文章编号：1003-0077(2007)01-0109-06

基于大规模日志分析的搜索引擎用户行为分析

余慧佳[1]，刘奕群[1]，张敏[1]，茹立云[2]，马少平[1]

（1.清华大学 智能技术与系统国家重点实验室，北京 100084；2.搜狗公司 研发中心，北京 100084）

摘 要：用户行为分析是网络信息检索技术得以前进的重要基石，也是能够在商用搜索引擎中发挥重要作用的各种算法的基本出发点之一。为了更好的理解中文搜索用户的检索行为，本文对搜狗搜索引擎在一个月内的近5 000万条查询日志进行了分析。我们从独立查询词分布、同一session内的用户查询习惯及用户是否使用高级检索功能等方面对用户行为进行了分析。分析结论对于改进中文搜索引擎的检索算法和更准确的评测检索效果都有较好的指导意义。

关键词：计算机应用；中文信息处理；网络信息检索；搜索引擎；用户行为分析；点击信息分析

中图分类号：TP391 文献标识码：A

Research in Search Engine User Behavior Based on Log Analysis

YU Hui-jia[1], LIU Yi-qun[1], ZHANG Min[1], RU Li-yun[2], MA Shao-ping[1]

(1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084;
2. Sogou R&D Centre, Beijing 100084)

Abstract: User log analysis is important for both Web information retrieval technologies and commercial search engine algorithms. In order to better understand search behavior of Chinese Web search users, we presents an analysis of Sogou Search Engine query log consisting of approximately 50 million entries for search requests over a period of one month. The analysis includes search retrieval behavior in individual queries distribution, user request customs in the same session and whether using advanced search functions. Conclusions may help improve Web information retrieval algorithms and search performance evaluation methods.

Key words: computer application; Chinese information processing; web information retrieval; search engine; user behavior analysis; click through data analysis

1 引言

随着网络与信息资源的飞速发展，网络搜索引擎已经成为人们获取网络信息的主要途径。但现在人们通常只是简单地通过短短几个词的查询与检索系统进行沟通，而在网络信息资源规模如此庞大的情况下这种沟通是远远不够的，检索系统往往不能比较准确地返回用户所真正需求的信息。因此，进行搜索引擎的用户行为分析是非常必要的。

搜索引擎日志是网络搜索引擎用户行为的重要载体，国内外的不少研究者都针对网络搜索引擎的用户日志进行了相关的研究。网络信息检索工具得到普及之后，面向网络信息检索的用户行为分析得到了更多的关注，文献[1~3]就分别在90年代中期左右对Web用户的浏览行为进行了调研和分析；到1998年前后，部分研究者如文献[4,5]等就开始对商业搜索引擎的用户日志进行大规模的分析。但由于各方面条件的限制，这种研究，例如查询词频分布规律的研究等，都很少集中在中文网络用户的行

收稿日期：2006-07-26 定稿日期：2006-10-11

基金项目：国家重点基础研究(973)资助项目(2004CB318108)；国家自然科学基金资助项目(60223004，60321002，60303005，60503064)；教育部科学技术研究重点资助项目(104236)

作者简介：余慧佳(1985—)，女，本科生，主要研究方向为信息检索。

为分析上。

中文网络数据环境与英文的有较大的差异，除了数据上的，还有使用群体的文化、语言习惯等差异，这些都造成了中文搜索引擎用户行为上的特异性。因此有必要对中文搜索引擎的用户行为进行分析，以针对中文搜索引擎的算法或检索性能评测方法等指出有益的方向。

本文将对为期一个月的真实规模中文搜索引擎网络日志进行研究，从较大规模的数据中分析中文搜索引擎用户行为的一些特点，为中文搜索引擎算法的改进和检索性能的评测等提供一定依据和方向。因为日志数据规模较大，所以更具一般性，更能反映出大部分用户的行为特征。在下文中我们将首先对已有工作和搜索引擎的日志设计等作简单的介绍；然后对基于日志的搜索引擎用户的行为进行分析；最后针对中文与英文搜索引擎用户行为差异尝试提出一些对于中文搜索引擎算法设计、评测方法设计有益的启示。

2 已有工作概述

2003 年第十四、十五期中国互联网络发展状况统计报告[6] 指出，2004 年中国搜索引擎用户已占互联网用户的 95.2%，每天的搜索请求量达到近 1.9 亿次。而根据最近发布的壹期中国互联网络发展状况统计报告[8]，截止到 2005 年 12 月 31 日，我国的网民数达到了 1.11 亿，65%的用户指出搜索引擎是他们经常使用的网络服务功能。另外，根据 Sullivan 的统计[9]，2004 年底，Google 作为世界上索引量最大和访问频率最高的搜索引擎，能够索引到超过 80 亿的网络页面，而其每天处理的用户查询则超过 2.5 亿个。

面对如此庞大的搜索需求，深入挖掘发现用户行为特点，进而提高搜索引擎算法的效率和准确率显得尤其重要。1998 年，Craig Silverstein 等人对大规模英文搜索日志进行了分析[4]，结论指出 85%的查询用户都只翻看了查询结果的第一页内容等。这些结论都对英文搜索引擎的算法改进和发展起到了有益的作用。由于中文网络数据的特殊性以及中英文网民行为的差异，对中文搜索引擎进行较大规模的分析以找出中文搜索引擎用户的行为特征是很有必要的。

此外，对用户检索目的的分析也是近年来用户行为分析研究的热点之一，IBM 研究院的 Broder 首先提出了“任务驱动”的概念，在他构想的用户检索流程模型中，查询任务决定了用户的查询需求，进而反映在查询词上。他在文献[10]中指出，用户的查询任务包括导航类、信息类和事物类三类。对查询任务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所区别。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。

3 搜索引擎用户的行为构成与日志设计

搜索引擎用户的行为构成可由图 1 表示。

用于分析的搜狗网络日志由一系列查询需求组成，每个查询需求都包括如表 1 所示条目。

利用查询词和用户点击页面的信息，我们可以分析出用户提交的查询一般有什么特点，如长短、频度等；而由用户点击结果页面的信息我们能得到用户的点击习惯等。我们的实验主要是建立在对大量的用户需求进行统计的基础上的宏观分析，主要目的是寻找用户需求中的热点、词频分布规律、查询行为特点等，进而对检索系统的系统结构和算法设计做出改进。

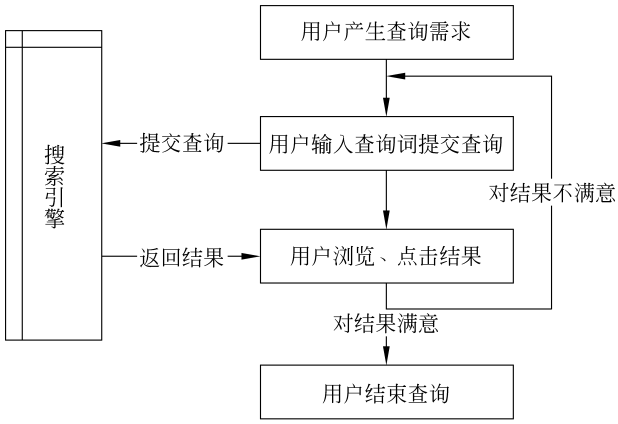

图 1 搜索引擎用户的行为流程

表 1 搜狗网络日志的内容

名 称	记 录 内 容
query	用户提交的查询
URL	用户点击的结果地址
time	用户点击发生时的日期、时间
rank	该 URL 在返回结果中的排名
order	用户点击的顺序号(这是用户点击的第几个页面)
id	由系统自动分配的用户标识号
submitter information	浏览器信息，计算机信息

4　基于日志的搜索引擎用户的行为分析

4.1　独立的查询分析

实验所使用日志包括搜狗搜索引擎在 2006 年 2 月 1 日至 2 月 28 日的 28 天内的所有查询。其中非空查询共 45 745 985 个，含非重复查询共有 4 345 557 个，session 个数为 26 255 952 个。

4.1.1　查询的长度

查询的长度主要指的是用户提交的查询中包含几个词语或字(用空格隔开的)，分析结果中，长度不超过 3 个词的查询占了总查询数的 93.15%，平均长度为 1.85 个词，这说明用户输入的查询通常都比较短。而且平均长度与 Craig Silverstein 等人[4] 分析的英文查询长度结果的 2.35 个词相比更短，这说明中文搜索引擎得到的用户需求信息更少，需要对用户需求有更多的分析和经验，才能更加准确地返回用户需求的信息。

4.1.2　查询的频度

查询的频度是指在整个 2006 年 2 月份的网络搜索日志中，该查询一共被提交过多少次。对于出现次数最多的前 150 个查询，我们将其出现次数及排名绘成图 2。

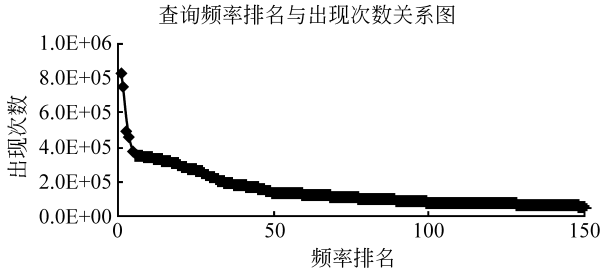

图 2　查询的频度排名与出现次数的关系

从图 2 中可以看出少数查询出现的次数很多，而我们得到的结果是：出现次数大于 100 次的 query 总数为 35 177 个，占非重复查询总数的 0.8%，但其总的出现次数却为 59 736 863 次，占总查询数目的近 70%。这说明在搜索引擎每天处理的大量查询中，有很多查询都是重复的，很少一部分查询就占了用户需求的大部分。如果搜索引擎能够通过某些方法提高这少部分经常出现的词的查询质量，就能使整体的检索质量提高不少。同时也证明了在搜索引擎设计中引入缓存(cache)机制或人为干涉的必要性与可行性。

而在对查询 term(term 指的是用户提交的查询中被空格分隔开的单个的词或字)的统计中发现，在出现次数最多(均大于 500 000)的 12 个 term 中，有 50%的 term 与图片相关，表明现在人们对图片信息的需求量越来越大，因此搜索引擎在图片搜索方面应予以重视。

4.2　Session 相关分析

一个 session 指的是同一个用户在某一小段时间内的连续查询。对于某一小段时间的定义，是由搜狗搜索引擎的网络日志决定的，即对于同一用户，在他开始使用搜索引擎检索到他关闭浏览器的那段时间就定义为一个 session。

4.2.1　每个 session 中的查询数分析

在实验中我们对每个 session 中所含的查询个数进行了分析，得出的结果中在同一个 session 内查询的平均个数为 1.75，有 66.46%的 session 只含有一个查询，即在那小段时间内，大部分用户只提交了一个查询且没有对该查询进行修改。造成这种情况的原因可能是用户对检索结果表示满意，找到自己想要找的信息后结束查找，也可能是对检索结果不满意，但又不想修改查询词后再次搜索了。这与 Craig Silverstein 等人分析的结果 63.7%基本一致[4]。

4.2.2　一个 session 内修改查询方式所占比例

当用户提交一个查询后，如果对搜索引擎返回的结果不满意时，用户有可能会在原有查询词的基础上进行增加或删减字词。另外一种更普遍的情况是，在一个 session 内，用户很可能彻底更换查询内容。对于那些一个 session 内提交了 2 个以上查询的情况(即用户对原查询进行了修改)，我们分析了用户修改查询词的各种方式所占的比例，详见表 2。

表 2　在一个 session 内中文搜索引擎用户对查询的修改方式分布

查询的不同修改方式	平均占 session 中修改过查询数的比例
Adding terms	9.00%
Deleting terms	1.43%
Totally changing the query	83.27%
其他修改方式	6.30%

注：Adding terms 和 Deleting terms 包括在任意位置增加或删除的改动。

当用户对查询不满意而适当修改时(除去全部改变的情况)，很大程度是因为返回结果的搜索范围较大，因此用户会选择增加查询词以限制搜索范围，

搜索结果过于冗余是搜索算法应该重视的一个问题。

4.3 其他用户行为相关分析

在不同的用户群体中表现出来的用户行为特征是有所不同的。例如用户习惯点击的结果在搜索引擎返回结果中的大体位置如何，或者中文用户有多少会提交含有英文的查询、有多少用户会采用高级检索或直接键入 URL 地址作查询词等。这个部分将对这些问题进行一定分析。

4.3.1 点击次数与 rank 之间的关系分析

用户提交一个查询后，搜索引擎可能会返回很多页结果，但是并不见得这些结果都会对用户有用，因为用户一般不会将这些结果点击浏览过。在我们的实验中，我们分析了搜索引擎返回的结果的顺序排名(rank)与被点击次数的关系如图 3，取对数值后则得图 4。

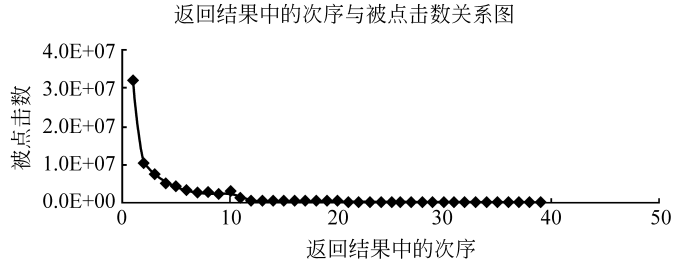

图 3 返回结果的顺序排名与被点击次数的关系

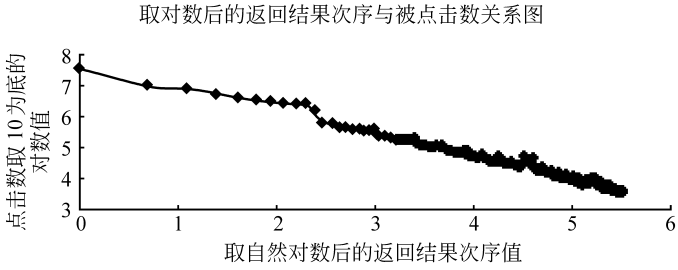

图 4 对返回结果的顺序排名与被点击次数取对数值后的关系

实验数据显示约 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。这个用户行为决定了尽管搜索引擎返回的结果数目十分庞大，但真正可能被绝大部分用户所浏览的，只有排在最前面的很小一部分而已。所以传统的基于整个结果集合查准率和查全率的评价方式不再适用于网络信息检索的评价，我们需要着重强调在评价指标中有关最靠前结果文档与用户查询需求的相关度的部分。TREC 在近年组织的网络信息检索评测[11, 12, 14, 15]中，以及针对中文网络信息检索的评测[15, 16]都采用了更重视检索结果最前的少数几篇文档是否满足用户需求。

4.3.2 用户使用高级检索的比例

在 1998 年 Craig Silverstein 等人分析的结果中[4]，超过 20%的查询中含有+、-、and、or 等符号以运用高级检索功能进行检索。但在我们的实验结果中，却只有约 0.73%的查询中含有用于高级检索功能的符号，即目前中文检索用户更多的检索方式只是简单地输入几个关键词用以查询。这说明在使用检索系统的过程中，简便仍是被用户所看重的，也说明了各种复杂功能的指定应从用户使用便利的角度出发。

4.3.3 直接输入 URL 作为查询词的比例

在实验结果中，有 2.82%的查询是用户直接输入 URL 部分或全部地址进行查询的。对这些包含 URL 的查询进行统计分析后发现，平均有 32.41% 的点击数点击的结果就是用户输入的 URL 的网址。从这个比例可以看出，很大一部分用户提交含有 URL 的查询是由于没有记全网址等原因而想借助搜索引擎来找到自己想浏览的网页。因此搜索引擎在处理这部分查询的时候，一个可能比较理想的方式是首先把相关的完整 URL 地址返回给用户，这样有较大可能符合用户的查询需求。

4.4 独立用户行为分析

从前面的结果中得知中文查询通常较短，而且

中文含义较多，因此搜索引擎很难选择到底该将什么样的结果返回给用户。通过表给出的两个例子，可以看到虽然两个用户都是提交了“仙剑奇侠传”这个查询，但是由于检索目标不同，用户查询的行为也有很大差别。用户 1 只点击了一次结果，而且从点击的 URL 可以看出他是想找有关仙剑这个游戏的相关网站或下载地址；而用户 2 则点击了 13 次之多，而且他点击的 URL 是一些导航类的网页，可以推测他的检索目标是和仙剑电视剧或游戏相关的信息或资讯。而这两种情况基本代表了提交仙剑奇侠传这个查询的用户的两种最主要的检索目标。因此，搜索引擎在返回结果的时候，可以根据这种情况，将这两类相关的网站交错排列放到返回结果的靠前位置，从而较好地满足两种不同用户的查询需求。

表 3　用户查询行为特征比较

用户编号	查询词	点击次数	首次点击的 rank 值	首次点击的 URL
1	仙剑奇侠传	1	6	games. tom. com /zhuanti/pal_3/index. html
2	仙剑奇侠传	13	12	news. 17173. com/zt /0707xj/

5　结论与讨论

在本文中，我们介绍了对搜狗搜索引擎一个月内的真实查询日志的分析情况。结果指出对于 85％的查询，用户只翻看搜索引擎返回结果的第一个页面，这与英文用户行为分析的结果[4] 基本一致。但中文检索用户行为有一些特征是区别于英文检索用户的。例如，中文检索用户提交的查询中只有 0. 73％是使用了高级检索功能，这个比例远低于英文用户行为分析中 20％的比例，说明中文检索用户更注重搜索引擎使用方法的简便，这也跟中国的文化发展特征和现状有关；另外，中文检索用户提交的查询中重复率比英文检索用户高得多，在本文分析的 4 500 万查询中有 4 000 万个查询是重复冗余的，即少数查询出现总数占了总查询数的绝大部分，这说明在中文检索算法中使用缓存机制及人为干预更加必要等。这些结果，都表现出中文检索用户行为的特点，并对中文搜索算法和评测标准的改进都起着非常重要的作用。

用户行为分析作为一种研究检索系统算法设计及性能评价的极为重要的工具，在网络信息检索研究领域具有极其重要的研究价值和广泛的应用背景。随着中文用户搜索需求的扩大，进一步地对中文搜索用户行为特征进行分析也显得更加重要。由于中文的特点，在查询词之间的联系比英文查询词更多，如何进一步利用这些中文搜索用户的特征进一步改善检索算法性能，则是我们今后工作的方向所在。

参考文献：

[1] Cockburn, A., & Jones, S. Which way now? Analyzing and easing inadequacies in WWW navigation [J]. International Journal of Human Computer Studies, 1996, 45, 105 129.

[2] Catledge, L. D., & Pitkow, J. E. Characterizing Browsing Strategies in the World Wide Web [J]. Computer Networks and ISDN Systems, 1995, 27, 1065 1073.

[3] Tauscher, L., & Greenberg, S. How people revisit web pages: Empirical findings and implications for the design of history systems [J]. International Journal of Human Computer Studies, 1997, 47, 97 137.

[4] Craig Silverstein, Monika Henzinger, Hannes Marais, et al. Analysis of a very large Web search engine query log [J]. In SIGIR Forum, fall 1998, Volume 33: Number 1, 6 12.

[5] Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. Real life information retrieval: A study of user queries on the Web [J]. SIGIR Forum, 1998, 32(1): 5 17.

[6] 第 14 次中国互联网络发展状况统计报告[R]. 中国互联网络信息中心(CNNIC)，2004 年 7 月.

[7] 第 15 次中国互联网络发展状况统计报告[R]. 中国互联网络信息中心(CNNIC)，2005 年 1 月.

[8] 第 17 次中国互联网络发展状况统计报告[R]. 中国互联网络中心(CNNIC)，2006 年 1 月.

[9] Danny Sullivan, Search Engine Sizes. In search engine watch website [J], http: //searchenginewatch. com /reports/article. php/2156481.

[10] Andrei Broder, A taxonomy of web search [J]. In SIGIR Forum, fall 2002, Volume 36 Number2.

[11] Ellen M. Voorhees, Donna Harman. Overview of TREC 2001 [A]. E. M. Voorhees and D. K. Harman, eds. In: Proceedings of the tenth Text Retrieval Conference [C]. Gaithersburg: National Institute of Standards and Technology, NIST, 2002, volume 10.

[12] Ellen M. Voorhees. Overview of TREC 2002. E. M [A]. Voorhees and Lori P. Buckland, eds. In: Proceedings of the eleventh Text Retrieval Conference [C]. Gaithersburg: National Institute of Standards and Technology, NIST, 2003, volume 11.

[13] D. Hawking and N. Craswell. Overview of the TREC 2002 web track [A]. E. M. Voorhees and Lori P. Buckland, eds. In: Proceedings of the eleventh Text Retrieval Conference [C]. Gaithersburg: National Institute of Standards and Technology, NIST, 2003.

[14] D. Hawking and N. Craswell. Overview of the TREC 2003 web track [A]. E. M. Voorhees, eds. In: Proceedings of the twelfth Text Retrieval Conference [C]. Gaithersburg: National Institute of Standards and Technology, NIST, 2004.

[15] 国家 863 计划基础资源与评测, 2003 年度信息检索评测大纲, http://www.863data.org.cn/src/863history/2003/2003fulltextretrieval_s.zip.

[16] 国家 863 计划基础资源与评测, 2004 年度信息检索评测大纲, http://www.863data.org.cn/src/2004eval/.

[17] Open Directory Project, http://www.dmoz.org.

《中文信息学报》征稿简则

一、《中文信息学报》主要刊登中文信息的基础理论、应用技术、中文信息处理系统及设备、中文信息的自动输入和人工编码输入、汉字字形信息、自然语言处理、计算语言学及民族语言文字信息处理及网上信息处理等方面的研究论文、技术报告、综述、通讯、简报、国内外学术活动等。

二、来稿要求和注意事项

1. 来稿内容力求正确, 论点明确, 文字简练, 数据可靠, 图表清晰, 字数不超过 8000 字。

2. 投稿要一式三份, 计算机打印, 亦接收电子投稿。文章题目不超过 20 个字, 须有 200 字中文摘要和英文摘要。英文文摘应符合英文语法, 概括论文内容, 包括研究目的、方法、结果和结论。中英文摘要均应包括题目、作者姓名、单位名称、城市名、邮编、摘要、关键词。写明中图分类号。

有基金项目支持的写明基金名称、编号。

给出作者信息, 包括姓名, 出生年, 性别, (学位), 职称, 主要研究方向。

3. 文中图、表放在文稿中相应位置, 并注明图号、图注。图中文字用六号宋体。

4. 文中外文字母、符号要分清大小写、正斜体; 上下角标的位置高低应区别明显; 容易混淆的字母数字, 用铅笔批清(如 O[英大] 0[数字]); 文中需用黑体字之处, 在字下加波纹线。全文计量单位要一致, 或中文, 或符号。

5. 参考文献只列最主要的, 必须是已公开发行的书刊才能列入, 最少不得少于 5 条。文献按文中出现先后次序编排, 书写格式为:

专著: [序号] 作者. 书名[M]. 出版地: 出版者, 出版年

期刊: [序号] 作者(多作者用逗号分开, 超过 3 个者用等). 文章题目[J]. 刊物名称, 年代, 卷数(期数): 起止页码

论文集: [序号] 作者. 题名[A]. 编者. 论文集名[C]. 出版地: 出版者, 出版年, 起止页码

学位论文: [序号] 作者. 题名[D]. 保存地点: 保存单位, 年份

报纸文章: [序号] 作者. 题名[N]. 报纸名, 出版日期(版次)

6. 来稿请勿一稿二投, 文责自负。不录用稿件概不退还, 请自留底稿。来稿一经发表, 按规定付给稿酬, 并赠送单行本 2 册。

来稿请寄: 北京 8718 信箱《中文信息学报》编辑部收, 邮政编码 100080, 电话: 010 62562916。本刊也接收电子投稿, 请以附件方式, 将 WORD 文档发至: cips@iscas.ac.cn。请写明作者工作单位、通信地址(邮政编码)、电话(手机)、E mail。