Web搜索引擎日志挖掘研究框架*

□ 王继民 李雷明子 / 北京大学信息管理系 北京 100871
孟涛 / 北京大学信息科学技术学院 北京 100871

摘要：搜索引擎日志记录了用户与系统交互的整个过程。对日志文件进行挖掘，可以发现用户进行Web搜索的行为特征与规律，有效改善搜索引擎系统的性能。在对国内外相关研究进行系统梳理和总结的基础上，文章提出了一个Web搜索引擎日志挖掘的研究框架，主要包括日志挖掘的研究内容、数据集的选择方法、数据预处理的方法、不同地域用户行为的特征与比较、如何应用于系统性能的改善等内容。

关键词：搜索引擎，用户日志，Web使用挖掘，用户搜索行为
DOI：10.3772/j.issn.1673—2286.2011.08.007

1 引言

搜索引擎（Search Engine）是Web上的一种应用软件系统，它以一定的策略发现和搜集Web网页信息，进行处理和组织后，为用户提供信息查询服务[1]。在2003年，全球约有3200多个分布于211个国家的各类Web搜索引擎[2]。随着Web信息量的快速增长，各种综合性大型或面向主题的小型搜索引擎的数量也在持续不断地增加。

用户与搜索引擎的交互过程可简述为：用户在查询框内输入一个查询串（query），经搜索引擎内部进行分析和处理后得到几百甚至上万条相关记录，每若干条记录（如10个记录）组成一个查询结果页面，其中每条记录都代表着一个网页（文档）入口，它包含了该文档的标题、在Web上的位置（即网址，Uniform Resource Locator，URL）、网页内容摘要等信息。用户可以由此判断该记录所指向的网页是否包含自己感兴趣的内容，并决定是否点击该URL进行详细浏览。

搜索引擎日志记录了用户与系统交互的所有信息，不同搜索引擎的日志记录格式略有不同，但一般都包括用户的访问时间、用户的IP地址、输入的查询串、用户所点击的URL、点击的时间以及点击URL的序号等。这些信息通常按某种格式存入磁盘的文件系统中。

搜索引擎日志挖掘是Web使用记录挖掘（Web usage mining）的一种，它从用户的查询记录中抽取有意义的模式，主要包括研究用户如何使用Web搜索引擎；研究用户在Web上查找何种内容的信息；研究群体或单个用户的查询行为特征、规律及其演化趋势； 研究不同地域或不同主题搜索引擎的用户查询行为的异同； 以及如何利用日志分析改进系统的性能等。

在对国内外搜索引擎日志挖掘研究的理论、技术、方法与实证研究进行系统的分析和总结的基础上，本文提出了对该领域进行研究的一般框架，主要包括：数据集的选择方法、数据预处理的方法（第2节）；可从日志文件中挖掘的主要内容及其主要结果，不同地域的用户查询行为特征的比较分析（第3节）；利用日志挖掘提高搜索引擎系统性能的主要方法（第4节）；该领域的研究前景

* 本研究得到国家社会科学基金（10BTQ050）、教育部人文社会科学基金（09YJA870002）和核高基项目（2011ZX01042—001—001）的资助。

（第5节）等。

2 数据集与数据预处理

2.1 数据集

2.1.1 数据格式

许多大型搜索引擎系统将用户的查询与点击记录分开保存，即系统的日志文件由用户查询日志和用户点击日志组成。用户查询日志是在用户提交查询请求时记录的，它包括用户查询时提交的查询串、提交时间、用户IP地址、页号（查询结果分页显示，每页显示10个查询结果，用户首次查询页号为1，用户翻页时的页号即为用户选择的结果页面号）等信息。以北大天网搜索引擎的用户日志文件为例，用户查询日志的一个简单的记录格式为：

Fri Mar 11 10:36:02 2005 // 提交时间
162.105.146.* // 用户IP
Database // 是否在缓存中命中
北大 // 查询串
1 // 页号

用户点击日志是用户浏览查询结果并点击页面时记录的，它包括用户点击页面的时间、点击页面的URL、用户IP地址、点击页面的序号（该页面在查询结果中的位置）、该点击对应的查询串等信息。天网用户点击日志的一个简单的记录格式为：

Fri Mar 11 10:36:02 2005 // 点击时间
162.105.146.* // 用户IP
北大 // 查询串
http://www.pku.edu.cn // 点击的URL
2 // 点击页面的排序

2.1.2 数据集的选择

根据研究目的的不同，通常要选取不同的数据集进行分析。但基于商业竞争的考虑，主流商业搜索引擎系统一般不愿提供或是不愿完整地提供自己的日志数据，这在一定程度上制约了搜索引擎日志挖掘的研究。

目前公开发表的学术论文中，数据集的选取差异很大，主要表现在时间跨度上：选择1天的数据进行研究和分析的最多，如文献[3,4]；条件允许时，也可选择1周、几个月甚至几年的数据进行研究与分析，如文献[5,6]。

通常而言，短期内群体用户的访问规律、查询内容与点击URL等行为方式基本类似。例如用户的查询量、点击量和不同用户的访问量可用时间序列中的潜周期模型来描述[7]；用户的查询内容与点击URL的过程具有自相似性的特征[8,9]；因此，若研究某一时段（如一个月）用户查询行为的一般特征，时间跨度选取的影响较小，一周（甚至一天）的日志数据就够了。

2.2 特定术语

关于搜索引擎日志挖掘的研究，需要事先定义一些特定的术语。目前，如下的几个概念被广泛使用。

词项（term）：不含分隔符的一个字符序列；这里的分隔符包括逗号、句号、冒号、空格符等事先指定的符号。词项的例子如“search”或“搜索”等。分隔符的选择直接影响词项的计数结果。日志分析中以空格符作为分隔单位的居多。

查询串（Query）：用户在搜索框内输入的查询内容，由一个或多个词项组成。例如“search engine”或“中文 搜索”等。查询串中可能包括某种逻辑操作，如and、 or、 not等。对同一个用户的某次查询，首次输入的查询串称为初始查询串（initial query）；若随后的查询串等同于先前输入的某一查询串则称之为重复查询（repeat query）；若随后输入的查询串不同于先前的查询串则称之为修正查询（modified query）。

会话（Session）：单个用户在一段时间间隔内所提交的整个查询串序列，其中查询串的个数定义为会话长度。区分用户会话的时间间隔可以是若干分钟、若干小时或1天，典型的如5分钟、15分钟、30分钟或1天等。用不同的时间间隔进行会话分割，统计结果会有差异。但从用户日志角度来看，会话可能由单个用户、共同用户或者程序抓取而产生。

2.3 数据预处理方法

原始日志文件可能存在不完整的或不一致的噪音数据，因此需要在数据分析或模式挖掘之前进行数据预处理，主要包括数据清理、用户识别、会话识别、英文词干提取或中文分词等工作。进行有效的数据预处理可以提高挖掘模式的质量，降低挖掘所需要的时间。

（1）数据清理：删除日志中与挖掘任务无关的数

据，例如删除用户误操作所导致的空查询串，以及根据需要删除标点符号和查询串中多余的空格等。

（2）用户识别：通常用IP地址来区分不同的用户，但由于本地缓存（Cache）、代理服务器和防火墙的存在，仅从用户日志还无法确定来自某一IP的查询是否为真正的单用户。通常靠用户会话长度的大小来识别部分用户，例如删除一天内来自同一IP且查询次数超过某一阈值（如200次）的全部记录。

（3）会话识别：将一个用户的访问记录分为若干个单一的会话。一般利用用户对系统访问的时间间隔进行会话识别，例如，当一个用户的两次查询请求的时间间隔超过某一设定阈值（如30分钟），则认为用户开始了一个新的会话。

（4）英文词干提取或中文分词：在英文信息处理中，可以利用词干提取技术减少词语空间的大小。所谓词干是指将词的词缀（前缀或后缀）删除后剩余的部分，如“compute”是“computer”和“computing”的词干。由于中文词与词之间没有分界符，所以需要人为切分；不同切词软件由于其采用的分词算法不同，其分词结果略有差异。

（5）大小写英文字母转化：由于多数搜索引擎不区分大小写英文字母，挖掘某些模式时需要将查询串中的大写英文字母全部转化为小写，这有利于查询信息的聚集与合并。

3 挖掘的主要内容及其结果

搜索引擎日志挖掘的主要技术和方法包括：统计分析方法、建模分析与预测、序列模式发现、关联规则挖掘、聚类分析等；挖掘的主要内容包括：词项级、查询级和会话级的数据分析、用户结果页面的查看和点击URL的特征、用户查询行为的演化趋势、不同地域用户查询行为的比较，以及如何利用日志分析改进搜索引擎系统的性能等。

3.1 主要统计指标

对搜索引擎用户日志可以挖掘的模式有很多层面，目前已发表的研究论文中所包含的主要统计指标及结果如下。

（1）词项级（term level）：对单个词项的使用情况进行统计分析，包括词语本身及多语言的使用情况以及词项的误拼写情况等。例如我们考察了词项中所包括的中文、英文、中英文混合、纯数字的使用情况等[4]。更深层次的研究结果包括：英文词项的频次频级分布符合power-law分布（或类Zipf分布）的特征[10,11]。

（2）查询级（query level）：研究用户输入的查询串中所包含的词项个数，即查询长度；以及查询的复杂性，即用户使用布尔操作（AND、OR、NOT）或短语查询的情况。主要的研究结果[2,4,11-13]包括：英文搜索引擎的输入的查询串平均包含2.2到2.4个英文单词，多数为两个英文单词，查询串中所包含的英文单词的数量服从Poisson分布。多数中文用户输入的查询串中只含有一个词项并且包含中文字符，其中以2至4个汉字居多。Web搜索引擎用户使用复杂查询的比例较小。

（3）会话级（session level）：研究用户会话的长度、用户进行查询修订或重复查询的使用情况以及用户提交查询的时间间隔等。主要的结果[5,11,13-15]包括：多数用户会话只含有一个查询，少量用户进行查询修正；绝大多数用户的会话时间小于15分钟等。

（4）结果页面查看（results pages viewed）：研究用户查看结果页面的个数（如翻页等）、查看网页快照的情况，以及查看结果页面的时间间隔。主要统计结果[2,13,16]包括：绝大多数搜索引擎用户查看较少的结果页面，通常为1-2个；查看结果页面的时间间隔在2~3分钟之间；用户查看网页快照的比例较小，如天网数据分析显示：点击网页快照的数量只占总点击量的3.5%。

（5）点击URL（hit URL）：研究一次会话或一次查询中用户所点击结果页面中URL的个数、序号以及相关性等。主要结果[2,6,9,11]包括：用户点击不同URL的数量遵从Heaps定律，点击URL的频度频级服从类Zipf分布，点击URL与页面大小相关，点击URL具有时间局部性，其点击过程具有自相似性特征等。

3.2 不同地域用户查询的特征及比较

地域、文化背景和语言使用上的不同，可能导致用户群的查询行为方式以及查询内容上的不同。根据搜索引擎的主要用户群所在地进行划分，目前已被分析的搜索引擎日志约有10余个。

美国：Excite[17]；AltaVista[5]

南美洲：TodoCL[18]（智利）

欧洲：AlltheWeb[19]（挪威）；BWIE[20]（西班

牙）；Fireball[21]（德国）

亚洲：NAVER[6]（韩国）；GAIS[13]（中国台湾）；TianWang[4]（中国大陆）

尽管上述各文献所选取的搜索引擎日志的时间段不同和数据集的大小也有很大的差异，所采用的数据预处理方法也不尽相同，但它们都各自反映了某一地域和文化背景下的搜索引擎用户行为的一些基本特征。对这些研究结果进行对比和分析，我们可以发现它们的一些共同点：

（1）用户输入的查询串一般比较短，通常为1~3个词项，其中英文以及其他欧洲语言为两个单词，而中文、朝鲜语与西班牙语为1个词项。

（2）超过一半的用户每次只进行一次查询，且查看较少的结果页面，通常只查看1~2个结果页面。

（3）用户的查询结构比较简单，一般不使用高级检索功能。

进一步，我们还可以发现不同搜索引擎的用户行为在某些统计指标如平均每个查询串包含的词项个数、只查看1个结果页面的比例等相差较大。这启示我们不能将一个搜索引擎日志分析的结果完全推广到另一个搜索引擎上，特别地，对多地域服务的大型搜索引擎如google、yahoo!等，在系统设计与服务上应根据地域的不同而选择不同的策略。用户查询行为趋向简单化，就要求我们在界面的设计上尽量做到用户容易使用。

3.3 深度挖掘

对搜索引擎日志的深度挖掘主要包括用户查询的多任务性、用户查询的演化趋势、用户访问时间的分布等。这些研究需要一些其他的领域知识，如查询主题的分类、时间序列分析、数学模型的构建等。其他的深度挖掘研究还包括查询串中词语的共现情况[22]、发现热点查询事件[23]、命名实体识别[24]，以及具体到某一特定主题的用户查询行为研究，如多媒体查询、医药与健康查询、色情查询和经济类的信息查询等[2]。

3.3.1 用户查询的多任务性

用户的信息查询行为可能包含多个主题（多任务），例如用户先后查找“计算机”类和“娱乐”类的信息。文献[25]的研究显示：用户在不同信息环境（问卷调查、Web查询、在线数据库、学术图书馆）下都存在多主题信息查询。文献[3,25-27]分析了Excite和AlltheWeb这两个分别以美国和欧洲用户为主的搜索引擎的多任务查询的特征，结果显示每个多任务会话平均包括3个不同的主题，每个主题提交4~5个查询等。我们对2002年的天网中文搜索引擎的多任务Web查询进行了研究和分析[16]，结果显示：多于1/3的用户进行多任务Web查询；超过1/2的多任务会话包含两个不同的主题并进行2~7次查询；多任务会话时间的均值是一般会话时间均值的两倍；天网用户的多任务查询主要有三个主题：计算机、娱乐和教育，近1/4的多任务会话中包含不确定的信息。

多任务Web查询是人们进行信息查询时的一种常见模式，它揭示一些用户在有几个信息需求时才进入Web搜索引擎系统进行信息查询。多任务Web查询具有较复杂的查询模式，需要更有效的检索技术为如此复杂的查询结构提供服务。

3.3.2 用户查询的演化趋势

单个用户的信息需求在不断地变化，用户群在搜索引擎上的查询主题也在不断地迁移。为分析这种迁移，Spink曾抽样选取了Excite搜索引擎在1997、1999和2001年各2000多个用户输入的查询串[17]，进行手工分类（定义了11类）后发现：人们的信息需求正在“from e-sex to e-commerence”；也就是说在Excite的Web用户查询中，商业信息需求逐渐增加，娱乐类查询相对减少。

我们对2001-2005年的天网用户日志进行抽样分析[16]，结果显示：用户输入的查询串中所包含词项数量有明显增多的趋势；用户会话的长度逐年下降；用户查看的结果页面越来越少；查看的时间间隔逐渐缩短；查询串中所包含的汉字个数基本稳定，包含2-4个汉字的查询串居多；在查询结果中发生点击行为的比率呈递减的趋势；查询次数与点击次数的相关性逐渐减弱；Web用户查询的主题变化较快。

3.3.3 用户访问时间的分布

掌握用户对系统访问时间的分布有利于系统资源的配置。我们曾对天网用户对系统的访问时间的分布进行统计分析[7]，结果显示了用户访问具有极强的规律性，一天中用户到达的时间出现三个波峰，早

晨 10：30，下午 4：30和晚上 8：30，用户的最少访问量发生在凌晨 3：00—7：00，这与CNNIC的用户上网时间类似。进一步研究表明：短期内（如一两个月）天网用户每个工作日（周一至周五）的访问量基本相同，时间分布也非常类似，与公休日（周六和周日）的到达时间略有差异，整体呈现周期性波动， 用户的访问量可用时间序列中的潜周期模型来描述。

4 应用于搜索引擎系统性能的改善

对搜索引擎日志进行挖掘，可以有效地改善搜索引擎系统在效果、效率、服务等方面的性能。具体地，在系统效果方面，我们可以利用用户查询、用户点击URL等反馈信息，提高结果排序的质量；在系统效率方面，使用Cache（缓存）替换策略可以有效改进系统的应答速度；在系统服务方面，我们可以发现给定查询的一些相近Web查询，并提供查询推荐服务等。

4.1 提高结果排序的质量

在用户提交一个查询请求之后，如果点击了系统返回结果页面中的某个URL，一般表示用户对该URL的一个认可，并且多数情况下所包含相关信息差的URL不被点击。2002年Zhang Dell提出了一种利用用户点击记录提高结果排序质量的方法[28]，该方法首先应用于中文图像信息检索，随后Baeza-Yates将其推广应用于文本信息检索[29]，取得了不错的实验效果。Zhang称这种方法为MASEL（Matrix Analysis on Search Engine）算法。

MASEL方法试图寻找用户、查询与点击URL之间的关系。其基本假设是：好的用户提交好的查询、好的查询返回好的URL、好的URL被好的用户所点击。根据用户点击记录递归地定义这三个基本量，这一思想非常类似于Hits算法中页面的Authority和Hub之间所建立的递归关系[30]。

Baeza-Yates对智利Todo搜索引擎日志进行小规模实验时发现MASEL方法具有良好的效果，并且指出：日志周期的选取对实验结果的精度有一定的影响，特别地，对于多义词（查询的词项），当选取较长时间段的日志时其结果精度反而下降。

针对元搜索引擎的结果排序，Joachims提出了一种以点击数据作为训练集，学习检索函数的方法[31]；其实验的排序结果优于Google的检索结果。

4.2 Cache的替换策略

在搜索引擎系统的检索端或索引端使用Cache，可以大大改善用户查询的平均响应时间，提高系统的效率。用户查询分布的统计分析表明用户的查询是非常集中的（即查询的局部性特征），这揭示了我们在查询中使用Cache的可行性，把这些查询次数较高的词项的查询结果放在Cache中，使用容量很小的Cache就能命中大部分的用户查询，这样就可以用较小的空间取得较大的Cache命中率。

Xie在论文[10]中首先讨论了搜索引擎系统使用Cache可以有效降低服务器负载，减少系统应答时间。文献[8]对天网搜索引擎检索端Cache替换的几种策略进行了对比，结果显示LRU（Least Recently Used）和LFU（Least Frequently Used）的Cache命中率要明显好于FIFO（First In First Out），如果给LFU固定了一个衰减因子，其效果和LRU相差不多，如果选取好的衰减因子，可以得到比LRU稍微好一些的效果。考虑到实现的复杂性，LRU和FIFO都比较简单，而LFU在发生替换的时候要进行衰减，必须遍历整个Cache，其替换时间要远远大于LRU和FIFO，而其效果和LRU相差不是很多。所以该文献综合认为LRU是最好的Cache替换策略。

4.3 发现相关Web查询

由于搜索引擎用户输入的查询串通常比较短，而短词语所能表达的主题比较宽泛，容易存在歧义，并且用户时常不能准确地表达自己的信息需求，因此有时用户需要对自己的查询请求进行不断的修正，以求找到自己满意的信息。为方便用户进行查询修正，一些搜索引擎系统如Google、百度等已经在用户第一次提交查询请求后，在系统返回的结果页面中包含了一个相关查询列表，供用户进行查询修正时参考，这为用户更精确地表达自己的信息需求提供了便利。

传统的信息检索系统利用查询扩展来发现相关查询，主要方法有：基于用户反馈作查询扩展、基于局部或全部信息作查询扩展等[32,33]。这些方法一般依赖于文档集中各个文档的具体内容，实现上较为复杂，在实际的检索系统中使用并不多。

基于搜索引擎用户日志发现相关查询是一个实际可行的方法。目前的研究通常要考虑两种基本因素的影响[16, 34]：（1）如果两个查询串包含了相同词项，则它们可能是相关的，并且所含的共有词项越多，相似性就越高。（2）对两个不同的查询串，如果用户点击了查询结果中的相同URL，则它们可能是相关的。除此之外，其他的影响因素还包括：被查询次数，不同用户的查询分布，被点击相同URL的次数，点击URL对应文档间的类别相似性等。

随后，我们可以利用某种方法组合这些影响因素进而发现一给定查询的相关查询；方法之一是手工标记部分训练数据，建立回归模型，以相关度的大小确定相关Web查询[16]。通常选用的回归分析方法是多元线性回归和支持向量回归。一般来讲，线性回归方法的参数计算简单，能很快得到每一查询的预测结果，但预测精度略差一些；支持向量回归方法涉及较多的参数选择，训练时间较长，结果预测的精度较高。研究发现：对不同类型的查询（信息型、导航型、事物型[35]），其预测结果的准确度存在较大的差异。

仅通过用户点击日志，对查询串进行聚类是另一种发现相关查询的方法；该方法首先构造一个二部图，即依据用户的点击记录，连接查询集合与点击URL集合中的一些对应元素；然后用凝聚的（Agglomerative）迭代算法进行聚类，依次合并两个查询和两个URL直至迭代结束[36]。该方法的一个不足之处是不能有效处理“噪音”数据，即如果用户误点了某个URL，那么两个不相关的查询就可能被永远地聚在了一起[37]。

还可以利用关联规则确定相关Web查询[38]，具体地，将查询看作关联规则中的项（item），将查询日志中的会话看作事务（transaction），即在一定时间间隔内单个用户提交查询的集合为一个事务。然后利用关联规则挖掘算法，找出强关联规则，进而发现相关的Web查询。

5 结语与展望

基于搜索引擎日志发现用户Web查询的特征与规律，不仅能够改善搜索引擎的系统性能，而且对用户的信息行为研究具有重要的意义。大型的商业搜索引擎都很重视对系统日志挖掘的研究，如百度公司与北京大学合作成立了“中国人搜索行为研究室”，搜狐公司与清华大学合作成立了“搜狐搜索技术联合实验室”等。从事该领域的研究工作需要用到信息科学、计算机科学、数据挖掘、人工智能、人机交互、教育心理学、认知科学等各方面的知识。就目前公开发表的学术论文来看，这一领域的研究成果以英文搜索引擎的研究为主，欧洲、亚洲等区域的研究相对较少。

本文针对Web搜索引擎日志这一特定的数据集，提出了对其进行挖掘的一般流程（框架），贯穿了日志挖掘的整个过程，既涉及所使用的理论、技术与方法，也归纳总结了目前已有的研究成果。该框架的建立可以指导搜索引擎及其类似Web日志挖掘的研究等。

在该研究领域内仍有许多问题值得深入研究，其中的一个重点仍将是如何利用日志挖掘的结果进一步改善搜索引擎系统的性能；其他的一些研究主题包括：如何划分用户会话比较合理？不同地域用户的搜索行为与地域文化之间具有什么样的关系？如何利用用户查询行为的演化规律评估搜索引擎系统的性能？如何根据访问日志预测用户的信息需求？在复杂环境下与实验室环境下（或问卷调查的结果），用户的查询行为具有什么样的关系？如何发现并识别不同类型如不同性别、年龄、知识背景下的用户查询行为特征？在工作日与休息日这两个不同的时间段，用户的查询内容及行为特征有何异同？如何利用认知心理学的一些模型解释用户的各种信息查询行为？综合利用多领域知识对搜索引擎日志进行深度挖掘仍有许多挑战性的工作有待我们去研究。

参考文献

[1] 李晓明,闫宏飞,王继民. 搜索引擎原理、技术与系统[M]. 北京:科学出版社,2005.
[2] SPINK A, JANSEN B J. Web search : public searching on the Web [M]. Netherlands: Kluwer Academic Publishers, 2004.
[3] OZMUTLU S, SPINK A, OZMUTLU H. A day in the life of Web searching: an exploratory study [J]. Information Processing and Management, 2004,40:319-345.
[4] 王继民,陈翀,彭波. 大规模中文搜索引擎的用户日志分析[J]. 华南理工大学学报, 2004,32(S):1-5.
[5] SILVERSTEIN C. Analysis of a very large altavista query log [J]. ACM SIGIR Forum, 1999,33(1):6-12.
[6] PARK S, LEE J H, BAE H J. End user searching: A Web log analysis of NAVER, a Korean Web search engine [J]. Library & Information Science Research, 2005,27:203-221.

[7] 王继民,彭波. 搜索引擎用户访问量模型[J]. 计算机工程与应用,2004,40(25):9-11.
[8] 王建勇,李晓明,等. 海量web搜索引擎系统中用户行为的分布特征及其启示[J]. 中国科学(E),2001,31(4):372-384.
[9] 王继民,彭波. 搜索引擎用户点击行为分析[J]. 情报学报, 2006,25(2):154-162.
[10] XIE YINGLIAN, O'HALLARON D. Locality in Search Engine Queries and Its Implications for Caching [C]// Proc. IEEE Infocom 2002, New Jersey: IEEE Press, 2002:1238-1247.
[11] BALDI P, FRASCONI P, SMYTH P. Modeling the Internet and the Web, probabilistic methods and algorithms [M]. John Wiley, 2003.
[12] JANSEN B J, SPINK A, SARACEVIC T. Real life, real users, and real needs: a study and analysis of user queries on the web [J]. Information Processing and Management, 2000,36:207-227.
[13] JANSEN B J, SPINK A. How are we searching the World Wide Web? A comparison of nine search engine transaction logs [J]. Information Processing and Management, 2006,42:248-263.
[14] LAU T, HORVITZ E. Patterns of Search: Analyzing and Modeling Web Query Refinement [C]// 7th Int. Conf. on User Modeling, Springer, 1999.
[15] HE D, GOKER A. Detecting session boundaries from Web user logs [C]// 22nd Annual Colloquium of IR Research 2000, UK: Cambridge, 2000.
[16] 王继民. 中文搜索引擎日志挖掘研究[R]. 北京大学信息科学技术学院博士后研究工作报告,2005.
[17] SPINK A, JANSEN B J, et al. From e-sex to e-commerce: Web search changes [J]. IEEE Computer, 2002,35(3):133-135.
[18] BAEZA-YATES R. Applications of Web Query Mining [C]// European Conference on Information Retrieval (ECIR'05), Spain: Springer,2005.
[19] SPINK A, OZMUTLU S, et al. US versus European Web searching trends [J]. SIGIR Forum, 2002,32(1):30-37.
[20] CACHEDA F, VINA A. Understanding how people use search engines: a statistical analysis for e-business [C]// Proceedings of the e-Business and e-Work Conference and Exhibition, Italy, 2001.
[21] HOELSCHER C, STRUBE G. Web search behavior of internet experts and newbies [J]. International Journal of Computer and Telecommunications Networking, 2000,33:337-346.
[22] WANG P, BERRY M W, YANG Y, Mining longitudinal Web queries: Trends and patterns [J]. Journal of the American Society for Information Science and Technology, 2003,54:743-758.
[23] GU Y Q，CUI J W, et al. Detecting Hot Events from Web Search Logs[C]// Proceedings of 11th International Conference on Web-Age Information Management, China, 2010.
[24] XU G, YANG S H, et al. Named Entity Mining from Click-Through Data Using Weakly Supervised Latent Dirichlet Allocation[C]// Proceedings of 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2009.
[25] SPINK A, OZMUTLU H C, OZMUTLU S. Multitasking information seeking and searching processes [J]. Journal of the American Society for Information Sciences and Technology, 2002,53(8):639-652.
[26] OZMUTLU S, OZMUTLU H C, SPINK A. Multitasking Web Searching and Implications for Design [C]// ASIST'03: Annual Meeting of the American Society for Information Science and Technology. Long Beach, CA, 2003.
[27] OZMUTLU S, OZMUTLU H C, SPINK A. A Study of Multitasking Web Searching [C]// IEEE ITCC'03: International Conference on Information Technology: Coding and Computing, Las Vegas, NV, USA, 2003.
[28] ZHANG D, DONG Y. A novel Web usage mining approach for search engines [J]. Computer Networks, 2002,39:303-310.
[29] BAEZA-YATES R. Query Usage Mining in Search Engines [M]// SCIME A. Web Mining: Applications and Techniques. Idea Group, 2004:307-321.
[30] CHAKRABARTI S. Mining the Web: Discovering Knowledge from Hypertext Data [M]. Morgan-Kaufmann, 2003.
[31] JOACHIMS T. Optimizing search engines using clickthrough data [C]// Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), 2002.
[32] BAEZA-YATES R, RIBEIRO-NETO B. Modern Information Retrieval [M]. Addison-Wesley-Longman, 1999.
[33] CUI HANG, WEN JI-RONG, et al. Query Expansion by Mining User Logs [J]. IEEE transactions on knowledge and data engineering, 2003,15(4): 829-839.
[34] WEN JI-RONG, NIE JIAN-YUN, ZHANG HONG-JIANG. Query clustering using userlogs [C]// Proceedings of the 10th World Wide Web conference. New York: ACM Press, 2001.
[35] KANG I H, KIM G. Query Type Classification for Web Document Retrieval[C]// Proceedings of the Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR, 2003.
[36] BEEFERMAN D, BERGER A. Agglomerative clustering of a search engine query log[C]// Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 2000.
[37] CHAN W S, LEUNG W T, LEE D L. Clustering Search Engine Query Log Containing Noisy Clickthroughs [C]// Proceedings of the 2004 International Symposium on Applications and the Internet (SAINT'04), 2004.
[38] FONSECA B M, GOLGHERET P B, et al. Using association rules to discovery search engines related queries [C]// First Latin American Web Congress (LA-WEB'03). Santiago, Chile, 2003.

作者简介

王继民（1966-），北京大学信息管理系副教授，博士，研究方向：搜索引擎与Web挖掘。E-mail: wjm@pku.edu.cn

A Research Framework of Web Search Engine Usage Mining

Wang Jimin, Lilei Mingzi / Department of Information Management, Peking University, Beijing, 100871
Meng Tao / School of Electronics Engineering and Computer Science, Peking University, Beijing, 100871

Abstract: Log files of search engines record the interactive procedure between users and the system completely. Mining the logs can help us to discover the characteristics of user behaviors and to improve the performance of search systems. This paper gives a framework on Web search engine usage mining, which includes the choice of data collections, the methods of data preprocessing, and an analysis and comparison of search behaviors from different countries. We also explore its applications on improving the effectiveness and efficiency of search engines.

Keywords: Search engine, User log, Web usage mining, User search behaviors