基于用户行为分析的搜索引擎自动性能评价*

刘奕群[1+], 岑荣伟[2], 张敏[3], 茹立云[4], 马少平[5]

[1, 2, 3, 5](清华大学 智能技术与系统国家重点实验室,北京 100084)

[4](搜狐公司研发中心,北京 100084)

Automatic Search Engine Evaluation Based On User Behavior Analysis*

LIU Yi-qun[1+], CEN Rongwei[2], ZHANG Min[3], RU Liyun[4], MA Shao-ping[5]

[1, 2, 3,5](State Key Lab of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China)

[4](Sohu Inc. Research and Development Center, Beijing 100084, China)

+ Corresponding author: Phn: +86-10-62777702, Fax: +86-10-62771138, E-mail: liuyiqun03@mails.tsinghua.edu.cn. , http://www.csai.tsinghua.edu.cn

Abstract: Performance evaluation is an important issue in Web search engine researches. Traditional evaluation methods rely on much human efforts and are therefore quite time-consuming. With click-through data analysis, we proposed an automatic search engine performance evaluation method. This method generates navigational type query topics and answers automatically based on search users' querying and clicking behavior. Experimental results based on a commercial Chinese search engine's user logs show that the automatically method gets a similar evaluation result with traditional assessor-based ones.

Key words: Web Information Retrieval; Performance Evaluation; User Behavior Analysis.

摘 要: 性能评价一直是网络信息检索研究中的核心课题之一。传统的评价方式需要花费大量的人力物力，时间效率也较低。基于用户行为分析的思路，提出了一种自动进行搜索引擎性能评价的方法，此方法能够自动生成导航类查询测试集合并对查询对应的标准答案实现自动标注。实验结果证明，此方法能与人工标注的评价取得基本一致的评价效果，同时大大减少了评价所需的人力、物力资源，并加快了评价反馈周期。

关键词: 网络信息检索;性能评价;用户行为分析

中图法分类号: TP391 TP393 **文献标识码**: A

1 引言

检索系统的评价问题一直是信息检索研究中的最核心问题之一，Saracevic[1]指出：“评价问题在信息检索研发过程中处于如此重要的地位，以致于任何一种新方法与他们的评价方式是融为一体的”。Kent首先提出了精确率—召回率的信息检索评价框架（根据[1]），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

* 得到国家重点基础研究（973）(2004CB318108)、自然科学基金（60621062, 60503064）和 863 高科技项目（2006AA01Z141）资助.

作者简介: 刘奕群（1981－），男, 山东济南人, 博士研究生, 主要研究方向是信息检索,机器学习; 岑荣伟（1982 －），男, 浙江慈溪人, 博士研究生, 主要研究方向是信息检索,机器学习; 张敏(1977－),女,博士,讲师,主要研究方向为机器学习, 信息检索. 马少平(1961－),男,教授,博士生导师,主要研究领域为知识工程,信息检索,汉字识别与后处理以及中文古籍数字化.

方面的研究，而英国Cranfield工程在上世纪五十年代末到六十年代中期所建立的基于查询样例集、标准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[1]，其评价框架一般被称为Cranfield方法（A Cranfield-like approach）。

Cranfield方法指出，信息检索系统的评价应由如下几个环节组成：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，即进行标准答案集合的标注；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，再利用检索评价指标对检索结果和标准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的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。由美国国防部高等研究计划署(Defense Advanced Research Projects Agency，简称DARPA)与美国国家标准和技术局(National Institute of Standards and Technology，简称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信息检索评测和技术交流论坛。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开始尝试运作，如NTCIR (NACSIS Test Collection for IR Systems)计划与IREX (Information Retrieval and Extraction Exercise)计划等。

随着万维网的不断发展与互联网信息量的增加，如何评价网络信息检索系统的性能逐渐成为近年信息检索评价中的热点关注方向，而进行这方面评价时， Cranfield方法遇到了巨大的障碍。困难主要反映在针对查询样例集合的标准答案标注上，根据Voorhees[2]的估计，对一个规模为800万文档的语料库进行某个查询样例的标准答案标注需要耗费9个评测人员一个月的工作时间。尽管Voorhees提出了诸如Pooling[2]这样的标注方法来缓解标注压力，但当前针对海量规模网络文档的答案标注仍是十分困难的。如TREC海量规模检索任务（Terabyte Track）一般需要耗费十余名标注人员2-3个月的时间进行约几十个查询样例的标注，而其语料库数据规模不过1000万文档左右。考虑到当前搜索引擎涉及到的索引页面都在几十亿页面以上（Yahoo!报告为192亿网页，中文方面Sogou声称的索引量也超过百亿），利用手工标注答案的方式进行网络信息检索系统的评价会是一个既耗费人力、又耗费时间的过程。由于搜索引擎算法改进、运营维护的需要，检索效果评价反馈时间需要尽量缩短，因此提高搜索引擎性能评价的自动化水平是当前检索系统评价研究中的热点。

本文按照如下方式组织：第二部分讨论相关研究工作,阐明搜索引擎自动评价方面的已有工作成果和问题,第三部分简要介绍查询信息需求与搜索引擎评价之间的关系,第四部分对搜索引擎自动评价算法进行推导,并说明利用这个算法进行导航类查询自动评价的具体操作,第五部分给出标准答案标注实验和性能评价实验结果,最后总结并列出主要结论.

2 相关研究工作概述

为了解决 Cranfield 方法在网络信息检索系统评价中所面临的困境，不少研究人员提出了一些自动进行搜索引擎性能评估的方案，其工作集中在两个方面：基于 Cranfield 框架，只是使用自动化方法进行答案自动标注；或采用不同于 Cranfield 方法的评价框架进行自动化评价。

前一方面的研究工作中，研究者尝试使用检索系统反馈的结果信息进行自动标注。Soboroff[3]在基于TREC实验平台的研究中发现：评价人员对于结果池内文档的标注结果差异基本不影响检索系统性能排序的结果，因而随机挑选结果池内文档作为标准答案也有可能达到评价检索系统性能的作用。他因而提出可以在检索系统结果池中，随机挑选一定数量的结果作为答案集合进行评价。实验效果证明，按这种方式实现的检索系统评价结果与基于手工标注集合的评价结果正相关，但对于检索系统性能排序的影响较大因而难以投入使用。Nuray[4]提出对Soboroff方法的修正方案，即选择结果池中原本在搜索引擎结果序列中排序较前的文档作为标准答案，他们的方法也没有取得与手工评价方法相类似的评价结果。

我们认为，这类基于搜索引擎结果反馈信息（伪相关反馈信息）进行搜索引擎评价的尝试很难获得成功。这是由于伪相关反馈信息本身就是一种不可靠的信息源，它只能对搜索引擎处理性能较高的查询进行正确的结果标注，而事实上由于针对这部分查询的评价不会对搜索引擎的性能提高起到指导作用，因此很少需要对

其进行性能评价。这就形成了需要进行评价的查询标注不好，不需要进行评价的查询反而标注的较好的情况，因此这种自动标注的思路很难应用于实际搜索引擎评价中。

也有部分研究人员基于已有的网页目录资源进行结果的自动标注，如Chowdhury[5]和Beitzel[6]提出的利用开放目录计划（ODP计划）所整理的网页目录和对应的网页摘要资源进行性能评测的工作。其方法的优势在于答案标注的正确性比较单纯使用搜索引擎结果反馈信息更高，但使用网页对应的摘要信息作为用户查询的模拟还是一个不合理的假设，因而其工作也没有得到大规模的普及应用。

第二方面的研究工作中，比较有代表性的有 IBM Haifa 研究院研发的“相关词集合评价方法”与 Joachims 提出的基于用户点击行为的评价方法等。

Amitay[7]提出了“相关词集合评价方法”（*Term Relevance Sets,* 简称Trels方法）。方法首先选择一定量的代表用户查询需求的查询词；随后针对每一个查询词，手工标注尽量多的与此查询词相关联的词项；施行评价时，通过待评测文档中关联词项的分布情况判定文档的相关程度及检索结果的可靠性。这种方法将大量手工工作从收集检索结果的过程之后转移到收集结果之前，作者也认为其标注的关联词项能够较长时间发挥稳定的评价作用。Trels方法一定程度上解决了评价结果反馈时间过长的问题，但丝毫没有减少甚至增加了相关性标注的难度。同时，词与词的相关程度本身就是一个难以界定的问题。作者基于TREC小规模数据的实验取得了一定的效果，但并没有将之使用在大规模的网络信息检索系统评价中。

Joachims[8]第一次提出了使用用户点击行为信息评价搜索引擎性能的思路。他设计了一个元搜索引擎，用户输入查询词后，将查询词在几个著名搜索引擎中的查询结果随机混合反馈给用户，并收集随后用户的结果点击行为信息。根据用户不同的点击倾向性，就可以判断搜索引擎返回结果的优劣，Joachims同时证明了这种评价方法与传统Cranfield方法评价结果具有较高的相关性。由于记录用户选择检索结果的行为是一个不耗费人力的过程，因此可以避免传统Cranfield方法反馈过慢的问题。但这之前，必须首先评判用户点击行为的可靠性，即用户的点击是否意味着其认为被点击的结果与查询相关。Joachims在这方面并没有给出一个完善的解决方案，其随机混合答案的方式尽管避免了所谓的“排序偏置”（即减少用户因为结果排列在前面就点击它的可能性），但也与用户正常使用搜索引擎的体验产生差异，因此收集到的用户行为可信程度降低；同时，使用这个元搜索引擎本身并无法为用户带来更加快捷方便的搜索体验，因此其必然无法吸引足够多的用户提供点击信息，进而影响到评价结果的可信程度。

综上所述，研究人员基于 Cranfield 框架进行了自动结果标注的尝试，但由于选择的标注方式不可靠而没有获得成功；在 Cranfield 框架之外进行的各种尝试，尽管自动化程度都较高，但其评价方法的可靠性问题还有待商榷。我们认为，Cranfield 的检索系统评价方式是经过相当程度的理论和实践检验，因而在其面临搜索引擎评价的困境时将其抛弃是不明智的选择。而发展 Joachims 的用户点击行为分析方法，将其扩展到查询样例集合的结果自动标注过程中，是一个可行的解决方案。

3 查询信息需求与自动性能评价

上一节，我们对搜索引擎自动评价的研究成果进行了综述，并提出了使用用户点击行为分析的方法进行答案自动标注的问题。这个想法的出发点在于：由于现有的绝大多数搜索引擎用户还是能够通过搜索引擎找到满足其查询需求的答案（尽管可能需要花费较多的精力），因此用户的点击行为中肯定蕴含了其对检索结果相关性的评价。

从个体用户的行为上讲，有可能由于个人知识水平、网络使用习惯的不同而点击某些与查询需求无关的页面，甚至有可能被垃圾页面，SEO 页面等所欺骗；但从用户群体的宏观行为规律上讲，这些无关点击可以被认为是随机噪声而滤除掉。因而当用户群体足够大，收集到的点击信息足够完善时，点击信息的可靠程度还是能够得到一定的保证。

对于搜索引擎而言，其网络服务供应商的身份同时也为其收集了海量规模的用户日志信息。在我们之前的工作[9]中，我们利用这部分用户日志信息实现了用户查询信息需求的分类，那么，利用这些信息中蕴含的用户群体点击行为信息实现答案自动标注也是一个自然的解决问题的思路。

然而，用户群体行为的可靠性尽管可以得到保证，但对于性能评价中的答案标注而言，标注出正确的结果并不是唯一需要考虑的问题，是否标注出了所有正确的结果同样值得考虑，这就需要具体考虑用户查询信息需求的问题。

Broder（2002）指出，用户的查询信息需求包括以下三类：

导航类（Navigational）：目标是查找某个特定的站点或者网页。如“上海市政府网站”、“清华大学招生简章”等（摘自百度网站“搜索风向标”栏目，下同）。

信息类（Informational）：目标是获取可能位于一个或某几个网页上的信息。如“现代企业制度的形式”、“农村党员队伍状况”等。

事务类（Transactional）：目标是查找能够处理某些以 Web 为媒介的事务的网页。如“连连看下载”、“歌词查询”等。

对查询信息需求进行划分的出发点在于，针对三类检索可以使用不同的检索模型、参数，甚至评价方法也随着检索类别的变化而有区别。因此实现检索类别的自动划分对于提高检索性能和增加检索评价的可信度都有非常重要的意义。

对于导航类查询而言，其正确答案唯一，因而不需考虑答案全面性的问题；其对应的搜索引擎检索性能也较高，因此用户点击行为的可靠性也比较容易保证。即：用户在进行导航类查询时，较容易发现并点击结果列表中对应的答案，因而我们所进行的主要工作，只是将用户点击行为中反映出的答案挑选出来。对于信息或者事务类查询（统称信息事务类查询）而言，情况则要复杂的多，其正确答案不唯一，因此必须考虑答案全面性的问题；而其对应的搜索引擎检索性能相对较低，用户能否点击到即使是正确的答案也较难保证。

为了考察用户点击行为是否适用于进行信息事务类查询的答案标注，我们考察了提交查询词“电影”的四个常用中文搜索引擎（百度，谷歌，雅虎，搜狗）用户在 2006 年 12 月 10 日的点击情况，如下图所示：

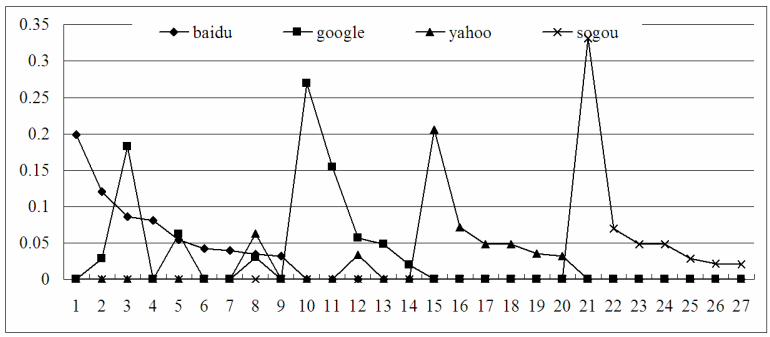

Figure 1 Differences in click-through behavior of four Chinese search engines using keyword “movie”

图 1 针对查询词“电影”的不同搜索引擎用户点击情况

实验收集了四个搜索引擎针对查询词返回的前 10 位结果，取并集后共 27 个结果，图 1 中的横轴对应这 27 个结果，而曲线上的点则是结果对应的不同搜索引擎的用户点击频度信息。如第 21 号结果对应的搜狗搜索引擎曲线（用“×”表示）上的数值约为 34%，即代表第 21 号结果在搜狗搜索引擎上被 34%的查询“电影”的用户所点击。本实验数据的获得是通过搜狗公司采集的用户搜索反馈信息，共涉及了近 200 名用户的搜索引擎访问信息。

图中，我们可以发现，不同搜索引擎用户针对这个查询的点击情况差异非常大，如百度用户的点击多集中在第 1 号结果上，而谷歌用户点击第 3 号和第 10 号的最多；各个搜索引擎结果尽管有一定交集，如第 3、5、8 号结果均被多个搜索引擎用户所关注，但关注程度却有较大差异。

尽管“电影”这个查询词仅仅是信息事务类查询的一个简单样例，但它可以反映出这种类型的查询需求对应的检索结果反馈现象：提交同一个信息事务类查询需求时，用户在不同搜索引擎上得到的结果是不同的。这种差异既来源于搜索引擎的页面索引差异（即不同搜索引擎索引到的页面集合不同），也来源于搜索引擎的结果排序策略差异，因而对于查询目标页面不唯一的信息事务类查询是难以避免的。

这说明，对于信息事务类而言，用户期望的正确答案可能有多个，但某单个搜索引擎则很难反馈全所有的结果，因此使用某个搜索引擎的用户行为信息去评价其他搜索引擎信息事务类查询的性能是不合理的。

对于研究人员而言，获取多家搜索引擎的用户日志有较高的难度，对于搜索引擎自身来讲，获取其它供应商的日志更是难上加难，因此在现有的实验环境和商业运行模式下，实现信息事务类查询的自动评价可能是不现实的选择。

4 导航类查询的自动性能评价算法设计

上节的论述中，我们明确了在当前的实际应用条件限制下，搜索引擎性能自动评价的对象只能限制于导航类检索，因此本节我们来讨论导航类自动性能评价系统的算法设计。依照 Cranfield 方法框架，查询样例集合、标准答案集合和语料库是性能评价必备的三要素，对于网络信息检索系统而言，Web 数据集合即其面对的语料对象，因此实现查询样例集合和标准答案集合的自动生成，就成了我们所主要关心的问题，包括这两个环节在内的搜索引擎自动评价方法的整体运行流程如下图所示。

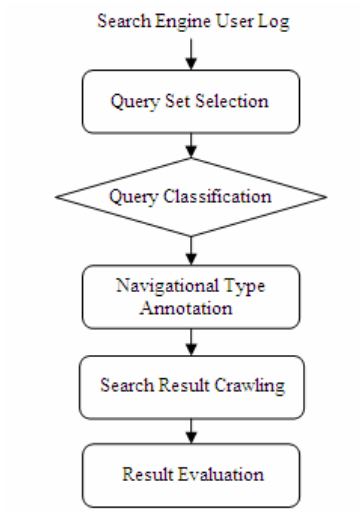

Figure 2 Automatic Search Engine Performance Evaluation

图 2 搜索引擎自动评测方法流程

搜索引擎日志首先经过数据预处理，获得必需的用户点击行为特征，随后进行查询样例集合的自动选取，并依据第三章所述的搜索引擎用户查询信息需求分类方法进行查询需求分类，其中的导航类需求被挑选进行自动标准答案标注，此后进行搜索引擎结果的抓取和性能评价指标的计算。

在上述评测方法流程中，搜索引擎结果的抓取与过滤是指将查询样例集合中的样例提交给搜索引擎进行查询，并收集其结果页面，过滤出结果 URL 列表。而搜索引擎的性能评价指标计算则是指根据搜索引擎返回的结果 URL 列表与自动标注出的答案集合，计算性能评价指标的过程。对于导航类查询需求而言，性能评价指标使用“首现正确结果排序倒数”（RR）进行计算。

RR 是指检索系统返回的结果序列中第一个满足用户需求文档出现序号的倒数。RR=1 表示检索系统返回的结果中，第一个结果就可以满足用户需求。这个指标通常用来评价导航类检索的性能，因为这类检索只有一个标准答案可依满足用户需求。

4.1 传统决策树算法处理关键资源判定的优势与困境

构建有合适代表性的查询样例集合对于搜索引擎评价结果的可靠性也是至关重要的。在传统的性能评价研究如 TREC 相关工作中，查询样例集一般是由评测人员专门挑选出的，部分任务的查询主题可能来自于对搜索引擎日志的筛选，但大部分是专门设计的用于评测系统性能的查询。此外，由于手工标注工作量的限制，

因此查询样例集合的规模一般较小，每单个 TREC 检索任务的查询样例集合约包括几十到一、二百查询不等。

由于我们所进行的是自动性能评测系统的查询样例集合设计，因此可以较少考虑人工标注所导致的查询数量限制，因此我们重点考察查询样例集合的代表性问题，即多大规模的样例集合足够代表搜索引擎用户的实际查询情况。为此，我们对 Sogou 搜索引擎 2006 年 2 月全月的用户日志集合进行了查询频度分析，分析结果如图 3 所示。

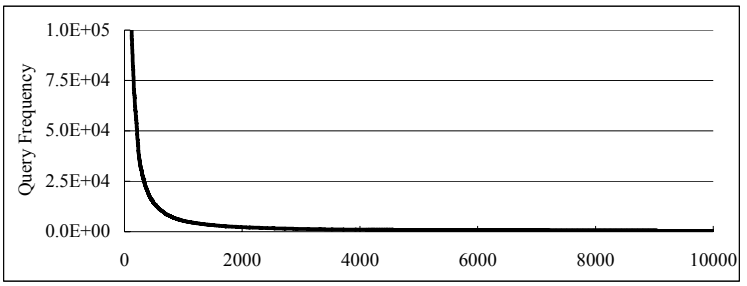

Figure 3 Query Frequency Distribution in Sogou Search Engine Log

图 3 查询日志中的查询频度分布情况

图 3 中，我们选择了查询频度最高的 10000 个查询词，并观察其频度的分布情况。图中的横坐标为按频度进行排序的序号，而纵坐标为对应排序的查询的查询频度多少。从图中，我们可以发现频度绝对数值随排序增加下降的非常迅速，这意味着少数查询即可能代表相当大一部分的用户查询需求。根据统计，此查询词集合中频度高于 100 的查询仅有 35177 个，占查询总数目的不足 1%，但此 1%的查询却覆盖了 69%的用户查询需求。这说明使用一个较小规模的查询样例集合代表搜索引擎大部分用户的信息需求是完全可行的。

尽管标准答案集合的标注将自动完成，但由于搜索引擎结果抓取速度受到网速、搜索引擎服务策略等多方面的限制，因此查询样例集合的整体规模不宜过大。考虑到实际施行难易程度和用户需求代表性两方面因素，我们认为选择约 10000-15000 查询词作为查询样例集合较为合适。这个规模的样例集合能够代表相当大比例的用户需求（一个月用户需求总数的约 50%），处理时间也可以接受（当程序运行硬件环境为 1.8G 主频 CPU、1G 内存与 100MLAN 网络时，约需 1 天时间完成性能评估）。

因此，选取一段时间内搜索引擎查询频度最高的一定数量查询，是我们构建查询样例集合的核心策略。

4.2 连续属性值的离散化

比较起查询样例集合的自动生成而言，标准答案集合的标注是更为困难的研究课题。在第 1 节的论述中，我们也指出这个标注过程是 Cranfield 方法在评价搜索引擎性能时面临的最大困难所在。上节中我们将查询样例集合的规模控制在 10000-15000 查询左右，这个规模的样例集合对于手工标注答案而言是不可完成的任务，因此构建快捷准确的答案标注算法势在必行。

Lee[10]首先给出了点击集中度的定义，对于某个查询Q，定义R_{most}为查询Q的搜索引擎用户点击的最多的一个结果，而点击集中度则为对应R_{most}的点击数与针对Q的总点击数的比例值。Lee提出点击集中度的概念，更多的是从查询信息需求分类的角度进行考虑，而我们把注意力转向对于R_{most}的考察。

对于导航类查询而言，由于用户的信息需求唯一，因而用户的点击一般会集中在其查询目标页面上，即R_{most}有很大可能性成为查询目标页面。但当搜索引擎无法把查询目标页面反馈在结果序列中较靠前的位置时，由于Silverstein[11]和余慧佳[12]指出的“绝大部分用户只点击第一页搜索结果”的情况，上述推断也可能出现错误。由于搜索引擎针对导航类查询的检索效果较好，在前一、两页结果中没有返回查询目标页面的概率很小，因此这种推断错误的情况应该认为很少发生。

如果定义网页 r 针对查询 Q 的“点击比率”为：

$$\text{点击比率}\ (Q, r) = \frac{\text{查询Q的用户中，点击 } r \text{ 的次数}}{\text{查询Q的用户进行点击的总次 数}} \quad (1)$$

则我们有下式成立：

$$\text{点击比率}\ (Q, R_{most}) = \text{点击集中度}\ (Q) \quad (2)$$

即点击集中度等于点击比率的最大值，而点击比率最大的 r 则很可能成为导航类查询 Q 的查询目标页面。依照上述推断，我们可以设计下面的标准答案自动标注算法：

对待标注的查询样例集合中给定的查询Q 及其被点击过的查询结果$r_1, r_2, \ldots, r_M$：

IF Q 为导航类查询

在$r_1, r_2, \ldots, r_M$中定位R，使之满足点击比率(Q,R) = 点击集中度(Q);

IF 点击集中度$(Q) > T$

标注 R 为 Q 的标准答案;

EXIT;

ELSE

Q 不可被标注;

END IF

ELSE //Q 不是导航类查询

Q 不可被标注;

END IF

在算法过程中设计点击集中度最小阈值 T 的目的，在于避免出现前文提到的搜索引擎无法把查询目标页面反馈在结果序列中较靠前的位置的情况，在这种情况下，无法依靠用户行为日志实现答案自动标注，因此算法需要给出对应的提示信息。

5 实验与结果分析

5.1 实验环境和方法简述

本节针对上述提出的搜索引擎自动性能评价算法，利用实验方法验证其可靠性。实验数据采集自 Sogou 搜索引擎 2006 年 6 月至 2007 年 1 月半年多的查询、点击日志，每日的查询和点击行为数量约为 150 万条。采用海量真实规模搜索引擎数据进行算法有效性的验证，可以充分考察算法的可靠性及施行效率。

如上文所述，实验所采用的硬件平台是一台普通 PC 级别计算机，整体花费约 6000 元人民币，实验在 100M 局域网内进行。算法的时间花销主要集中在待评测搜索引擎结果序列的抓取上，每小时处理的查询个数约为 400 个。比较传统性能评价工作中十几个标注人员工作数月的效率而言，有了质的飞跃性提高。

下面我们将分别从标准答案标注的正确性与性能评价指标的准确性两方面来衡量搜索引擎自动评价算法的性能。

5.2 答案自动标注实验结果

为控制查询样例集合的规模、同时考察算法针对用户行为时间变化的鲁棒性，我们没有整体使用上述日志数据，而是将实验所用的用户行为日志数据按时间段分成三部分，分别施行查询样例提取和标准答案标注。针对每一个部分的答案标注结果，我们随机抽取约 5%的数据进行手工验证，实验结果如下表所示：

Table 1 Automatic Answer Annotation Experimental Results

表 1 答案自动标注实验结果

时间段	查询样例集合规模	手工验证数据规模	精确率
2006.6 - 2006.8	13,902	695	98.13%
2006.9 - 2006.11	13,884	694	97.41%
2006.12 - 2007.1	11,296	565	96.64%

每个时间段导航类查询样例集合的规模都控制在了 10000-15000 个查询，最后一个时间段的时间跨度略短，因此集合规模也略小。而手工验证的实验结果则说明，答案标注的准确度相当高，每个样例集合的标注精确率都超过了 95%，考虑到即使手工标注也很难避免错误，自动答案标注可以说满足了搜索引擎性能自动评价的需要。

我们进一步分析了若干标注错误的样例，发现绝大部分的错误都是如下情况导致的：正确结果应当是某个站点的主页，而自动标注的结果为此站点的某个子站点主页。例如查询“163”对应的标准答案应当是 http://www.163.com，而自动标注的答案是 http://mail.163.com；查询“搜狗”对应的标准答案应当是 http://www.sogou.com，而自动标注的答案是 http://d.sogou.com。

通过进一步分析，我们发现，这种自动标注答案发生错误的情形是由用户点击行为的倾向性造成的。由于绝大部分用户事实上并非想访问其查询词对应的站点主页，而是这个站点最具吸引力的某个子站点首页，因此主页的用户点击集中度反而不及子站点首页高。查询“163”的用户大都点击了其邮箱主页是因为 163 的免费邮箱服务是中文网络环境中最受欢迎的，而搜狗 mp3 搜索比搜狗网页搜索产品质量更高也是一般网络用户的共识。

因此，这种错误的产生事实上并不是自动标注算法的谬误，而是用户真实行为和选择倾向性的体现。从这个角度讲，是否只有标注站点主页才算标注正确，也是值得商榷的问题，但这并非本文讨论的重点，在下文的实验中，我们还是以传统意义上的主页作为查询对应的标准答案。

5.3 性能评价实验结果

依照第 4 节中所述的搜索引擎自动性能评价方法，即可以自动生成性能评价所需的查询样例集合和标准答案集合，并使用 Cranfield 方法对搜索引擎的处理导航类查询需求的性能进行评价。

我们选用了搜狗搜索引擎之外的五个中文搜索引擎作为性能评价的对象，它们是：百度、谷歌、雅虎中文、新浪爱问搜索与中国搜索。之所以没有选择提供用户行为日志的搜狗搜索作为被评价的目标，是考虑到用户行为日志的提供者有可能在评价的过程中被给予不应有的偏向。

评价指标方面，我们选择了“首现正确结果排序倒数”（参见第 4 节）指标进行性能计算与比较。而为了验证评价实验效果的正确性，我们也建立了一个手工评价集合与自动评价结果进行比较。

手工评价的查询样例集合由评价人员从查询日志中手工挑选的 320 个导航类查询词组成，查询样例集合被投入上述五个搜索引擎进行查询并收集结果，将结果取并组成结果池，并使用 pooling 方法由评价人员进行了答案标注，以此获得了查询样例对应的标准答案集合和进一步的搜索引擎评价结果。

具体的评价实验结果如下图所示：

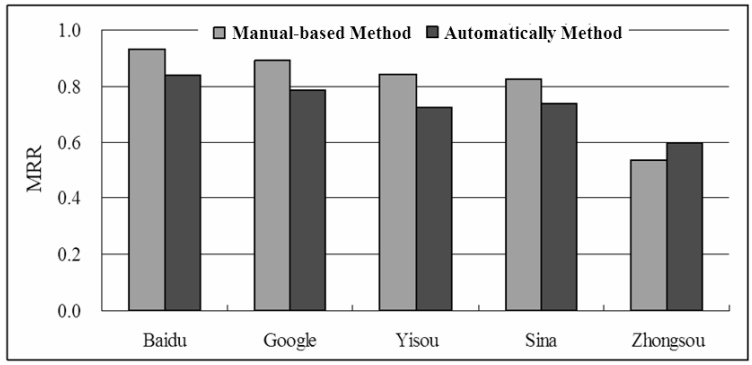

Figure 4 Comparison between automatic and mannual performance evaluation results

图 4 搜索引擎自动性能评价结果与手工评价结果的比较

图 4 介绍了搜索引擎自动评价的实验结果，并将其与手工评价结果进行了比较。由图可见，自动评价的搜索引擎性能排序结果与手工评价的结果完全相同，两种方法给出的 MRR 值之间的相关系数为 0.965，这意

味着两种方法给出的评价结果非常相似。

从图中我们可以发现，尽管从总体性能评价排序而言，手工方法与自动评价方法保持一致，但两种评价方式对应的 MRR 绝对数值则有一定差别。这是由所采用的查询样例集合不一致造成的，手工评价的 MRR 数值较高，是由于其手工选择的样例较少，可能都是搜索引擎处理的较好的样例，因而性能绝对数值都较高；而自动评价所选用的样例数量多（是手工评价样例的 40 倍以上），代表性好，其数值更能反映搜索引擎的实际处理能力的高低。

5.4 实验结果讨论

正如我们在第 2 节中所介绍的，我们的研究工作并不是首次把用户行为信息应用到搜索引擎评价中的尝试， Joachims[8]就提出了利用元搜索引擎接口评价搜索引擎性能的思路。但我们对以往工作中的核心假设，即用户点击的结果即为相关结果产生了怀疑，这促使我们从宏观而不是个体的角度考察用户行为，并提出了本章中上述的搜索引擎自动评价方法。

利用用户行为信息进行搜索引擎评价的工作容易被两方面的问题所困扰，其一是我们已经提到的用户点击的可信度问题，其二则是行为信息提供者（在本文的工作中即为搜狗搜索）本身的评价问题。

对于第一个问题，用户个体的点击行为确实容易受到多种因素的干扰而发生偏移。这是由于用户点击行为的作出是基于搜索引擎返回结果的标题与部分摘要文字作出的，但标题和摘要文字并无法完整的代表网页全貌，甚至有部分标题与摘要被专门设计用于欺骗用户点击（这在 SEO 操作中屡见不鲜），这就使个体用户点击行为的质量变得更不可信。

但从宏观而言，用户群体的点击行为还是可以信任的，如果某个页面吸引大量用户进行点击，但其并不是真正的查询目标页面，则这个页面要么是我们上文提到的与查询目标页面相关的同样吸引用户的一个页面；要么则是设计的极为出色的垃圾页面。从搜索引擎设计的角度，这样的垃圾页面会对用户体验造成毁灭性的打击，因此一般会被及时处理，而不会对用户的行为产生过大的影响。因此，用户群体在一定长度时间段内的行为特征是值得信任的。我们在本节的实验结果也充分验证了这一点。

对于第二个问题，我们认为用户行为信息的提供者不适宜同时作为被评价的对象。这是因为作为日志提供者的搜索引擎在评价过程中会具有评价偏向：由于只有出现在日志提供者索引中的网页才有可能被选择为标准答案，因此未出现在日志提供者索引中的正确答案也不可能被标注，即使其他搜索引擎返回了这样的答案，自动评价方法也无法辨识。例如http://www.sysu.edu.cn/和http://www.zsu.edu.cn/是中山大学主页的两个不同镜像，因此均应当被标注为“中山大学”的标准答案，但由于只有后者出现在了搜狗搜索引擎的索引中，因此只把前者作为查询结果返回的搜索引擎自然会被给予不合理的评价。这也正是我们没有在 5.3 节的实验中纳入搜狗搜索引擎作为被评价对象的原因。

6 结论与未来工作

性能评价一直是信息检索系统研究中的核心问题之一，传统的基于标准语料库、查询样例集合和标准答案集合的 Cranfield 方法，在处理搜索引擎的性能评价问题上面临着巨大的困难。基于用户查询点击行为挖掘的方法自动化的评价搜索引擎的查询性能，是本文要解决的主要问题。

为了实现提高检索系统查询处理能力的目的，我们提出基于用户群体行为分析的搜索引擎自动评价方法。该方法利用搜索引擎用户查询、点击行为的宏观分析，自动挑选适用于搜索引擎评价的查询集合，并进一步自动定位对应这些查询的标准答案。由于挑选查询集合和标准答案的过程由计算机自动完成，因此可以及时、准确、客观的反映搜索引擎的真实性能。

利用海量规模搜索引擎网络日志数据的实验结果证明，上述基于用户群体行为分析的搜索引擎自动性能评价方法能够自动提取用户查询需求样例，并准确标注超过 95%的导航类查询答案。针对中文搜索引擎性能评价的实验也发现，该自动方法性能评价的结果与手工评价的结果具有很高的一致性。比较传统的手工评价方法而言，自动评价方法具有及时、客观、准确、全面的特点。

综上所述，我们基于用户群体行为分析的思路实现了一种有效的搜索引擎性能自动评价方法，该方法适应大规模网络信息检索系统的应用需求，同时能够客观准确的实现搜索引擎系统的自动性能评价。

References:

[1] Saracevic, T. 1995. Evaluation of evaluation in information retrieval. In Proceedings of the 18th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, United States, July 09 - 13, 1995). E. A. Fox, P. Ingwersen, and R. Fidel, Eds. SIGIR '95. ACM Press, New York, NY, 138-146.

[2] Voorhees. E. M. 2001. The philosophy of information retrieval evaluation. In Proceedings of the Second Workshop of the Cross-Language Evaluation Forum, (CLEF 2001), pages 355-370.

[3] Soboroff, I., Nicholas, C., and Cahan, P. 2001. Ranking retrieval systems without relevance judgments. In Proceedings of the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (New Orleans, Louisiana, United States). SIGIR '01. ACM Press, New York, NY, 66-73.

[4] Nuray, R. and Can, F. 2003. Automatic ranking of retrieval systems in imperfect environments. In Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in informaion Retrieval (Toronto, Canada, July 28 - August 01, 2003). SIGIR '03. ACM Press, New York, NY, 379-380.

[5] Chowdhury, A., and Soboroff, I. 2002. Automatic Evaluation of World Wide Web Search Services. In Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Tampere, Finland, August 11 - 15, 2002). SIGIR '02. ACM Press, New York, NY, 421-422.

[6] Beitzel S. M., Jensen E. C., Chowdhury A., and Grossman D. 2003. Using titles and category names from editor-driven taxonomies for automatic evaluation. In Proceedings of the twelfth international conference on Information and knowledge management, 17-23.

[7] Amitay, E., Carmel, D., Lempel, R., and Soffer, A. 2004. Scaling IR-system evaluation using term relevance sets. In Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Sheffield, United Kingdom, July 25 - 29, 2004). SIGIR '04. ACM Press, New York, NY, 10-17.

[8] Joachims T. 2002. Evaluating Retrieval Performance Using Clickthrough Data. In Proceedings of the SIGIR Workshop on Mathematical/FormalMethods in Information Retrieval.

[9] Liu Y., Zhang M., Ru L., Ma S. 2006. Automatic Query Type Identification Based on Click Through Information, AIRS 2006, LNCS 4182, pp. 593–600.

[10] Lee, U., Liu, Z., and Cho, J. 2005. Automatic identification of user goals in Web search. In Proceedings of the 14th international Conference on World Wide Web (Chiba, Japan, May 10 - 14, 2005). WWW '05. ACM Press, New York, NY, 391-400.

[11] Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. 1999. Analysis of a very large web search engine query log. SIGIR Forum 33, 1 (Sep. 1999), 6-12.

[12] Yu, H., Liu, Y., Zhang, M., Ru, L., Ma, S. 2007. Research in Search Engine User Behavior Based On Log Analysis. Journal of Chinese Information Processing. Vol. 21(1): pp. 109-114, 2007.

附中文参考文献:

[12] 余慧佳，刘奕群，张敏，茹立云，马少平. 2007. 基于大规模日志分析的网络搜索引擎用户行为研究. 中文信息学报 Vol. 21(1): pp. 109-114, 2007.