第 34 卷 第 21 期 计 算 机 工 程 2008 年 11 月
Vol.34 No.21 Computer Engineering November 2008
·软件技术与数据库· 文章编号：1000—3428(2008)21—0040—02 文献标识码：A 中图分类号：TP393

大规模中文搜索日志中查询重复性分析

窦志成[1]，袁晓洁[1]，何松柏[2]
(1. 南开大学信息技术科学学院，天津 300071；2. 军事交通学院汽车指挥系，天津 300161)

摘 要：分析大规模中文搜索日志中的查询重复性，通过对查询重复率和用户个体查询重复率等数据的统计发现：查询串的查询频率、文档的点击频率及用户查询频率均符合 Zipf 分布，查询重复率较高。查询历史越长，查询重复率越高。高查询频率用户的查询重复率较高。以上数据为中文搜索引擎的改进提供了有力的依据。

关键词：搜索引擎；日志分析；重复性；Zipf 分布

Analysis of Query Repetition in Large-scale Chinese Search Log

DOU Zhi-cheng[1], YUAN Xiao-jie[1], HE Song-bai[2]
(1. College of Information Technical Science, Nankai University, Tianjin 300071;
2. Automobile Transport Command Department, Academy of Military Transport, Tianjin 300161)

【Abstract】This paper analyzes query repetition in a large-scale Chinese search engine log. It provides detailed statistics about query repetition and individual query repetition. Key conclusions include: query frequency, document click frequency and user frequency follow Zipf distributions. Queries are with high repetition ratios. Query repetition ratio increases when users' search histories become rich. The users who search more frequently have higher query repetition ratios. These conclusions are useful for improving search performance of Chinese search engines.

【Key words】 search engine; log analysis; repetition; Zipf distribution

搜索引擎日志中记录了用户的查询和点击信息。对搜索引擎日志进行分析，从中挖掘出查询特征和用户行为规律，能够为改进搜索引擎效率、提高搜索精度提供依据和指导方向。随着中文网民数量的增加和中文搜索引擎的发展，对中文搜索引擎中用户查询重复性进行统计分析，成为一项非常有意义的工作。大部分现有搜索日志分析工作[1-7]主要分析查询串长度、查询频率分布、用户平均浏览结果页数、会话长度等，针对查询重复性的分析较为简略，一般仅对日志集上的整体查询重复率进行统计。本文对搜索日志中的重复特征进行了详细的分析，统计了不同历史日志天数下日查询请求中查询重复率的变化、用户个体查询重复率的变化以及不同查询频率的用户的查询重复率分布。

1 相关研究

目前关于 Web 搜索引擎中查询重复特征的研究包括：文献[2]指出搜索引擎 AltaVista 中存在较高的查询重复率，大约有 1/3 的查询串在 6 个星期中被用户重复使用；文献[3]也指出，Web 搜索引擎查询中具有高度的局部性和重复性，少量查询串被大量用户频繁使用；文献[4]发现，用户查询内容和 URL 点击表现出明显的局部性；文献[5-6]对天网中文搜索引擎一天的查询日志分析指出，日志中查询串数量满足 Heaps 定律，少量查询串被频繁查询；文献[7]对搜狗中文搜索引擎 2006 年 2 月份的日志进行了统计分析，发现该日志集中整体查询重复比例高达 91%。这些研究虽然指出了搜索引擎中存在较高的查询重复率，但它们仅对整体查询重复率进行了简单统计。

2 数据集

本文采用搜狗实验室发布的搜索日志集(下载地址为 http://www.sogou.com/labs/dl/q.html)。该日志集包括 2006 年 8 月份 30 天内(数据集中不包括 2006 年 8 月 25 日的日志)搜狗搜索引擎网页查询请求及用户点击记录，包括查询日志的所有点击、查询信息和次序信息，仅滤去了没有点击的查询及涉及不良内容的查询词。该日志中包含的用户行为信息有：(1)用户提交的查询词；(2)用户点击的结果 URL；(3)该 URL 在返回结果中的排名；(4)用户点击的顺序号(即用户点击的第几个页面)；(5)由系统自动分配的用户标识号。日志中不涉及用户的个人信息，如 IP 地址。

在对日志进行统计分析前，本文先对查询日志进行了预处理。用户输入一个查询词，并且点击若干返回结果(可能包括多个页面)的整个过程被识别为一个查询。去除日志中少量错误数据后，该日志数据集中包含点击记录 21 426 131 条，查询 10 812 075 次。详细的统计数据见表 1。

表 1 日志集基本统计信息

统计项	值	统计项	值
日志天数	30	用户数	5 130 767
查询次数	10 812 075	查询串个数	3 118 907
点击次数	21 426 131	文档(URL)数	8 621 580

3 查询重复性研究

3.1 查询串、文档与用户频率分布

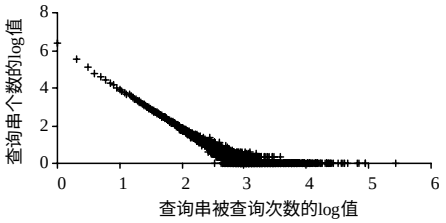
查询频率分布在一定程度上决定了查询重复特征。图 1(a)为查询串查询频度分布，其中，X 轴为查询串被查询的次数；

基金项目：天津市科技发展计划基金资助项目(06YFGZGX05700)；天津市应用基础研究计划基金资助项目(07JCYBJC14500)

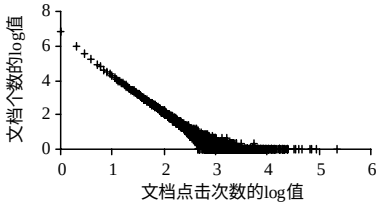
作者简介：窦志成(1980－)，男，博士研究生，主研方向：Web 信息检索，数据挖掘，日志分析；袁晓洁，教授、博士生导师；何松柏，讲师

收稿日期：2007-12-20 **E-mail：**douzc@hotmail.com

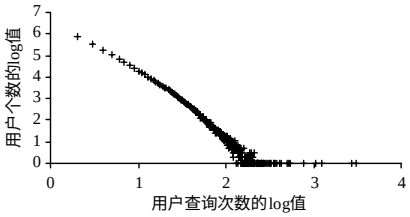
Y 轴为被查询了指定次数的查询串数。可以看出，查询串的查询频率是符合 Zipf 分布的。大多数查询串具有很低的查询频率，小部分高频查询串被频繁重复使用。图 1(b)显示了文档点击次数分布。和查询串查询频率分布类似，文档点击频率同样符合 Zipf 分布，少数文档被大量用户频繁点击。图 1(c)显示在 30 天的搜索日志中，用户查询次数也接近于 Zipf 分布。以上分析说明，搜狗搜索引擎中查询串的查询频率、文档的点击频率及用户的查询频率基本上都符合 Zipf 分布，因此，该日志集中必然存在较高的查询重复现象。



(a)查询串查询频率分布



(b)文档点击次数分布



(c)用户查询次数分布

图 1　查询串查询频率、文档点击频率及用户频率分布

3.2　查询重复率

如果一个查询使用的查询串曾经被用户自己或其他用户使用过，则认为这次查询是一个重复的查询。表 2 给出了日志集中不同查询次数的查询串的总数及其所占比例。结果显示，在全部 30 天的日志记录中，大约 21.87%的查询串被至少提交过 2 次，这些重复查询串对应的查询次数约占总查询次数的 77%，其中，重复查询占总查询次数的 71.15%。

表 2　查询串查询重复统计

查询次数	查询串个数	总查询次数	重复查询次数
1	2 436 699(78.12%)	2 436 699(22.54%)	0(0.00%)
≥ 2	682 208(21.87%)	8 375 376(77.46%)	7 693 168(71.15%)
≥ 5	186 655(5.98%)	7 147 868(66.11%)	6 961 213(64.38%)
≥ 10	82 256(2.64%)	6 481 311(59.95%)	6 399 055(59.18%)
≥ 50	14 188(0.45%)	5 185 164(47.96%)	5 170 976(47.83%)

在实际应用时，历史查询日志一般被收集起来，经过处理后作为训练集。实时记录下来的用户查询日志一般不能马上整理及应用。假设实际应用中以天为单位处理查询日志，即当日的查询日志可于次日完成整理并使用。第 N 天的查询中使用的查询串在第 1~第 N-1 天的日志(即历史日志)中出现，才被认为是重复查询；仅在第 N 天当天重复使用的查询，不认为是重复查询。

图 2 显示了日志中日查询重复率的分布。例如 2006 年 8 月 3 日(横坐标值为 2)有 60.64%的查询所使用的查询串(占当天所有查询串的 22.49%)曾经在 2006 年 8 月 1 日或 8 月 2 日被用户自己或其他用户使用过。因为查询日志从 2006 年 8 月 1 日开始，所以 8 月 1 日这天没有重复查询。

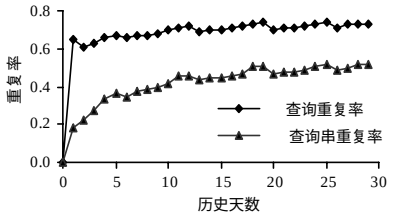


图 2　每日查询串及查询重复率

从图 2 看出，查询重复率与日志集日期长短有关。总体上，历史日志日期越长，新查询的重复率越高。当仅有一天历史查询日志时，查询的重复率超过了 60%(重复查询串的比例为 17.88%)。随着历史日志增加，每天的重复率在慢慢增加。在具有 29 天的历史查询日志时，第 30 天的查询重复率约为 73%，重复查询串的比例超过 50%。按这种方式计算，这 30 天的平均查询重复率为 70.09%，略低于表 2 中给出的 71.15%。这是因为查询串首次出现后，该查询串出现当日的重复查询并没有被计算。如果应用系统中仅保存 30 天的查询日志，30 天之前的查询日志被清空，则每天新达到的查询串中大概超过 50%是在 30 天的查询历史中存在的；每天用户提交的查询中将有超过 70%是重复查询。如此高的查询重复率说明可以利用历史查询日志帮助用户重新查找查询过的信息。图 2 同样说明，如果保留 30 天的查询历史，则搜狗搜索引擎每天新增查询串的比例约为 50%。

3.3　用户个体查询重复率

如果一个查询中使用的查询串曾经被用户自己使用过，则称这个查询为个体重复查询。统计发现，在全部 30 天的 10 812 075 次查询中，945 590 次(约 8.7%)为个体重复查询。与日重复查询比例计算方式类似，图 3 给出了每日个体重复查询比例。当只有一天查询历史时，仅有 2.5%的查询为个体重复查询。重复比例随历史查询天数的增长而增加，当有 29 天查询历史时，用户查询重复比例达到了 8%。以上统计数据说明，日志中存在较高的用户个体查询重复比率，但相对于整体查询重复率，个体查询重复比例略低。

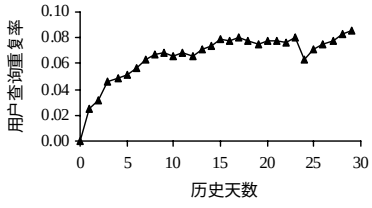


图 3　每日个体查询重复率

考虑到不同查询频率的用户其查询重复比例可能不一致，本文统计了不同查询频率的用户的平均查询重复率。图 4 给出了用户查询频率与查询重复率的关系。可以看出，查询频率较低的用户，其平均查询重复率也较低；用户查询频率越高，其查询重复率也越高，但增长趋势变缓。如在 30 天中，查询次数为 10 次的用户平均查询重复率约为 13%；查询次数为 40 次的用户平均查询重复率约为 28%；当用户查询频率增加到 100 次左右，其查询重复率逐渐稳定，接近于

（下转第 44 页）