通过这个差异，我们能联想到什么呢？

首先，中国饮食文化的核心是定性，而西方饮食文化的核心是定量。西方厨房里常见的定时器、天平和温度计在中国人的厨房里不是“常客”。

其次，中国饮食文化更注重“人”的作用。每人的口味不同，正所谓“众口难调”，“左右”、“适量”、“少许”正是需要根据各人口味进行调整的过程。

最后，“评价”是完成一道成功的中国菜肴不可缺少的环节。做菜的过程中需要通过“尝”的方式进行评价；完成之后更是要通过“享用”的环节来进行评价。鸭子是否足够酥烂，咸淡是否适中，色泽是否红润，都是需要评价的对象。通过评价，可以总结出下次做菜需要改进的环节，从而达到日臻完美的境界。

综上所述，中国式的菜谱，提供了厨师自我发挥的空间，他可以根据个人喜好对菜式的烹调方式进行调整，并通过评价得到的结果改进自己的厨艺。正是由于这个原因，当烹饪一道中国菜失误时，人们往往会埋怨厨师的技艺，而不会像国外的不少食客一样，去怀疑菜谱的正确性。

从以上的例子中可以发现，“评价”是完成一道漂亮的菜肴乃至培养一个技艺高超的厨师的必由环节。“实验”之后的“评价”是“改进”的必由环节，它不一定会导致正确的“改进”，但是脱离了“评价”的“改进”必然是无的放矢，浪费时间和精力而已。

在搜索引擎相关研究领域，“评价”同样是性能改进乃至保证其顺利运营的至关重要的环节。对互联网用户而言，性能评价意味着选择最有效的信息获取途径；对广告商而言，性能评价协助其选择最有利的广告投放平台；对于研究人员而言，性能评价则是算法改进的指导和正常运营的重要保证。“实验”、“评价”、“改进”三者在搜索引擎性能提高中发挥的作用如图 3.1 所示。

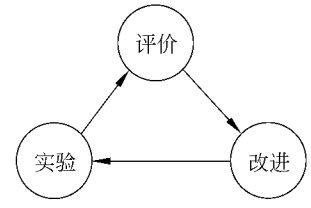


图 3.1　搜索引擎性能改进三部曲：“实验”、“评价”、“改进”

由于性能评价在搜索引擎发展过程中发挥着重要的作用，因此关于搜索引擎性能评价的研究成果与新闻报道也格外丰富，其中不少内容还存在相互矛盾，如表 3.1 中涉及的各个关于搜索引擎性能评价的新闻，其中矛盾的内容不在少数。

表 3.1　近年来关于搜索引擎性能评价的部分新闻报道

| 时　间 | 来　源 | 新 闻 题 目 |
|---|---|---|
| 2007/06/14 | 网易科技 | 中国雅虎副总裁炮轰百度谷歌，称都不如雅虎搜索 |
| 2007/09/07 | eNet 硅谷动力 | 中文搜索引擎流量百度占 74.88%，远超谷歌、雅虎 |
| 2008/01/10 | 艾瑞网 | 谷歌每天处理 2 万 TB 数据，与雅虎、微软相比优势明显 |
| 2008/05/22 | 和讯网 | 雅虎称其数据库超谷歌为全球最大 |
| 2008/11/24 | 华军资讯 | 谷歌网站搜索速度提升，雅虎末路 |

由表 3.1 可见，搜索引擎性能相互比较的话题是各大商业搜索引擎公关宣传活动的重点，然而，对于同样若干个搜索引擎产品，为什么不同信息来源的新闻提供的评价结果却完全不同？这是由于这些新闻大都缺乏一个客观公正的视角，它们往往是从搜索引擎性能的某一个方面（如索引量、用户访问量、搜索反馈速度等）而非搜索引擎系统的整体角度来对搜索引擎的性能进行评价。在第 1 章中，我们提到过这样的观点：搜索引擎是互联网上最重要的应用系统，也是人类历史上最大规模的信息集散平台。对于这样庞杂的一个系统，使用一两个单一指标对其性能进行评价显然是犯了管中窥豹的错误。

那么，面对搜索引擎系统评价的问题，怎样做才能够足够客观、公允、全面，才能够站在搜索引擎研究的角度科学地进行评价呢？我们尝试站在互联网用户的角度，使用信息检索

系统评价中沿用了近半个世纪的 Cranfield 评价体系，结合当前互联网应用的实际总结出一套评价方案。首先，让我们从 Cranfield 评价体系谈起。

## 3.1 搜索引擎评价与 Cranfield 评价体系

按照考察角度的不同，搜索引擎评价可以从两个不同的角度展开。

一方面，搜索引擎是一类网络服务的供应商，通过对接受这种服务的用户使用服务体验的调查，可以获取到对不同搜索引擎的评价数据。这种评价方法被大量的互联网咨询公司和研究机构所采用，国内这方面比较著名的评价报告如中国互联网络信息中心(CNNIC)发布的中国搜索引擎市场调查报告、艾瑞咨询集团(iResearch)发布的搜索引擎行业发展报告、搜索引擎市场监测报告等。

另一方面，搜索引擎与网络信息检索相关技术有着密不可分的关系。尽管搜索引擎包括的研究范畴要大大广于传统的信息检索工具，但使用“黑箱模型[①]”单纯考察其输入、输出关系时，搜索引擎与信息检索工具的性质非常类似，因而也可以将其作为“网络信息检索工具”对其性能进行评价。由于信息检索相关技术的发展已经经历了半个世纪的时间，其评价技术也相对成熟，因此基于传统信息检索评价技术的框架对搜索引擎进行性能评价是学术界经常采用的方法。

“评价”是信息检索领域研究的核心问题之一。1995 年的 SIGIR(ACM Special Interest Group on Information Retrieval，国际计算机协会信息检索专委会，http://www.sigir.org)会议上就有人提出了这样的观点：“评价”在信息检索系统的研发中一直处于核心的地位，以致于算法与其效果评价方式是合二为一的。

值得一提的是，尽管信息检索研究的开展过程中“评价”具有十分重要的作用，但是大部分针对信息检索系统评价的工作都是针对系统的检索效果(Effectiveness)而非效率(Efficiency)进行的。对于搜索引擎性能评价而言，效率层面的考察主要包括用户需求是否得到了很快的响应，为满足用户需求耗费了多大规模的硬件资源等。这方面的研究内容对于搜索引擎系统的构建至关重要。但是对于搜索引擎用户而言，这方面的内容在搜索引擎技术当今的发展阶段并非关注的核心内容。因此，我们所重点研讨的内容也将集中在对搜索引擎系统检索效果(Effectiveness)的评价上进行。

在这方面的研究过程中，Kent 于 1955 年首先提出了“准确率/召回率”的信息检索评价框架(将在第 3.4 节中详述)，随后，美国政府所属的研究机构开始大力支持关于检索评价方面的研究，而英国 Cranfield 工程在 20 世纪 50 年代末到 60 年代中期所建立的基于查询样例集、正确答案集和语料库的评测方案，则真正使信息检索成为了一门实证性质的学科，也由此确立了“评价”在信息检索研究中的核心地位。其评价方法一般被称为 Cranfield 方法框架(A Cranfield-like approach)如图 3.2 所示。

---

① 所谓黑箱(Black Box)模型，又称“闭盒(Closed Box)模型”，是指对于内部结构尚不能直接观测，只能从外部去认识的客体，通过观测其输入输出来认识其性质的方法。最典型的应用之一是在水文学领域，黑箱模型指不考虑流域物理过程，而是基于输入和输出时间序列的分析建立模型的分析方法。在本文中，指只关注搜索引擎系统的输入输出，而不关注搜索引擎的具体实现方式的评价方式。

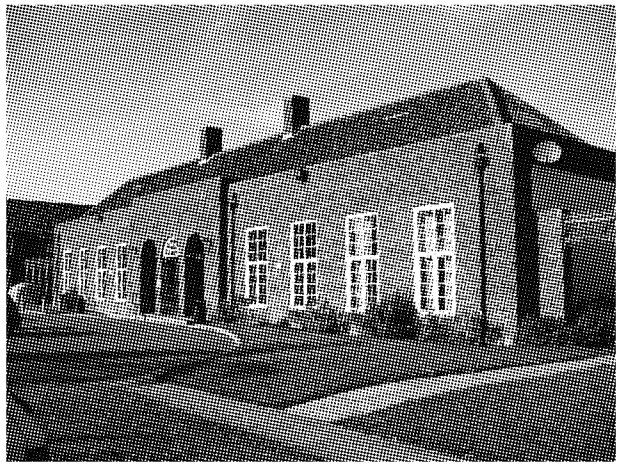

图 3.2　Cranfield 评价方法的发源地：英国 Cranfield 大学

Cranfield 方法指出，信息检索系统的“评价”应由如下 3 个环节组成：首先，确定查询样例集合，抽取最能表示用户信息需求的一部分查询样例构建一个规模恰当的集合；其次，针对查询样例集合，在检索系统需要检索的语料库中寻找对应的答案，即进行正确答案集合的标注；最后，将查询样例集合和语料库输入检索系统，系统反馈检索结果，再利用评价指标对检索系统结果和正确答案的接近程度进行评价，给出最终的用数值表示的评价结果。

Cranfield 方法一直到今天也被广泛地应用于包括搜索引擎在内的大多数信息检索系统评价工作中。由美国国防部高级研究计划署（Defense Advanced Research Projects Agency，DARPA）与美国国家标准和技术局（National Institute of Standards and Technology，NIST）共同举办的 TREC（文本信息检索会议，http://trec.nist.gov）就是一直基于此方法组织的信息检索评测和技术交流论坛。TREC 是目前最大规模的文本信息检索评测平台，迄今为止已有 300 余家单位参与，其中国内的清华大学、复旦大学、中科院计算所、北京大学等单位从 21 世纪初开始参与 TREC 评测，也都取得了较为出色的评测成绩。这说明国内的信息检索相关技术研究水平在国际上处于领先地位，这也是与在搜索引擎产业领域百度等国内公司足以与谷歌、微软、雅虎等巨头分庭抗礼的情况相一致的。

除 TREC 之外，也有一些针对不同语言设计的基于 Cranfield 方法的检索评价论坛在尝试运作，如由日本国家科学信息中心（National Center for Science Information Systems，NACSIS）组织的 NTCIR（NACSIS Test Collection for IR Systems）评测计划与美国纽约大学组织的 IREX（Information Retrieval and Extraction Exercise）计划等。在中文信息检索评测领域，北大天网实验室从 2003 年开始组织 SEWM 中文网络信息检索评测[①]，而清华-搜狐搜索技术联合实验室 2008 年推出的 SogouT 互联网语料库[②]，则是目前研究领域最大规模的中文互联网语料库，包含了 1.3 亿中文网页，10 000 个中文互联网真实用户查询和对应的正确答案。

① 2009 年度评测会议的主页是 http://sewm2009.dlut.edu.cn。

② 数据资源介绍参见：http://www.sogou.com/labs/dl/t.html。

由此可见，使用 Cranfield 评价体系进行检索系统的性能评价已有近半个世纪的历史，该评价体系经历了众多研究项目的考验，也体现出了较强的生命力。借助这一评价体系进行搜索引擎性能评价，是一个合理和可行的方案。

通过对 Cranfield 体系施行过程的描述，可以看出，使用 Cranfield 体系对信息检索系统进行性能评价需要以下 4 个必要的因素：

- 语料库集合；
- 查询样例集合；
- 正确答案集合；
- 评价指标。

其中，语料库集合是指与信息检索系统应用目标相一致的语料数据集合，对于搜索引擎系统而言，语料库集合就是指万维网数据的全体。与传统语料库集合是整理完善的语料内容不同，万维网数据内容繁杂，变化迅速，获取、更新都有一定的难度，但面对同样的万维网数据集合，如何合理利用有限的存储、计算、带宽资源满足用户需求，则是不同搜索引擎各显神通的舞台。因此，对于搜索引擎性能评价任务而言，作为评价者的我们不需准备专门的语料库集合，搜索引擎需要利用网页抓取子系统自行获取万维网数据。

除语料库集合外，在使用 Cranfield 评价体系评价搜索引擎性能时，查询样例集合、正确答案集合和评价指标都是需要着手构建的内容。它们的构建方式将在本章的随后几节中逐一介绍。

## 3.2 查询样例集合构建

查询样例集合是指使用 Cranfield 评价体系评价搜索引擎性能时，模拟用户实际需求向搜索引擎系统提出的查询(query)集合。查询样例集合构建中需要遵循的 3 个重要原则是：集合构建的真实性、代表性和信息需求表述的完整性。

### 3.2.1 查询样例集合构建中的真实性

真实性，是指构建查询样例集合时需要采用真实的搜索引擎用户查询，也就是那些能够反映普遍用户的真实信息需求(而不是出于某种目的专门设计的特定信息需求)的查询。采用真实的用户查询，对于保证搜索引擎性能评价的结果与普通用户的真实使用感受吻合是至关重要的。如果仅采用个别评测人员依照自己的知识背景、兴趣爱好而设计出的查询，则查询样例集合乃至最后的评价结果就缺乏了客观性和公正性。

国内外组织的各项网络信息检索系统评测中，都十分重视查询样例集合的真实性，如文本信息检索会议(TREC)的 Web 检索任务(2002—2004，2009)、Terabyte 检索任务(2004—2006)、Million Query 检索任务(2007—2009)均采用了微软 Live 或 Yahoo! 等搜索引擎提供的真实查询样例(如图 3.3 所示)。而北京大学网络中心组织的搜索引擎与网络信息挖掘(SEWM)评测也采用了来自天网搜索引擎(http://www.sowang.com/beidatianwang.htm)的真实查询样例。

wt09-1：obama family tree
wt09-2：french lick resort and casino
wt09-3：getting organized
wt09-4：toilet
wt09-5：mitchell college
wt09-6：kcs
wt09-7：air travel information
wt09-8：appraisals

图 3.3　TREC 会议采用的部分查询样例

采自 2009 年度 Web 检索任务，任务组织者声称“Topics for this task will be developed from information extracted from the logs of a commercial Web search engine.”，即“本任务所涉及的查询样例采集自某商业搜索引擎的查询日志”。

对于绝大多数搜索引擎领域的研究人员而言，获取到真实的搜索引擎用户日志是非常困难甚至完全不可能的，这就使得如何构建具有真实性的查询样例集合变得十分困难。为了解决这一问题，一方面可以通过某些搜索引擎公开发布的用户行为日志样例[①]来获取真实查询；另一方面也可以借助不少搜索引擎提供的热门查询展示服务来获取真实的用户查询。

如图 3.4 所示的“搜索风云榜”业务，可以使我们方便地得到用户提交给搜索引擎的查询中最热门的部分，如 2009 年 7 月 21 日的热点就包括将于次日发生的“日全食”，以及当时比较热门的一些网站查询如“开心网”、“校内网”等。与百度“搜索风云榜”类似的业务还包

图 3.4　搜索引擎发布的热门查询（采集自百度“搜索风云榜”）

① 如搜狐公司搜狗实验室发布的 SogouQ 查询日志集合 http://www.sogou.com/labs/dl/q.html，该集合包括约 2008 年 2 月份 1 个月内的 Sogou 搜索引擎部分网页查询需求及用户点击情况的网页查询日志数据集合。为进行中文搜索引擎用户行为分析的研究者提供基准研究语料。

括“搜狗热搜榜”(http://top.sogou.com),“雅虎排行榜”(http://top.cn.yahoo.com)等。通过这些业务,我们可以获得真实的热门用户查询。但这就引出了查询样例集合构建中的第二个问题:这些查询即使是真实的,但它们能否在用户群体中具有足够的代表性呢?

### 3.2.2 查询样例集合构建中的代表性

代表性,是指构建出的查询样例集合要能够反映出搜索引擎用户群体的查询偏好,而不能只反映少数用户的需求。由于人力、物力资源的限制,能够构建的查询样例集合规模往往不能太大(这与用户目标页面集合构建的难易程度息息相关,将在 3.3 节中详述)。如何用少量的查询样例集合代表大多数用户的查询偏好,是亟待解决的问题。为了解决这一问题,需要我们对搜索引擎用户查询频率的分布情况进行一个较深入的探讨。

搜索引擎每日需要处理的用户查询数目十分庞大,根据国际上比较著名的搜索引擎技术网站 SearchEngineWatch.com 的统计,2003 年时谷歌每日处理的查询请求数就达到了 2.5 亿个。我们也尝试对某中文搜索引擎网站一个月内的部分查询日志进行了分析,分析的结果表明,这部分查询日志的查询请求数达到了 10 多亿个。然而,这些数量庞大的查询请求并非两两不同,而是集中在若干个查询上。如我们分析的搜索引擎查询日志中,“百度”这一查询的查询请求次数就高达 18 万次以上。我们进一步观察查询频率的分布时发现,查询频率最高的一部分查询集中了大多数的用户查询请求。查询日志涉及的独立查询数共有 1500 多万个,其中查询频度最高的 10 000 个查询(仅占独立查询总量的万分之六点五)就集中了超过 56%的用户查询请求。这说明搜索引擎查询频度的分布在很大程度上符合“二八定律”[①](Pareto principle)而不是“长尾定律”[②](long tail)。

用户查询频率的分布规律启示我们:可以使用少量的高频查询样例集合来代表大多数用户的查询请求。这种将性能评价的关注重点集中在高频查询词上的查询样例集合的构建方法能够保证对用户群体的代表性,从而反映搜索引擎性能的主要情况。由于 3.2.1 节中所提到的搜索引擎“搜索风云榜”业务的存在,这种构建方法也可以十分方便地获取热门(即高频)查询样例数据,这使得其既具有理论层面的合理性,也具有操作层面的可行性。

当然,这种查询样例集合的构建方式也并非尽善尽美。尽管查询频度最高的一部分查询代表了搜索引擎用户的大部分查询需求,但其只能占到搜索引擎所要处理的独立用户查询中的很少一部分(像上面举的例子中,查询频度最高的 10 000 个查询只占独立查询总量的万分之六点五),这使得该查询样例集合在构建过程中较少顾及占独立查询总量中绝大部分的查询频度较低的用户查询。然而,一方面,如果要提高对查询频度较低的用户查询的代表性,则查询样例集合的规模必然有较大地增加;另一方面,从操作层面而言,获取查询频度较低的用户查询数据也有较大的难度。如何在构建查询样例集合的过程中既控制好样例集合的规模,又保证各种查询频度的用户查询数据能够被采样到,这也是搜索引擎研究领域重点关注的研究方向之一。在这个问题没有得到很好地解决之前,使用查询频度最高的一部

① 二八定律是指这样一种现象:在一组事物中,最重要的只占其中一小部分,约 20%,其余 80%的尽管是多数,却是次要的。

② 长尾定律最初是指一种商业现象,即不受到重视的销量小,种类多的产品或服务由于总量巨大,累积起来的总收益超过主流产品的现象。目前也被广泛应用于互联网产品的设计与分析。

分查询作为查询样例集合是一个既合理又实际可行的操作方式。

### 3.2.3　查询样例集合构建中信息需求表述的完整性

相比真实性和代表性而言，信息需求表述的完整性相对更容易被人忽略。信息需求表述的完整性是与当前搜索引擎的用户交互方式密切相关的，当前的搜索引擎交互方式主要是以谷歌为代表的“关键词查询＋选择性浏览”的交互方式。这种交互方式的特点是：用户用简单的关键词作为查询提交给搜索引擎，搜索引擎并非直接把检索目标页面反馈给用户，而是反馈给用户一个可能的检索目标页面列表，用户浏览该列表并从中选择出能够满足信息需求的内容加以浏览。

“关键词查询＋选择性浏览”的交互方式可以说是与互联网用户的行为习惯和搜索引擎当前的技术发展水平相适应的。一方面，互联网用户希望减少键盘输入，而尽可能只通过最简单的方式（鼠标点击）与计算机进行交互[①]，因而只希望输入最简单的关键词；另一方面，搜索引擎难以通过简单的关键词准确的理解用户的查询意图，因此只能将有可能满足用户需求的结果集合以列表的形式返回给用户，而无法提供给用户准确的检索目标。搜索引擎发展的历史上，也有不少试图改变这种交互方式的尝试，如 Konopnicki(1995) 的 W3QL 系统，Mendelzon(1997)的 WebSQL 系统等通过制定专门查询语言的方式来确保搜索引擎对用户需求的准确理解，但这些努力却往往使搜索引擎系统本身的使用变得过于复杂而难以普及。根据 Silverstein(1999)和余慧佳(2007)的用户行为调查工作，近 80％的英文用户和 95％以上的中文用户甚至根本不使用当前搜索引擎提供的十分有限的“高级查询”功能，因此，期望大多数用户去学习使用专门的查询语言进行信息获取是不现实的要求。

在“关键词查询＋选择性浏览”的交互方式下，搜索引擎面对的查询是十分简单的字词组合。根据我们对某中文搜索引擎 2009 年 5 月的用户查询进行的分词分析，查询所包含的平均词数为 3.11 个，简短的 3 个多词要承载十分复杂的用户信息需求，这确实造成了不少查询信息需求表述不清的现象。

图 3.5 是我们从搜索引擎用户日志中采集到的同样查询“魔兽争霸”的三个不同用户的点击行为序列，通过观察这三个行为序列，可以尝试归纳出用户信息需求的不同类型。

图 3.5 标示的三个用户的行为序列，能够在一定程度上反映出此三个用户在进行“魔兽争霸”查询时不同的信息需求。

用户 A 只点击了排名第 1 位的题为“魔兽争霸官方资讯网”的结果就结束了查询，很有可能是因为其查询目标即为这个结果，用户 A 之所以进行查询，只是因为不记得这个结果页面的 URL，而借助于搜索引擎定位这个自己已知（或以前访问过）的网络资源。以用户 A 为代表的这种信息需求，一般被称为“导航类”(Navigational)信息需求。

用户 B 的行为与用户 A 大相径庭，他首先点击了若干个结果（排名分别为第 1、2、5 位），随后更换查询为“魔兽争霸下载”，点击这个查询对应的某个结果后结束了查询。如果我们仔细观察用户在“魔兽争霸”结果页面里点击的三个结果（第 1、2、5 位检索结果），我们会发现，这三个结果也都在标题和摘要中注明了“魔兽争霸下载”或“魔兽争霸 3 下载”的字样。这说明：用户尽管输入的关键词只有“魔兽争霸”，但其原本的信息需求就是下载魔

① 由于相当一部分的中老年网民尚不熟悉打字输入，所以这种情况在我国的网民群体中更为普遍。

| 点击次序 | 被点击结果的排序 | 结果标题 | 结果 URL |
|---|---|---|---|
| 1 | 1 | 魔兽争霸官方资讯网|魔兽争霸 1.20D|冰封王座|魔兽争霸下载 | war.aomeisoft.com |
| 结束查询 | | | |

(a) 用户 A

| 点击次序 | 被点击结果的排序 | 结果标题 | 结果 URL |
|---|---|---|---|
| 1 | 1 | 魔兽争霸官方资讯网|魔兽争霸 1.20D|冰封王座|魔兽争霸下载 | war.aomeisoft.com |
| 2 | 2 | 魔兽争霸 3:冰封王座_魔兽 1.23_魔兽争霸 3 秘籍_魔兽争霸 3 下载 | fight.pcgames.com.cn/warcraft |
| 3 | 5 | 【魔兽争霸 3:冰封王座】单机游戏下载 | youxi.zol.com.cn/...x3767.html |
| 更换查询:魔兽争霸下载 | | | |
| 4 | | 魔兽争霸Ⅲ下载 v1.20E 冰封王座中文绿色版 单机游戏 | www.52z.com/soft/2330.html |
| 结束查询 | | | |

(b) 用户 B

| 点击次序 | 被点击结果的排序 | 结果标题 | 结果 URL |
|---|---|---|---|
| 1 | 1 | 魔兽争霸官方资讯网|魔兽争霸 1.20D|冰封王座|魔兽争霸下载 | war.aomeisoft.com |
| 2 | 2 | 魔兽争霸 3:冰封王座_魔兽 1.23_魔兽争霸 3 秘籍_魔兽争霸 3 下载 | fight.pcgames.com.cn/warcraft |
| 3 | 3 | 魔兽争霸 3:冰封王座_dota_魔兽地图_单机游戏_新浪游戏 | games.sina.com.cn/z/war3 |
| 4 | 6 | 魔兽争霸 3,魔兽争霸秘籍,魔兽争霸下载,魔兽争霸地图 | wcg.yesky.com/war3 |
| 5 | 8 | 浩方对战平台-全球最大的电子竞技平台之一-魔兽争霸子站 | war3.cga.com.cn |
| 6 | 9 | 魔兽争霸 3 冰封王座专题站 攻略 下载 心得 地图 战报 录像 | games.enet.com.cn/...nti/war3 |
| 结束查询 | | | |

(c) 用户 C

图 3.5 不同用户进行“魔兽争霸”查询时与搜索引擎的交互情况比较

兽争霸的游戏。最终,也通过修改查询为“魔兽争霸下载”而进一步表明了自己的信息需求。以用户 B 为代表的这类用户进行搜索引擎查询的主要目的是寻找某种类型的网络资源(音乐、视频、软件、数据库等),进而通过下载、查询等进一步的交互方式获取该资源。这种信息需求的类型一般被称为“事务类”(Transactional)信息需求。

用户 C 的行为与用户 A、B 也有差异，他进行查询后，先后点击了多个检索结果（排名分别为第 1、2、3、6、8、9 位），随后结束了查询。该用户点击的检索结果，均为各大权威资讯平台关于魔兽争霸游戏的站点（魔兽争霸官方资讯站点、新浪游戏魔兽争霸频道、太平洋游戏魔兽争霸频道、天极游戏魔兽争霸频道等）。此用户有很大的可能是想通过搜索引擎获取关于“魔兽争霸”游戏相关的资讯信息，因此其点击的结果都是权威性较高的介绍“魔兽争霸”游戏信息的网站。以用户 C 为代表的这类用户主要的目的是获取与某个主题相关的信息，这类需求一般被称为“信息类”(Informational)信息需求。

同样的一个查询请求“魔兽争霸”，不同用户的信息需求却截然不同，这说明在现有的“关键词查询＋选择性浏览”的用户交互方式下，查询词并无法很准确地描述用户的信息需求。这为搜索引擎处理用户查询造成了很大的困难，也为构建查询样例集合的过程设置了障碍。如果没有确定用户的信息需求，则无法准确地确定用户的查询目标页面，例如对于“魔兽争霸”的“导航类”信息需求而言，仅有排名第 1 位的结果“魔兽争霸官方资讯网”才是目标页面；而对于“信息类”信息需求而言，结果列表中的多个结果都可以被认为是目标页面。

针对这一问题，可以采用如下 3 种方式明确用户信息需求类别，进而加以解决。

(1) 可以尽量选取信息需求描述得比较明确的用户查询。如：“清华大学本科招生网”是明确的“导航类”信息需求，“潜伏在线观看”是明确的“事务类”信息需求，而“手足口病症状”则是“信息类”信息需求。当然，这一选取查询的过程需要在满足“真实性”和“代表性”的前提下进行。

(2) 可以对查询样例集合的信息需求类别加以规定。如 TREC 组织的 Home Page Finding(主页查找)、Named Page Finding(命名网页查找)子任务就是专门针对“导航类”信息需求设定的评测任务；而 Topic Distillation(主题提取)子任务则是针对“信息类”需求而设定的评测任务。国内的 SEWM 评测也沿袭了这一评测体系，将评测任务分为主题提取和导航搜索两个子任务，如图 3.6 就列出了 SEWM 评测中的一部分查询样例。

```
<top>
<num>Number:NP890
<title>中国政法大学学工部</title>
</top>

<top>
<num>Number:NP891
<title>中国电子工业标准化技术协会网</title>
</top>

<top>
<num>Number:NP892
<title>国道 111 改建工程招标公告</title>
</top>
```

图 3.6　北京大学网络实验室组织的 SEWM 评测所采用的评测样例

（采集自 2007 年该评测对应的命名网页与主页查找任务）