基于用户日志挖掘的搜索引擎广告效果分析

陈磊，茹立云，马少平

智能技术与系统国家重点实验室，清华信息科学与技术国家实验室，清华大学计算机系，北京 100084

Email: lei99@mails.tsinghua.edu.cn

摘要：随着搜索引擎市场的飞速发展，竞价排名广告以其有效、低风险、灵活等特点逐渐受到中小企业用户的青睐，成为搜索引擎稳定的收益增长点。然而竞价排名广告是否会影响用户体验，从而削弱其宣传效果并且影响用户对于搜索引擎的忠实度成为了企业及搜索引擎所担忧的问题。本文从网络用户日志中挖掘出了网络用户对于广告的实际交互行为，并给出了各大搜索引擎竞价排名广告方面的统计数据。对于企业用户如何更有效地利用竞价排名广告以及搜索引擎如何平衡广告的经济效益和用户体验之间的关系都有较高的指导意义。

关键字：搜索引擎，用户行为分析，竞价排名广告

Effectiveness of Online Sponsored Search Based on User Log Analysis

Chen Lei, Ru Liyun, Ma Shaoping

State Key Lab of Intelligent Technology and Systems, Tsinghua State Lab of Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084

Email: lei99@mails.tsinghua.edu.cn

Abstract: With the explosive growth of information available on the Web, more and more users adopt search engines to collection information on the internet. Meanwhile, sponsored search has become one of the most popular forms of Internet advertising because of its effectiveness and feasibility. However, it remains questions to us whether the sponsored search results become obstacles in users' information acquisition process. With analysis into large scale Web user access logs, we obtained several Chinese commercial search engines' sponsored search statistics. We also look into users' interaction behavior with sponsored search results and find out that search engine is hardly affected by sponsored search in meeting users' information needs.

Keywords: Search Engine; User Behavior Analysis; Sponsored search

1. 引言

近几年无论是国际市场还是国内市场搜索市场规模都在持续高速增长。《2007 年中国搜索引擎市场研究专题报告》指出“从全球市场来看搜索引擎市场规模持续快速增长，2007 年以 17.3%高速增长实现了 28.5 亿美元的规模。……2007 年中国搜索市场规模以 76.5%的高速增长达到了 29.3 亿元。”搜索引擎成为人们日常工作和生活中不可或缺的信息获取手段。根据中国互联网络

图 1：竞价排名广告样例

信息中心（CNNIC）2008 年 1 月的统计，我国的 2.1 亿网民群体中搜索引擎的普及率达到 72.4%，已经超过了电子邮件服务的普及率，其中更是有 84.5%的搜索引擎用户将搜索引擎作为得知新网站的主要途径。

因此，在搜索引擎检索结果中取得有利的排名已经成为网络资源尽快获得用户关注的最有效途径之一。而以 Google 为代表的竞价排名广告服务则无疑为广告商宣传网络资源提供了一个最直接的途径，竞价排名广告（sponsored search）也成为了搜索引擎盈利的主要途径之一。

竞价排名广告指的是企业通过向搜索引擎付费使企业信息出现在搜索结果页上显眼的位置的广告方式，所需支付的费用与位置的显眼程度与该信息实际被点击的次数有关。竞价排名广告通常以“赞助商链接”的“推广”的形式出现在搜索结果页的顶端或侧边栏，见图 1。

搜索引擎的竞价广告之所以成为众多商家尤其是中小企业所热衷的广告方式，主要是因为其有效、低风险以及灵活的特点。竞价广告的有效性体现在搜索引擎的用户众多，甚至很多用户以其作为浏览 internet 的主要入口，因此某些热门查询词的包含广告的搜索结果页的访问量丝毫不逊于其他门户网站中主页。而大部分在线用户（62%）并没有意识到自然搜索结果与竞价广告之间的区别，越是排名靠前的广告越是能得多更多的关注。[1][2]通过建立经济学模型论证了竞价广告的有效性，并对比了它与传统广告的差异。而搜索引擎按照点击次数收费的方式比传统媒体显示即收费的方式更加合理，可以在保证广告效果的同时降低企业成本和风险。企业用户可以自行选择投放广告的关键词，投放的位置，投放的时间，有的甚至还可以全部在网上操作完成，因此使用相当灵活。并且广告的效果可以及时现象，方便用户根据效果调整投放策略。

然而，随着竞价广告的数量不断增多，人们出现了这样的担忧：竞价广告越来越多地取代了传统的自然搜索结果前几名的位置，甚至有些查询词搜索结果的第一页全部都是广告，比如“机票”、“加盟”、“理财”等等，会不会因此影响了用户体验，使得网络用户认为搜索引擎的搜索结果质量下降从而降低用户忠实度，并且使得广告的宣传效果收到影响。

因此本文希望通过分析搜索引擎用户日志的方法了解网络用户面对竞价排名广告的实际行为反应，从而分析竞价排名广告的实际宣传效果。本文按照如下方式组织：第二部分介绍搜索引擎日志的收集方法、数据分析方法；第四部分根据用户行为数据分析竞价排名广告的实际效果；第五部分根据现状对于企业用户和搜索引擎给出建议。

2. 用户行为数据收集

2.1. 数据来源

随着搜索引擎技术的发展，由搜索引擎公司提供的浏览器工具栏越来越为广大网络用户所接受。浏览器工具栏可以为用户提供直接的搜索引擎访问接口，同时也可以提供弹出窗口过滤、下载加速、网络书签等多种附加功能。目前的主流搜索引擎公司如谷歌（http://toolbar.google.com/）、雅虎（http://toolbar.yahoo.com/）、百度（http://bar.baidu.com/）、微软（http://toolbar.live.com/）等都推出了自己的浏览器工具栏服务，不少公司还把工具条与其他软件产品捆绑发行以加强推广。与此同时，大多数搜索引擎供应商也通过工具栏基于匿名策略收集用户的 Web 访问行为数据，以便为工具栏用户提供更多个性化的增值服务。最近，一些研究人员也开始利用这部分 Web 访问行为数据对网络用户的行为特征加以研究和利用（如[3]）。

本文的研究工作中，我们在某商业搜索引擎公司的协助下，利用工具栏的方式于 2008 年 4 月 22 日至 2008 年 5 月 13 日期间除去 5 月 1 日共 21 日，收集了 9 亿次用户点击的 Web 访问日志。日志所记录的信息项目如下表所示：

表 1. 用户 Web 访问日志记录的信息

名称	内容
时间	用户点击行为发生的时间
源 URL	用户点击发生时，其正在访问的网页
目的 URL	用户点击发生时，其点击链向的网页
ID	系统自动分配的用户标识号

上表所记录的信息可以被绝大多数搜索引擎的工具栏软件所收集，因此本文工作所涉及的方法具有较高的可行性。同时，这部分日志完全不涉及用户的个人隐私信息，而是用系统自动分配的 ID 号标识用户，也保证了日志分析工作的合理性。

2.2. 搜索引擎 url 格式分析

根据对国内 6 大著名搜索引擎，包括：谷歌、雅虎、百度、搜狗、搜搜和有道的实际观察，我们发现按点击量排名由大到小依次是：百度、谷歌、搜搜、雅虎、搜狗、有道。有道没有置顶广告只有侧边栏广告，并且数据量过于稀疏，因此忽略。而搜搜尽管使用了谷歌的广告并且没有侧边栏广告，但因为搜搜的搜索点击率仅次于与百度与谷歌，因此仍然保留。

表 2. 文中使用的搜索引擎及其 url 格式

搜索格式	默认编码	相关广告格式
http://www.baidu.com/s?...wd=$KEYWORD... http://www.baidu.com/baidu?...word=$KEYWORD... http://www.baidu.com.cn/s?...wd=$KEYWORD...	GBK	http://www.baidu.com/baidu.php?...
http://www.google.cn/search?...q=$KEYWORD...ie=$CODE... http://www.google.com/search?...q=$KEYWORD...ie=$CODE...	UTF-8 自适应	http://www.google.cn/aclk?...
http://search.cn.yahoo.com/search?...p=$KEYWORD...ei=$CODE...	GBK	http://click.p4p.cn.yahoo.com/ci_im?...

http://www.yahoo.cn/s?...p=$KEYWORD...	中文	
http://www.sogou.com/web?...query=$KEYWORD... http://www.sogou.com/sohu?...query=$KEYWORD...	GBK 中文	http://click.cpc.sogou.com/bill_search? http://click.cpc.sogou.com/bill_biz?
http://www.soso.com/q…w=$KEYWORD	GBK	http://www.google.cn/aclk?...

3. 搜索引擎竞价排名广告效果分析

3.1. 不同搜索引擎的搜索与广告点击量

表 3 列出了 5 个搜索引擎各自每日平均搜索结果点击量与广告点击量，第二列是搜索结果点击量单位是次，其中包括了广告点击，第三列是广告点击量单位也是次，第四列是广告点击量占搜索结果点击量的百分比（本文将其简称为“广告点击率”），第五列是 21 日广告点击率的标准差，标准差比平均值小一个数量级，表明 21 日内广告点击率变化幅度不大。

根据表 3 可以看到百度无论是搜索结果点击量还是广告点击量都遥遥领先于其他搜索引擎，它的搜索结果点击量占到所有搜索引擎的 79%，而广告点击量更占到了 84%。然而无论是哪个搜索引擎，其每日的广告点击量仍然只占总搜索结果点击量的 0.12~0.27%，可见竞价广告的市场空间还很大。

值得注意的是尽管搜搜使用了谷歌的广告，然而它的广告点击率却高出谷歌自身广告点击率的一倍以上。这主要是因为搜搜没有设侧边栏广告，而将原本出现在谷歌侧边栏的赞助商链接显示在了结果页面的最上端，尽管广告数量通常不超过 3 条，但置顶广告比侧边栏广告更能吸引用户的点击。

表 3. 不同搜索引擎的搜索结果点击量与广告点击量（21 日平均值）

搜索引擎	所有搜索结果点击量	广告点击量	广告点击率	标准差
百度	3,087,709	5,983	0.19%	2.26E-04
谷歌	627,424	733	0.12%	1.14E-04
雅虎	43,088	76	0.18%	3.09E-04
搜狗	23,445	37	0.16%	2.88E-04
搜搜	103,322	279	0.27%	3.71E-04

3.2. 广告点击量分布

表 4 统计了 21 日来出现过的所有查询词的数量、总点击次数、单词最高点击次数，平均点击次数由总点击次数除以总词数得到。

表 4. 查询词数量统计（21 日合计）

	词数	总点击数	平均点击数	最高点击数
所有搜索结果	**1,304,211**	**81,185,985**	**6.102277317**	**384,798**
广告	**69,938**	**145,186**	**2.07592439**	**812**
百分比	**0.53%**	**0.18%**	-	-

首先，根据表 4 的统计可以看出，有广告点击的查询词只占很小的一部分。有广告点击的查询词的广告点击次数平均仅为 2 次，小于查询词总体 6 次的平均结果点击次数。单个查询词最高广告点击次数更是远远小于所有结果点击次数。这主要是因为竞价广告的数量比起搜索结果数量要少得多。

图 2(a)是查询词数量根据不同词频的分布。横轴是单个查询词的点击次数（指数轴），纵轴

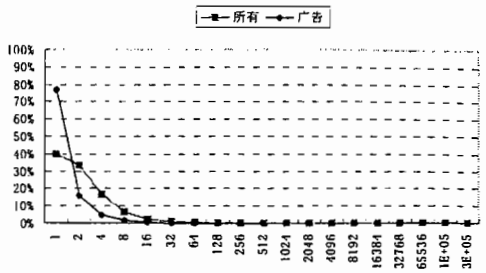

图 2(a). 查询词数量根据词频的分布

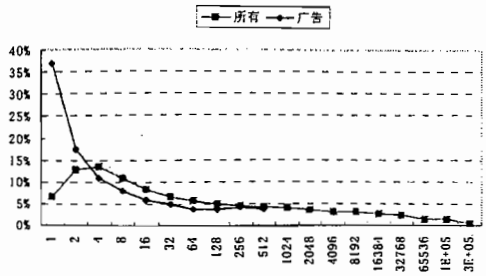

图 2(b). 查询词点击量根据词频的分布

是查询词个数的百分比，图中每个点表示点击数在 2n~2n+1-1 之间的查询词个数占查询词总数的百分比。图中可见，点击次数小于 8 次的查询词占到了总数的 90%左右，而广告点击次数小于 4 次的查询词就占到了总数的 90%以上。图 2(b)是查询词点击总数根据不同词频的分布。图 2（b）与图 2(a)的不同在于纵轴是点击数的百分比，即将点击次数在 2n~2n+1-1 之间的查询词的点击数总除以总点击数。从这张图上可以更清楚的看到广告点击次数小于 4 次的查询词不仅占了所有包含广告点击的查询词的绝大大部分，并且吸引了大部分的点击量（大于 50%）。这也说明了查询词的广告点击量普遍很低。

3.3. 查询词点击量与热门程度的关系

定义： 热门程度 = 点击次数 / 平均点击数

我们认为热门程度大于 10 的查询词是热门查询词。根据统计，广告点击次数大于 20 次的查询词（493 个）都是热门查询词，这些查询词点击次数平均为 3744 次，最高 65456 次，最低 64 次，高于热门搜索查询词要求的点击次数在平均点击数的 10 倍以上。

然而还有很大一部分热门查询词没有广告点击。热门查询词（共 98099 个）中只有 2090 个有广告点击，只占总数的 28.08%；这主要是因为热门查询词中大部分是网站导航类型的查询词，比如百度、淘宝、校内、新浪等等，在前 100 个热门查询词中有 50 个是这样的网站导航类查询词。这些词在自然搜索结果的前几页都能找到相应的网址，所以投放广告意义不大。而另一方面很多网络用户以搜索引擎作为浏览 internet 的入口，而对于网站的描述通常相对单一，比如如果要上新浪，常见的查询词只有新浪、sina，而不像其他搜索问题可能有多种的描述方式。因此网站导航类查询词成为了点击量最高的查询词。

3.4. 广告点击量与排名的关系

“推广”与“赞助商链接”是最主要的两种竞价排名广告形式。谷歌主要采用“赞助商链接”，并且大部分“赞助商链接”都在侧边栏，只有少数查询的“赞助商链接”会以置顶方式出现，并用特殊背景颜色标注。百度的侧边栏和谷歌类似，而出现在搜索结果里置顶的广告项目称为“推广”，格式与自然搜索结果几乎一致。搜狗的排版与百度基本上一样，但偶尔用特殊背景颜色标注的“赞助商链接”会出现在搜索结果置顶里，但“赞助商链接”与“推广”不同时出现。但是雅虎的侧边栏比较复杂，而置顶广告称为“推广”，用特殊背景颜色标出，并且广告项目不多于 3 个。搜搜没有侧边栏广告，以置顶的方式显示来自谷歌的“赞助商链接”，但不总是用特殊颜色加以区分，同样也是不超过 3 个。

综上，我们可以这样区分两种广告形式，尽管名称可能不尽相同：

“推广”是混于自然搜索结果里的竞价排名广告，其格式与自然搜索结果几乎完全一致，没有特殊的颜色标示，仅在最后以“推广”二字标示；

“赞助商链接”以区别于自然搜索结果的形式出现在搜索结果页面上，比如出现在侧边栏，或者以特殊背景颜色标出。因此“赞助商链接”的数量通常不会影响每页显示的自然搜索结果的数量。

由于“85%的用户只翻看搜索引擎返回结果的前 10 个结果，即返回结果页面的第一页”。[4] 我们分别对这两种不同类型的广告——推广、赞助商链接从第一名到第十名的点击量分布进行了统计，见图 3。“推广”和“赞助商链接”其第一位比第二位的点击量分别高出了 7 和 9 倍，分别占到前十位总数的 69%和 75%。而实际上结果页面里仅显示 1 个广告的查询词只占很少的比例（小于 1%）。也就是说大部分用户仅点击处于第一位的广告。

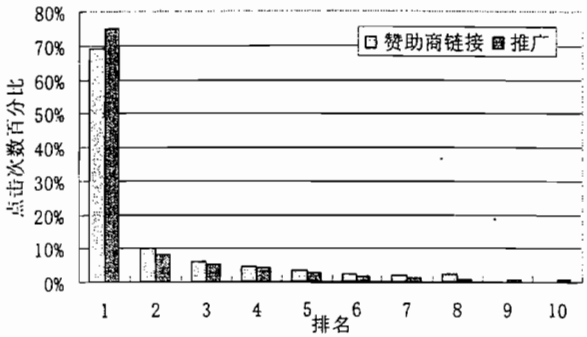

图 3. 不同类型广告点击量根据排名的分布

3.5. 广告的存在对于用户体验的影响

在数据的观察，我们发现对于有些查询词，点击量不是随排名的增加而急剧减小，而出现某些倒置，也就是说广告的出现可能影响了用户体验。因此我们对所有点击量大于 10 次的查询词的搜索结果页面进行了实时抓取，检查它是否包含广告，再从日志里统计出包含“推广”的页面、包含“赞助商链接”的页面和所有搜索结果首页根据排名的点击量分布，见图 4。与图 3 不同的是，在这次统计中加入了页面中包含广告但却没有发生广告点击的情况。另外，由于“推广”与“赞助商链接”在展示上采用了不同的策略（见 4.4），在统计中“推广”项目的排名与自然搜索结果的排名是统一的，而“赞助商链接”排名与之不统一。也就是说包含“推广”的页面里排名第一的应该是广告，而包含“赞助商链接”的页面里排名第一的并不是“赞助商链接”而是自然搜索结果或者“推广”广告。这样我们也可以看出不同的展示方式对于用户的不同影响。

从图 4 可以看出尽管在某些个例上出现了前面说的点击量倒置的情况，但从全局来看点击量与排名仍然是成反比的。对于包含“推广”的页面，用户反倒更倾向于点击排在第一位的广告；而对于包含“赞助商链接”的页面，搜索结果的点击量还是受到了排在前面的“赞助商链接”的影响，点击量随排名的衰减相对平缓。

这说明了实际的情况是用户并没有对竞价排名广告产生排斥，相反用户是乐于接受广告的。可能的原因有：

1. 搜索引擎对于排名的优化提高了广告的质量。比如谷歌的广告排名就不单纯的与企业出价有关，如果排名靠前的广告并没有得到足够的点击量，那么它的排名就会下降。[5] 的研究也发现广告的相关度要比自然搜索结果高，而广告的排名也与相关度一致。
2. 投放广告的查询词通常都是与产品相关的，比如：彩票、电影、手机等，用户点击这样的查询词目的主要是查询产品信息和网上购物，而广告投放商的目的就是宣传他们有这种服务，并且为了更高的提供服务吸引用户都编辑了详细的产品描述，正好能够符合网络用户的需求，自然点击量也会较高。

当然对于某些个例的点击量倒置现象也应该引起搜索引擎供应商的足够重视，避免进一步影响用户体验。

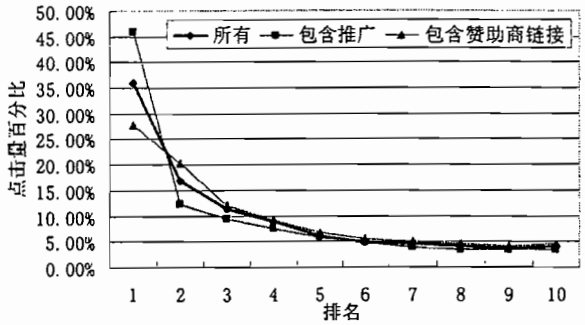

图 4. 包含广告与所有搜索结果首页点击量根据排名的分布

4. 结论

本文分析了谷歌、百度、雅虎、搜搜、搜狗五大搜索引擎的广告推送形式，并从 9 亿次点击的网络用户日志中挖掘了用户对于这五大搜索引擎的广告反应行为。百度作为最大的中文搜索引擎，日点击量逾千万，而日广告点击量也只占所有搜索结果点击量的 0.19%。竞价排名的市场空间还很大。有广告投放的相对热门（点击量大于 10 次）的查询词只占总数的 5%左右。而每个查询词的平均点击只有 2 次，大部分用户只会浏览点击搜索结果页面中第一页，并且仅点击排在首位的广告。因此与其集中在几个热门词投放广告，还不如多投放一些相对不那么热门的查询词广告并取得较高的排名，而这样所需的费用也许更低。

目前来看用户并没有排斥广告的倾向，对于有广告投放的页面首位广告反倒吸引了接近50%的点击量。但是某些点击量倒置的个例也需要引起搜索引擎供应商的重视，怎样优化能使广告排名与相关度一致是需要考虑的问题。竞价广告的推送形式也是搜索引擎需要关注的问题。“赞助商链接”不如“推广”能够吸引广告点击量。搜搜尽管引用了谷歌的赞助商链接，但由于改变了展示策略广告点击率达到 0.27%，为五大搜索引擎之首。

参考文献

[1] Animesh A, Vandana R, Siva V. An Empirical Investigation of the Performance of Online Sponsored Search Markets. ICEC'07, 2007, p153-160.

[2] Anindya G, Sha Y. An Empirical Analysis of Sponsored Search Performance in Search Engine Advertising. NET Institute Working Paper, 2007, p7-35

[3] Bilenko, M. and White, R. W. Mining the search trails of surfing crowds: identifying relevant websites from user activity. In Proceeding of the 17th international Conference on World Wide Web (Beijing, China, April 21 - 25, 2008). WWW '08. ACM, New York, NY, p51-60.

[4] 余慧佳，刘奕群，张敏等. 基于大规模日志分析的网络搜索引擎用户行为研究. 第三届学生计算语言学研讨会论文集，2006，p202-207.

[5] Bernard J. The Comparative Effectiveness of Sponsored and Nonsponsored Links for Web E-commerce Queries. ACM Transactions on the Web,2007, Vol. 1, Article 3.