

基于用户行为分析的搜索引擎 优化策略*

费巍 黄如花

武汉大学信息管理学院 武汉 430072

[摘要] 搜索引擎在给广大网络用户带来便捷的同时,也暴露出其不足。从网络用户利用搜索引擎的角度分析搜索引擎存在的问题以及用户利用搜索引擎时出现的障碍,在此基础上提出搜索引擎的优化模式。此外,提出由用户、知识生产者与知识组织者三者与搜索引擎共同组成一个信息系统的建议,以达到优化搜索引擎的目的。

[关键词] 搜索引擎 搜索引擎优化 用户行为 信息分析

[分类号] G250

Optimizing Search Engine Based on Analysis of User's Behavior

Fei Wei Huang Ruhua

School of Information Management, Wuhan University, Wuhan 430072

[Abstract] Search engine brings great convenient to netizens, but in the meanwhile, its weakness is also obvious. Search engine optimization has now become one of important tasks in the area of information retrievals. This article discusses the problems of search engine and the obstacles of users in using search engine. It also presents an optimization model, which regards users, knowledge producers and knowledge organizers as the exterior environment of search engine optimization, and composes an information system in combination with search engine, in order to optimize the search result.

[Keywords] search engine search engine optimization (SEO) user's behavior information analysis

2004年被称作“搜索引擎年”,搜索引擎之所以能成为互联网光环之下的焦点是因为它的实用性,尤以其直接著称。《中国互联网络发展状况统计报告(2005/1)》^[1]显示,在网上用户经常使用的网络服务/功能中,搜索引擎仅次于电子邮箱(85.6%),占到65%。几乎每一个上网的人,每一天都会使用搜索引擎。这背后蕴藏着巨大的商机,于是在国内搜索引擎界,外有Google虎视眈眈,内有百度、3721、中国搜索、搜狐、雅虎中国等群雄争霸。但是,现有搜索引擎系统所提供的Web查询技术还不能完全满足人们的应用需求,这主要体现在:检准率低,太多的检索结果使用户无所适从,现有的引擎系统对中文信息的处理不够有效,对学术信息的检索力不从心。

1 搜索引擎存在的问题

搜索引擎在给网络用户带来巨大便捷的同时,由于其信息检索技术智能水平的限制以及对自然语言理解的制约,对

网络信息的检索存在许多不足之处,主要表现在如下几方面。

1.1 技术的限制

现有的搜索引擎主要通过“网络蜘蛛”程序(spider)自动地在互联网中搜索信息,并将网页的全部或部分内容下载到自建索引库中,由于下载的页面许多是无用或暂时信息,既影响检索速度,又增加了用户检索负担。据调查^[2],有73.3%的用户认为,搜索结果重复率高,搜索到的网页打不开等是一个令人讨厌的现象。由于受技术的限制,搜索引擎在检索专业知识和多媒体信息方面做得尚不尽人意,有48.3%的用户认为专业/行业搜索功能差,有30.1%的用户认为多媒体搜索功能弱。而且,信息更新的速度也比较慢,有49.1%的用户这样认为。

1.2 要求用户掌握检索规则

搜索引擎一般都采用关键词检索方式,但许多情况下,用户很难简单地用关键词或关键词之间的组配来准确地表达真正需要的信息内容,表达困难导致检索困难。用户若想

* 本文系国家社会科学基金资助项目“网络信息组织模式的优化研究”(项目编号:03BTQ021)研究成果之一。

更便捷地获取信息,获取更高质量的信息,就需要掌握一定的检索规则,而不是仅仅通过关键词之间的组配进行检索。用户需要的是“傻瓜化”的检索系统,希望无需掌握纷繁芜杂的检索规则也可以用搜索引擎进行信息检索。

1.3 索引的比例有限

每个引擎的覆盖面都相当有限,没有一个搜索引擎的索引量超过整个网页的 1/6。即便搜索引擎界的龙头 Google 也是如此。全球 75% 的网上信息搜索是通过 Google 来完成的,通过 Google,全球网络用户能够使用 86 种语言,搜索 80 多亿个网页以及网页快照,11 亿多张图片。即使有如此大的信息量,Google 也并不是索引互联网上的每一个页面,它只倾向于索引包含博客、教育类信息以及新闻和信息类站点的最新页面(在三个月内创建的页面),一般会索引主流媒体站点在最近三个月内创建的页面而通常忽略那些信息量很少的某些类型的站点的页面。

1.4 搜索的结果不准确

搜索结果的准确性是由检索词与网页的相关性来确定的,用户输入的一个检索词能返回数万篇结果,或者零篇结果。在返回的大量信息中,只需粗略浏览便会发现有很多并非用户所需信息。很多情况下甚至还会发生网页跳转的现象,搜索引擎返回的结果标引为所需信息,但点击后却跳转到另一个网站或页面,这种情况的出现很大程度上源于网页或网站为了提高在搜索引擎中的排名而实施的作弊行为。

2 用户行为分析

用户既是搜索引擎的直接使用者,也是服务质量好坏的最终评判者。对用户使用搜索引擎行为的调查是搜索引擎优化尤为需要的,互联网为广大的网络用户提供了一个庞大的信息空间和自由获取信息的机会,而搜索引擎为用户找寻信息提供了指南。但搜索引擎给网络用户带来巨大便捷的同时也暴露出了不少问题,及时地解决这些问题,对搜索引擎进行优化则需要大量的用户信息。尤其要关注用户在使用搜索引擎时不满意的方面,并通过相关的软件技术对用户使用搜索引擎的行为进行跟踪,并对大量的资料进行分析,制定出优化搜索引擎的措施。

2.1 流量分析

网站流量统计最流行的术语是“点击”,监控和分析网站的流量是很重要的,可在获得网站访问量基本资料的情况下,对有关资料进行统计、分析,从中发现用户访问网站的规律,并将这些规律与搜索引擎优化相结合,从而发现目前搜索引擎中可能存在的问题,并为进一步优化搜索引擎提供依据,特别是针对网页和网站的排序优化。Alexa 网站^[3]提供全球网站访问流量排名,这也是 Alexa 最有影响力的一项服务。在该网站上可以查到全球各网站的流量排名及近 3 个

月来的流量和变化情况。如网站 google.com 近 3 个月在 Alexa 的排名为第三,而且从去年 12 月以来一直处于第三的位置(访问时间:2005/03/15)。通过流量分析还可以得知:

- 网站的主页浏览数。尽管许多用户并非通过主页进入网站,但是大多数通常还是会进入到主页,因此,主页浏览数可用于测量访问者的整体状况。

- 最主要的进入页面。对用户最主要的进入页面进行优化并进行重点维护,可有效地提高搜索引擎查询、标引的效率。

- 最主要的离开页面。了解用户最主要的离开页面的相关信息有助于改进页面的设计,甚至删除这些页面,这些措施能减少搜索引擎的工作强度。

- 每个访问的平均停留时间。

- 最多或者最少的访问页面。

- 每天访问高峰期。网站是需要不断维护和更新的,在访问高峰到来之前维护和更新更能获得搜索引擎的青睐。

2.2 用户利用搜索引擎的主要障碍

搜索引擎的工作是一个复杂的过程,其中的影响因素很多,存在着各种各样的障碍。如前所述,搜索引擎还存在不少问题,加之搜索引擎用户在信息检索方面缺少相应的专业知识支持,因此,在利用搜索引擎时遇到各种障碍是在所难免的。

2.2.1 技术障碍 搜索引擎给用户提供了一条快捷方便的获取信息的通道,但其返回的结果纷繁复杂,良莠不齐,要上万条信息中找到所需信息,用户不仅需要具备一定的知识经验以及信息识别能力,以从大量的,司空见惯的现象中迅速捕捉有价值的信息,还需要能熟练操作计算机,并能运用一些检索技巧,掌握必要的检索技术和一定的网络技术。如果缺乏这些相应的技术作为信息检索的支撑,就不能有效地查询到所需信息。

2.2.2 语言障碍 截至 2004 年 12 月 31 日,中国上网用户已达 9 400 万^[1],居世界第二,而中文网络信息只占整个因特网信息的 1% 左右,导致中文网络信息资源相对不足。互联网信息在全球范围内是开放获取的,全球网民可以自由加以利用,但用户在查询信息时要打破信息获取的地域限制,真正实现网上信息交流和资源共享,不可避免地会产生语言障碍,而且这也是使用搜索引擎较为突出的问题。

2.2.3 信息分析障碍 用户在搜索引擎上进行信息查询时,并不十分关注返回结果的多少,而是看结果是否和自己的需求吻合。在用户最看重的优点中,83.2% 的用户首要选择的是搜索结果的准确,另有 65.9% 的用户选择的是搜索速度快。结果准确和搜索速度快是目前用户对搜索引擎的主要需求^[2]。对于一个查询,搜索引擎动辄返回上万篇文档,用户不得不在结果中筛选,而后对众多的返回结果逐一进行分析和筛选则是对时间的一种极大浪费。解决查询结果过

多、质量不均的现象也是搜索引擎需要进一步发展优化的。

2.3 检索过程分析

对用户检索过程分析的内容包括用户查询词的分布情况、雷同查询词的衰减统计、相邻 N 项查询项的偏差分析、用户点击 URL 的分布情况等,其中对 URL 的分析在前文已有论述,这里不再赘述。用户查询分布的统计分析表明,用户的查询词是非常集中的,用户雷同查询项的统计分析表明用户查询有一定的稳定性,而相邻 N 项查询项的查询频率偏差很小且非常稳定^[4]。

用户在检索信息时,使用的查询词高度集中,且多是热门搜索词和为人们所熟知的词语。如百度搜索 Top50 中,QQ 日平均搜索次数达 50 382 次,mp3 日搜索次数达 45 266 次,日搜索电影的次数为 43 966 次^[5]。不难看出这些检索词都是人们日常生活中非常关注和熟悉的,通过搜索次数可以得知哪些事物是人们所追捧的。毋庸置疑,搜索引擎已成为品牌受欢迎程度的晴雨表,被搜索最多的品牌自然可以视为最受消费者追捧的品牌。2004 年初,Google 公布了前一年的热门搜索词,其中被搜索次数最多的前十大品牌依次为:法拉利、索尼、宝马、迪斯尼、低价航空公司 Ryanair、惠普、戴尔、低价航空公司 easyjet、旅行服务公司 last minute、沃尔玛,其中,法拉利被搜索 6.75 百万次而高居第一^[6]。

3 搜索引擎优化

将用户行为记录与分析结果应用于搜索引擎系统的优化,可提高系统的查询速度和信息检索的服务质量,有研究者将搜索引擎使用的 4 种信息——网页本身信息(author)、超链信息(other author)、人工编辑产生的目录信息(editor)和用户行为信息(user behavior)进行了比较,发现用户行为信息的利用对提高检索的查全率和查准率最有优势^[7]。

3.1 研究用户行为的工具

目前,有专门的网站提供各网站的各项统计资料,特别是对用户访问信息的统计。如:

- eXTReme Tracking^[8]。该网站提供访问者的来源、主页浏览数、最主要的进入页面、最主要的离开页面、每个访问的平均停留时间、最多或者最少的访问页面、每天访问高峰期、访问者使用什么浏览器。该网站还提供无限量 URL 实时跟踪,每个网站都可以通过点击该标志查看自己的统计信息。

- Web Site Traffic Report^[9]。在该网站填写一个简单的表格进行登记后,会发送一段 HTML 代码(只有 2 行),该站会在每天最后时刻通过 email 发送一份免费的网站访问量统计报告,包括每天的统计概况、网站流量表以及每个访问者的详细资料。

- The Counter^[10]。统计数据虽没有上述网站全面,但

仍提供了许多重要信息。

- Web Site Tracker^[11]。该站提供免费服务,可以统计许多信息,如访问者地理区域(按国家)、IP 地址、浏览器、提交 URL、每小时平均点击数等。

3.2 基于用户行为分析的搜索引擎优化

搜索引擎的优化可从多方位进行研究,如知识生产者的角度、知识组织者的角度、搜索引擎经营商的角度和用户利用的角度。本文仅从用户利用的角度对搜索引擎的优化策略进行分析,并在此基础上提出了优化搜索引擎的基本模式(见图 1):

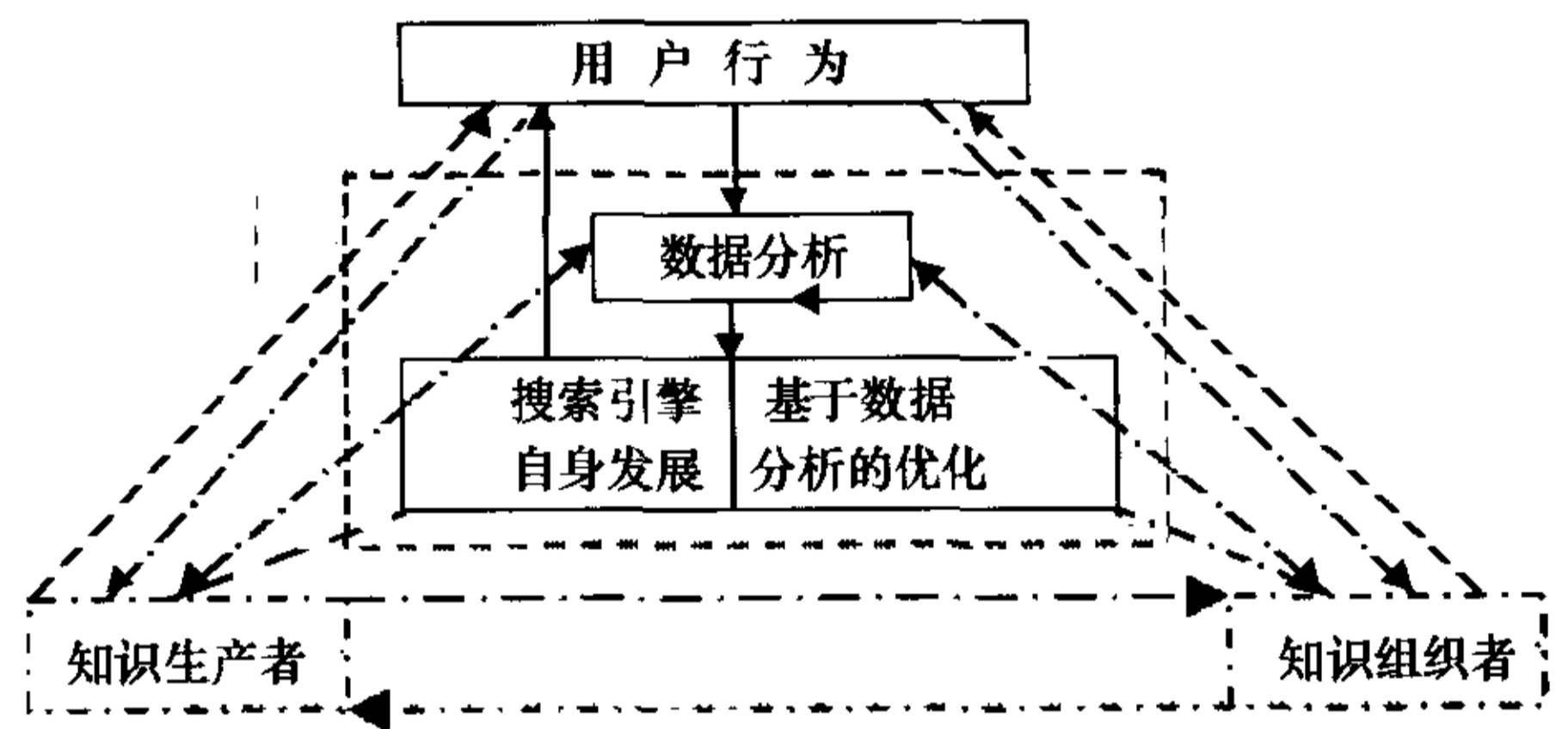


图 1 基于用户信息分析的搜索引擎优化模式

知识生产者、知识的组织者和网络用户者共同构成了搜索引擎优化的外部环境,三者与搜索引擎构成一个完整的信息链,避免各自为政,成为信息孤岛。从图 1 可以看出,用户与搜索引擎是互惠的关系。图中虚线方框内为搜索引擎优化的内部系统,由用户信息分析系统、基于用户信息的搜索引擎优化系统以及搜索引擎自身发展三部分组成,而后两部分则是整个优化系统的核心。用户信息首先被数据分析工具抓取,对信息进行统计和分析,再将分析所得的数据传递给搜索引擎,搜索引擎接受数据,并据此进行优化。如用户在利用搜索引擎检索信息时遇到的各种障碍便可由数据分析工具传递给搜索引擎,针对用户的困难,搜索引擎可采取各项措施来改进其自身能力。

针对目前用户利用搜索引擎出现的问题,可采取以下策略:

- 自然语言检索,以减少对用户检索技术的要求。
- 多语言兼容检索,可有效解决中文信息相对不足的状况。
- 基于本体和相似度的检索^[12],以提高检准率和检索效率。
- 基于内容的多媒体信息检索,满足用户多样化信息需求。

一旦搜索引擎得到优化改进,最大的收益者莫过于用户,用户在查取信息时将更加方便、快捷。因此,应使用户与搜索引擎优化之间形成一个良性循环。

参考文献:

1 中国互联网络信息中心. 中国互联网络发展状况统计报告. [2005 - 05 - 17]. <http://www.cnnic.net.cn/download/2005/2005011801.pdf> (下转第 110 页)

行评价与预测、优化地方文献馆藏与检索结构的有效工具。

2.4 建立资源开发的编研业务

档案编研是一种以馆藏档案为基础的编辑和研究业务,对于档案信息资源的开发利用具有重要的作用,是档案部门较具优势的业务领域。对一些确有价值、较为珍贵的地方文献来说,出于资源共享、便于传播、利于保护的目,图书情报部门也应开展专门的地方文献编研业务。而且由于文献信息形态在运作及观念上的差异,图书情报部门的地方文献编研工作应当比档案部门做得更有深度。对此,笔者认为:

- 吸收并运用档案编研的理念,正确认识“编”与“研”互为一体、相互递进的辩证关系。所谓“编”,就是对地方文献进行整理加工与汇集,是基础和前提性的工作;所谓“研”,主要是对地方文献的内容进行研究与考证,是对前者的深化与加工。惟有两者有机融合,才能为社会提供高质量的信息成果。在地方文献编研工作中,将逐渐形成的经验、程序、形式与方法上升为制度和规章,尽可能地规范化和制度化。

- 地方文献编研应以三次文献作为主要成果形式,强调分析、比较、综合的思维方式,着重揭示地方文献信息的核心和内在联系,注重对一次文献和二次文献进行再次整理和综合概括。依据特定的主题与形式,以满足用户信息需求为前提,尽可能形成多种类型的三次文献形态的参考性资料。结合图书情报部门的特点,一般可采用汇集、阐述、介绍、评说等多种方法,形成一批具有导航性的地方文献资源。

- 提供地方文献的查找线索,编制二次文献的检索和介绍性的工具。二次文献编研其实是信息要素的汇编,是对一次文献的部分信息材料进行科学的组合、集聚与排列。其

成果有概览、提要、文摘等介绍性文献和题录、索引等检索性文献。有条件的单位还应将散见于诸多搜索引擎中的反映地方文献的网页,采用链接索引、重新分类的方法,做成指南导航和专题目录。

- 按照一定的文献特征与逻辑次序,将地方文献的原文内容选编成一次文献形态的汇编、丛编、辑录或丛刊,其编研成果也属一次文献。这种文献对原文献内容而言,只是以特定的专题,在研究的基础上进行转录和考订,采用适当的体例编排组合成的。一次文献的编研成果,更有利于满足对地方文献专题性、真实性和系统性的检索需求。

- 与档案编研一样,以馆藏地方文献为基础参与史志的编修与有关著述的编著工作。

图书情报部门开展地方文献编研业务,有利于地方文献工作由单纯的文献服务型向文献资源开发型转换,同时也有利于地方文献工作者基本素质的提高。

参考文献:

- 1 林 岫. 图书馆地方文献信息资源的管理模式. 图书馆学刊, 2002(4): 22 - 24
- 2 许 萍. 西北地方文献收藏与开发研究. 图书情报工作, 2002(4): 77 - 78
- 3 王金夫. 图书馆地方文献与档案化因素. 图书馆理论与实践, 2004(1): 77 - 79
- 4 傅 虹. 关于地方文献信息服务工作的若干思考. 中国图书馆学报, 2000(6): 80 - 82
- 5 冯惠玲. 档案管理学. 北京: 中国人民大学出版社, 1999: 48 - 239
- 6 彭会兰. 开发地方特色档案信息资源的思考. 档案学研究, 1997(1): 46 - 48

[作者简介] 王金夫,男,1946年生,教授,中心主任,发表论文78篇,出版著作6部。

(上接第77页)

- 2 上海艾瑞市场咨询有限公司. 2004年中国搜索引擎研究报告(简版). [2005-05-17]. http://www.iresearch.com.cn/search_engine/detail_free.asp?id=12038
- 3 Alexa. Amazon.com company. [2005-05-17]. <http://www.alexa.com>
- 4 王建勇,李晓明,单松巍等. 海量Web搜索引擎系统中用户行为的分布特征及其启示. [2005-05-17]. <http://net.pku.edu.cn/~webg/papers/jwang-log.pdf>
- 5 百度热门搜索. [2005-05-17]. <http://top.baidu.com/2005/03/18>
- 6 是什么使它们成为最受追捧的品牌. [2005-05-17]. <http://sowang.com/news/20040216-1.htm>

- 7 Culliss G. User Popularity Ranked Search Engine. [2005-03-18]. <http://www.informotics.com/searchengines/boston1999/culliss/index.htm>
- 8 eXTReMe Tracking. [2005-05-17]. <http://www.extreme-dm.com>
- 9 WSTR - Website Traffic Report. [2005-05-17]. <http://www.websitetrafficreport.com>
- 10 The Counter.com. [2005-05-17]. <http://www.thecounter.com>
- 11 Website Tracker. [2005-05-17]. <http://www.websitetracker.com>
- 12 汪方胜,侯立文,蒋 馥. 基于本体和相似度的信息检索. 图书情报工作, 2005(2): 61 - 63

[作者简介] 费 巍,男,1981年生,硕士研究生。

黄如花,女,1968年生,副教授,副系主任,博士,发表论文60余篇,出版专著2部,参编著作4部,主编著作1部。