基于网页浏览日志的用户行为分析

郭俊霞 高 城 许南山 卢 罡
(北京化工大学信息科学与技术学院 北京 100029)

摘 要 随着问答社区信息的长期积累,越来越多的过时信息充斥在其中并被搜索引擎检索,给信息需求者带来不便。用户的网页浏览日志中隐性地包含用户的行为习惯,通过分析得到这些信息对判断网页信息时效性有着重要意义。文中提出针对网页浏览日志的查询过程划分方法,并在划分的基础之上对大量真实用户的浏览行为习惯做了统计分析。结果显示,用户查询一次信息平均浏览8.05个页面,用时6.28分钟,有将近1/3的查询在交替并发中进行,另外用户对于网站站内搜索的依赖较高。从浏览日志数据集中选取了一个社区网站的浏览记录来进行初步的网页信息时效性分析,结果表明造成用户不满意的原因主要是查询相关度不高,而过时信息只是其中一小部分。

关键词 网页浏览日志,用户行为分析,网页时效性,问答社区
中图法分类号 TP391.1 **文献标识码** A

User Behavior Analysis Based on Web Browsing Logs

GUO Jun-xia GAO Cheng XU Nan-shan LU Gang
(College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract With the long-term accumulation of the Q & A community information, there is more and more outdated information indexed by search engines, bringing inconvenience to users. The log of a user's browsing-behaviors contains the user's behavioral intentions and habits, which can help analyze timeliness of the information. This paper proposed a query-process-division method for users' browsing logs. Based on this method, a large number of real users' browsing historical records were statistically analyzed. The results show that in average, a user browses 8.05 Web pages in 6.28 minutes for one query. In addition, nearly 1/3 of total queries carry out concurrently and alternately. It is also found that users rely on inner-site searching more. By analyzing the browsing historical records of a community site, we found that the users are not satisfied with the query results mainly because of the non-high-related results. Out-of-date information is only a small part in the query results.

Keywords Web browsing logs, User behavior analysis, Web page timeliness, CQA

1 引言

进入Web2.0时代,互联网信息出现井喷式增长,特别是以问答社区(Community Question Answering,CQA)为代表的信息共享系统得到了广泛的应用。典型的社区如Yahoo! Answers、百度知道和搜搜问问等都已经拥有上亿级别的信息量,越来越多的Web用户通过各种途径从中获得信息。然而,随着信息量的不断积累和时间的推移,其中一些"陈旧"信息的价值开始逐渐降低,搜索引擎反馈给用户的查询结果中也经常会出现这些过时的信息,给用户的使用造成困扰。如何有效地检测和管理这些过时信息成为各种信息共享平台亟待解决的问题。

针对上述问题,我们从用户角度出发,通过对用户的网页浏览日志做分析,再结合网页信息本身的一些特征,力争对信息时效性做出更全面的判断。网络用户日志分析作为一种有效的手段被广泛应用于各类信息分析的研究中,研究人员根据用户的日志对用户行为、意图和使用习惯等做相应的判断,为各种网络应用提供指导,以更好地满足用户的使用需求。网络用户日志主要分为服务器端收集的用户日志和客户端收集的用户日志两大类,如搜索引擎的查询日志和网页浏览日志分别是两类用户日志的典型代表。与搜索引擎查询日志相比,网页浏览日志记录了用户浏览过程的每一个细节,包括在什么时间浏览了哪些网页。这些细节能更全面地反映用户的使用情况,不受单一搜索引擎的限制,因为搜索引擎日志收集的只是用户与单个搜索引擎交互的记录,而实际上用户在查询信息时很可能在几个搜索引擎之间切换查询。

在网页浏览日志中,浏览记录并不是孤立存在的,每一个信息查询的过程都包括一系列网页浏览记录,不同目的的查询过程交织在一起组成了用户浏览日志。支持本文研究所使用的数据集是24498个网络用户在一个月内的浏览日志,这

到稿日期:2013-05-02 返修日期:2013-08-26
郭俊霞(1977—),女,博士,讲师,主要研究方向为网络信息再利用技术、网页时效性分析,E-mail:gjxia@mail.buct.edu.cn;高 城(1988—),男,硕士生,主要研究方向为网页时效性分析;许南山(1956—),男,副教授,主要研究方向为过程工业实时监控(系统集成)、基于网络数据库的应用技术研究、计算机网络与MIS系统;卢 罡(1981—),男,博士,讲师,主要研究方向为复杂网络与复杂性科学、社会计算。

些日志数据来自于日本用户，是由尼尔森在线(http://www.netratings.co.jp/)收集的，日本 NII 研究所某实验室作为合作研究方购买该数据并提供使用，尼尔森在线拥有该数据的版权。为了能够有效地利用这些数据，本文提出了一种针对网页浏览日志数据划分用户查询过程的方法，即将一个用户连续的浏览记录按照不同的信息需求划分成若干次查询过程(详见4.2节)。这为更好地理解用户行为习惯提供了前提。在此基础之上，对浏览日志数据集做划分处理，并对划分结果做初步的统计分析，包括多个查询过程交替并发的频率、查询过程时间跨度分布以及搜索记录所占比例等。同时，本文选取数据集中访问量相对较多的 CQA 网站 Chiebukuro(雅虎日文问答社区 http://chiebukuro.yahoo.co.jp/)作为研究对象，结合相应的用户浏览日志对这些信息的时效性做了初步的分析。

本文第 2 节介绍目前国内外的相关研究工作；第 3 节和第 4 节分别介绍网页浏览日志数据的结构和处理方法；第 5 节给出对浏览日志划分的统计结果和讨论；第 6 节结合浏览日志分析给出针对 Chiebukuro 社区信息的时效分析；最后是总结。

2 相关研究

网络用户日志蕴含巨大的信息量，是各种网络应用研究的重要数据源。搜索引擎查询日志在搜索引擎的不断优化过程中扮演着重要角色，相关的研究一直被人们作为重点。其中对搜索引擎查询日志 Session 划分的研究，又被视为是深入研究用户行为和信息质量的前提和基础。目前，对查询日志 Session 划分的研究主要集中在时间间隔和查询内容两个方面。

在对搜索引擎查询日志 Session 划分的早期研究中，研究人员比较关注查询时间间隔。在 Silverstein 等人的研究[1,2]中定义 Session 为一定时间内连续提交的查询，并把时间分别设定为 1～50 分钟做相应的分析。后来，研究人员意识到 Session 的划分不应拘泥于固定的时间段，设定停顿时间间隔更合理，一般设定为 30 分钟左右[3,4]。同时，在对查询日志的研究过程中，研究人员还发现了用户浏览过程中的一些重要现象。Radlinski 和 Joachims 的研究[3]发现用户经常会连续提交一系列相近的查询，并称之为“查询链”；Spink 等人在后来的研究[5]中又发现用户查询信息时存在多任务并发进行的现象，即用户在查询某一信息的同时可能也在查询另外一些信息。

随着研究的深入，查询词和查询结果在 Session 的划分研究中也越来越受到重视。一些研究把一个用户与一个搜索引擎的所有交互看成一个整体，并利用用户查询词的不同在其中判断主题漂移来划分 Session[6-8]。在 Shen 等人的研究中[9]又利用查询结果对查询词进行语义扩展，以尽量保证对主题漂移判断的准确性。后来，研究人员结合使用停顿时间间隔和查询词进行 Session 的划分。其中 Joins 和 Klinkner 等人[10]使用二类分类法判断两次查询是否是同一个 Session；张磊等人[11]分别使用统计语言模型和决策树方法做 Session 边界的识别，得出的结论是前者不太适合数据稀疏问题，决策树方法则相对比较理想；文献[12]使用聚类方法实现了 Session 划分，并与采用时间间隔划分的结果进行了对比，结果显示使用聚类算法的划分效果更好。

对于用户浏览行为的研究，在 Session 划分的基础上，文献[1]发现 AltaVista 用户一次查询平均搜索 2.02 次，约 37%包含多次搜索；文献[12]统计分析了 AOL 用户的行为习惯，结果显示 Session 时长分布呈现典型的幂律分布，大部分用户提交的搜索都在 10 次以内，另外有 75%的查询存在多次交替并发的现象。在文献[5]的研究中发现，这种多任务并发查询的现象在搜索引擎用户中比较普遍，并且呈现增长态势。对问答社区中用户行为的研究，大都集中在提问者和回答者之间，主要以用户活跃性分布规律[13]、用户动机[14]和用户需求特点[15]等为主，对于浏览用户的行为习惯研究相对较少。然而，CQA 网站的问题与回答对于拥有相似问题的搜索用户同样有着很重要的意义。Liu Qiaoling 等人[16]将用户的搜索过程纳入到研究的范畴中，分析搜索用户对 CQA 网站信息的满意程度。在搜索引擎的帮助下，CQA 的使用人群不再局限于社区用户，如何更好地理解这些搜索用户的行为和使用习惯变得越发重要。

同样作为网络用户日志，网页浏览日志记录用户浏览查询信息的细节，对于分析用户行为习惯有着独特的优势。在本文的研究中，需要利用网页浏览日志来分析用户行为，以辅助判断 CQA 社区信息的时效性。本文借鉴 Session 划分使用的停顿时间和查询内容对浏览日志做划分处理。

3 网页浏览日志数据集

网页浏览日志通常由客户端的浏览器插件或代理的互联网服务提供商收集，主要记录用户浏览的网页 URL 和浏览时间等信息。相比于搜索引擎查询日志，网页浏览日志提供了一个更全面了解用户行为的途径[17]。

本研究所使用的网页浏览日志数据集是 24499 个网络用户在 2010 年 6 月一个月内的浏览日志，包括 81168263 条记录。这些日志记录包含以下信息：用户 ID、所访问的网页 URL、访问时间、网页停留时间、所访问的网页来源 Referrer。由于用户之间的日志相互独立，可以按照不同的用户 ID 将这些数据分组。按访问时间排序后，一个用户的浏览日志数据结构如图 1 所示。从图中可以看出记录与记录之间的联系是通过网页来源 Referrer 与 URL 的对应关系建立起来的，这种天然存在的前驱后继关系使得每条记录都能知道它的前驱在哪。如果把这种连续的浏览记录看成是一种特殊的树结构，每个用户的浏览日志记录集都是由多棵树组成的，那么就构成了一个森林结构，其中每一条记录都是其中的一个节点。

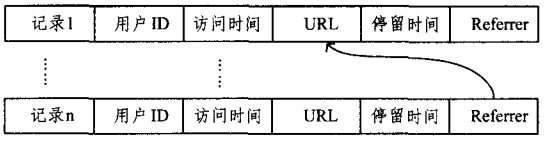


图 1 一个用户的浏览日志数据结构图

同时，节点与树、树与树之间的关系绝不仅仅依赖于这种结构。本研究从中筛选出存在搜索记录的部分作为研究分析的重点，将查询的相关性和数据本身的结构结合起来分析，对数据做相应的处理。

4 网页浏览日志处理过程

要从当前的网页浏览日志数据中获取有用信息，就要对这些看似杂乱的数据做有效的处理，才能使它们为后续的研究服务。首先需要确定用户的查询关键字是什么，围绕着不

同目的都浏览了哪些网页。其中涉及的主要分析处理过程有：

1. 截取 URL 中的搜索关键字。

2. 从杂乱的浏览日志中划分不同目的的查询过程(即：用户围绕某一关键字或者相似关键字进行查询的所有浏览记录)。

4.1 关键字的截取

在网页浏览日志中并不直接提供用户在浏览过程中的搜索关键字，而一般情况下搜索引擎会将查询关键字以 GET 的形式存放在 URL 中，需要通过某种方法从这些各式各样的 URL 中截取用户的搜索关键字。

根据 URL 规范[18]，一般的 URL 的结构如下：

scheme://host[:port]/path? query#fragment

客户端与服务器之间如果以 GET 的形式传递信息，信息会以变量的形式存放在 query 部分，也就是 URL 中“?”后面和“#”前面的部分。因此，只要弄清楚一般的搜索引擎都是使用哪些搜索关键字变量传值就可以处理绝大多数 URL 关键字的截取，而对于访问量比较大且传值方式和传值变量特殊的情况可以单独处理。

文中总结了常用 Web 通用搜索引擎和访问量较大的网站站内搜索系统使用的关键字变量，如表 1 所列。通过判断 URL 的 query 部分中是否含有这些关键字变量中的一个来定位搜索关键字的位置。但是，这样做有可能会出现同一个 URL 中存在两个以上可匹配的关键字变量的情况。为此，我们从数据集中抽取了含有两个以上可匹配关键字变量的 URL 样本，发现产生歧义的关键字变量绝大部分是 $q\backslash k\backslash w$ 中的两个或 3 个同时出现。而在 q 和 k 或 w 同时出现时，q 为关键字变量的频率更高，所以从顺序上 q 具有更高的优先级，同理 k 的优先级要高于 w。另外，对于一些访问量较高而关键字变量比较特殊的网站，单独进行处理，如 yahoo 相关的搜索 URL 一般以 p 为关键字等。

表 1 常见搜索引擎关键字变量

关键字变量名
MT、sobj_keyword、search_word、search_query、field_keywords、query、keyword、keywords、word、key、keyw、qry、kw、wd、text、request、searchword、search、qt、q、k、w

同时，我们发现还有少数网站的 URL 传值并不符合常规的 GET 传值方式。这些 URL 没有关键字变量，搜索关键字一般分布在 path 部分中。对此类 URL，本文也总结了其规律，如关键字分布在/word/、/keyword/、/keywords/、/wordlist/、/query/、&q=、/keyword=等标志性字符串后面，并以‘/’结束或直接结束。

为了验证该方法的有效性，我们从数据集中抽取了两个 URL 集，分别包含 1000 条以 GET 方式传参的 URL 和 1000 条以其他方式传参的 URL。经验证，第一个 URL 集的关键字截取正确率在 95%以上，另一个 URL 集的截取准确率也能达到 90%以上。

4.2 查询过程的划分

查询过程指的是一个用户为了满足某种信息需求而在一定时间内进行的网页浏览活动，其在浏览日志中对应一系列浏览记录，其中可能包括一次或者多次与搜索引擎交互的过程。而且，这些浏览记录不一定是连续的，可能出现几个查询过程交替并发进行[5,19]。

对于网页浏览日志查询过程的划分，据我们所知，目前并没有系统的研究可供参考。而在针对搜索引擎查询日志做 Session 划分时，除了人工逐条查看划分[19]之外，采用较多的是根据时间间隔、查询内容、查询结果以及点击信息等属性进行边界判定[11,12]。相比于搜索引擎查询日志，浏览日志虽然更加详细地记录了用户的浏览情况，但是可以收集到的信息种类相对单一。鉴于以上事实，本文提出了一种可以自动划分查询过程的方法，该方法结合使用时间间隔和查询相关度，并且可以处理多查询过程交替并发的情况。

为了便于描述，在本文中将不存在与搜索引擎交互的浏览过程视为无查询关键字的特殊查询过程。我们认为，在一个用户的浏览日志森林结构中，一棵树中间节点无关键字明显变化时，该树中所有节点归入一次查询过程；同时，时间上相邻(在一定时间间隔内)的两棵树查询关键字相近也应该认为是同一次查询过程。据此，给出如下定义：

定义 1(查询过程) 定义浏览日志数据集为 BL，其中用户集为 $U=\{u_1,u_2,\cdots,u_n\}$。用户 u_i 的网页浏览记录又可以表示成一个森林结构 $F_i=\{s_{i,1},s_{i,2},\cdots,s_{i,m}\}$，$s$ 表示一次查询过程。则有

$$BL=\sum\sum t_{i,j}=\sum\sum s_{i,k} \tag{1}$$

其中，$t_{i,j}$ 表示第 i 个用户的第 j 棵浏览记录“树”，查询过程 $s_{i,k}=\{t_{i,ok},\cdots,t_{i,vk}\}$，这里的 $\{t_{i,ok},\cdots,t_{i,vk}\}$ 必须满足在不超时的前提下查询关键字相近，并且不属于任何查询过程的子集。

根据以上说明，本文提出的划分方法描述如下：

(1)定义超时界限 TIMESPAN(用户查询间歇停顿时间上限)，按不同用户 ID 将记录分组，并对每个用户的记录按时间先后排序，分别对每个用户做以下 4 步处理。

(2)每个用户排序后的第一条记录节点作为当前树的树根，向下遍历查找后继节点记录。

(3)找到后继节点，判断查询关键字是否有明显变化，如果没有则加入到当前树中，继续向下查找。在第一次遍历到非后继节点(不存在前驱后继关系或者关键字明显变化)时将其标记为下一棵树的树根。

(4)在向下遍历查找的过程中，如果当前记录与当前树最后一个节点的时间间隔超过 TIMESPAN，就认为该树终结。同时判断该树是否与已存在的查询过程在不超时的前提下查询关键字相近，若相近则将当前树归入其中，否则当前树独立成一次查询过程并等待后来生成的树加入进来。

(5)然后，以标记好的下一棵树的树根为起点生成新的当前树，继续向下遍历查找后继。

(6)反复执行(3)－(5)步直至所有的记录都加入到相应的查询过程中为止。

以下是对一个用户浏览记录集 Records 的过程划分算法：

算法 1 浏览日志查询过程划分算法

```
const TIMESPAN=X;//定义超时上限
Root=Records[0];//将第一条记录作为树根
while(exist(Root)) do begin
  CurTree=new Tree(Root);//生成新树
  for(Note note∈UnlabelRecords) do begin
    if(超时)break;
    if(无前驱或关键字明显变化)continue;
```

```
    else curTree.add(note);//节点加入当前树
  end for
  if(存在树关键字与 CurTree 相近)合并两树;
  else CurTree 独立成树;
  标记 CurTree 的所有节点已加入树中;
  Root=UnlabelRecords[0];//新树根
end while
```

算法执行完成后，所有的查询过程都以 Tree 的形式存在。算法中，树与节点的存储结构如下：

```
class Tree{
    Note Root;//树根节点
    String Key;//树的关键字,由各节点关键字构成
    Note[] Leafs;//树种所有节点
    Datetime LastLeafTime;//树最后一个节点的时间
}
class Note{
    Datetime Time;//节点访问时间
    Uri Url;//节点所访问的站点 Url
    Uri Referrer;//节点来源的 Url,可能为空
}
```

在评价两次查询是否相近时，本着两次查询若包含相同部分则更可能相近的原则，本研究将这些查询关键字分词后去掉停用词，然后比较两次查询是否包含相同的词，来判断它们是否是相近查询。这里的相近仅反映字面上的相近。

为了确定 TIMESPAN 的值，以 5 分钟为超时界限(TIMESPAN=5 分钟)，从数据集中随机选取了 27259 条连续的浏览记录，并将这些记录按不同的查询过程划分开来。其中包含搜索关键字的查询过程有 980 个，占总数的 1/5。通过手工逐条分析，对其中前后关键字相近却被划分成两个查询过程的时间间隔分布情况做了统计，结果如图 2 所示。

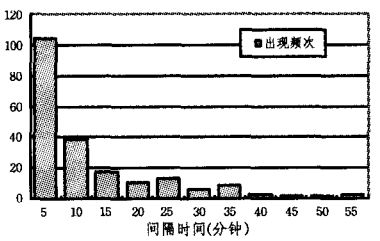


图 2　样本集中关键字相近却被分开的情况分布(以 5 分钟为超时上限)

由于频数较低，图中并没有反映间隔超过一小时的情况。从图中不难发现，这是一条典型的幂函数曲线，我们将其拟合成幂函数 $p(x)$，拟合优度为 0.934，其中 $b_0=1615.65$，$b_1=-1.63$，图 3 为拟合结果。

$$p(x)=b_0x^{b_1} \tag{2}$$

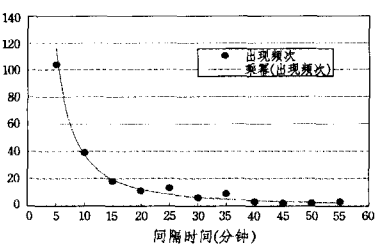


图 3　拟合结果

然后，构造拟合函数的概率密度函数 $f(x)$[12]，其中以 1(分钟)作为 $x_{\min}$。

$$f(x)=\frac{b_1-1}{x_{\min}}\left(\frac{x}{x_{\min}}\right)^{-b_1} \tag{3}$$

通过观察发现，停顿 30 分钟以后继续查询的可能性相对较低，即在停顿 30 分钟以后很少有用户重新发起与之前相近的查询。经计算 30 分钟后该事件发生的概率为11.57%，这与文献[12]以小于 14.9%的发生概率接受最大停顿 26 分钟相差不大。所以，在之后的实验中选择 30 分钟作为查询停顿时间上限 TIMESPAN。

4.3　查询过程划分的评价

对于划分结果的评价，主要考量划分的准确程度，以查询过程为单位判断其是否被正确划分[11]，分别计算划分结果的准确率 P、召回率 R 以及它们的 F 值(F-measure)。即：

$$P=\frac{\text{正确识别的查询过程数}}{\text{识别出的查询过程数}} \tag{4}$$

$$R=\frac{\text{正确识别的查询过程数}}{\text{样本集中的查询过程数}} \tag{5}$$

$$F=\frac{2PR}{P+R} \tag{6}$$

以 30 分钟为停顿时间间隔上限对样本集做查询过程划分，并从中选取包含搜索关键字(与搜索引擎存在交互)的查询过程集。经人工逐条查看，除去数据本身出现的页面遗失，评价结果如表 2 所列，准确率、召回率以及 F 值都能达到 80%左右。其中，错误划分的原因主要在于两方面：基于字面的关键字相似判定方法并不能对查询词的语义相似做出判定，造成部分划分结果不准确；对于查询过程的边缘记录，特别是不包含查询关键字的记录，属于哪个查询过程有时并不是很明确。从验证的结果来看，该方法可以用于接下来的数据分析研究。

表 2　查询过程划分结果的评价

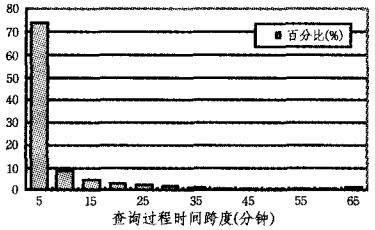
P	R	F
77.6%	86.3%	81.7%

5　基于用户网页浏览日志的统计分析

使用上节中描述的查询过程划分算法对第 3 节中提到的浏览日志数据集做划分处理，得到 23086093 个查询过程，其中含查询关键字的查询过程有 3064639 个，占总数的 12.62%。在含搜索的查询过程中，平均一次查询过程访问 8.05个页面才能找到答案或者放弃查询。图 4 是查询过程时间跨度的百分比分布情况，从图中可以看出绝大多数用户查询过程都会在 5～10 分钟内完成，超过这一时间用户就很少继续搜索下去了。用户查询过程中的浏览记录数分布如图 5 所示，大部分查询过程都只包含个位数的浏览页面数。

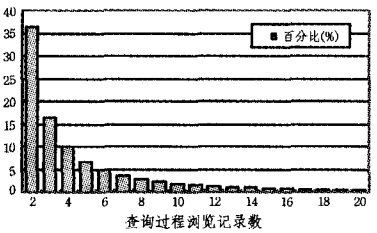
同时，为了能清楚地反映用户每次查询所经过的线路，本研究将所有 URL 分为 3 类：Web 通用搜索引擎类、站内搜索类和普通类，在本文中分别用 1 类网址、2 类网址和 3 类网址表示。根据维基百科对 Web 通用搜索引擎的总结，本研究认定 50 个 Web 通用搜索引擎类网站；对于可以截取关键字但不在一类网址中的 URL 则认为是站内搜索类；其余为普通类，也就是包含用户所需信息的网页对应的网址。在数据集中，3 类网址的浏览记录分别占到了 4.51%、9.87%、85.62%。用户平均每访问 6 个普通页面就要发起一次搜索，

同时用户使用站内搜索的比例要比使用通用搜索引擎的比例还高，这可能与站内搜索更有针对性有关。使用正则表达式对包含搜索的查询过程的不同查询路径做分析，结果如表3所列。从表中的统计结果可以看出，有超过半数的搜索查询过程以搜索页面终结，可以认为出现这种情况可能是很多查询没有强烈的目的性或是用户放弃搜索。在其他几类查询路径中，有接近6成的查询过程与站内搜索有关，可见这些站内搜索在用户查询信息时发挥着重要作用。



其中X轴表示一次查询过程的时间跨度，
Y轴表示频数所占百分比

图4　查询过程时间跨度的分布图



其中X表示一次查询过程所包含的浏览记录数，
Y轴表示所占百分比

图5　查询过程浏览记录数百分比分布图

表3　不同查询途径的分布情况

查询路径	数量	百分比
1类网址→3类网址	557424	18.19%
1类网址→2类网址→3类网址	58699	1.92%
2类网站→3类网址	649509	21.19%
2类网址→1类网址→3网址	29873	0.97%
以1类或2类网址结束查询	1769134	57.73%

另外，本研究还对用户在查询过程中的关键字变化、多次查询过程交替并发以及查询间断情况做了统计。结果显示，有17%的查询会在中途更换查询关键字；有30%的搜索查询过程是交替并发进行的，这与文献[12]的75%有比较大的差异，其原因可能与区域用户的使用习惯有关系；有18%的查询过程发生了查询间断，即并不是连续进行的。

6　CQA信息初步时效分析

信息时效性，即：信息价值与时间的相关程度。时效性低的信息，其价值会在很长时间内比较稳定；而时效性高的信息，一经发布其价值会在一段时间后逐渐降低，成为“过时”信息。主流的CQA网站经过长时间的发展，大都积累了海量信息，其中也不乏这些“过时”的信息。

在信息时效分析实验中，选取数据集中访问量相对较多的问答社区Chiebukuro（一个日文CQA网站）作为实验对象。在对网页浏览日志做查询过程划分的基础之上，结合用户所浏览网页的特征，对相应网页信息做初步的时效性分析。

通过分析一些能够反映Web用户使用情况的因素，再结合人工验证的结果，来推断Web用户对CQA所提供信息的满意程度，进而推断不满意浏览记录的原因与信息过时的关系。为此，本文将用于分析的满意度影响因素分为3方面：查询相关性、用户浏览行为特征、CQA非文本特征。其中用户浏览行为特征包括用户网页停留时间、当前查询过程是否终结、社区答案中的链接点击与否、推荐链接点击与否、其他搜索结果点击与否、是否对相近关键字做了新搜索等。CQA非文本特征能够较为准确地反映对应信息的质量[20]，本文总结并使用了在CQA研究中经常使用的项目：问题接受的回答数量、提问者级别、最佳答案提供者级别、最佳答案的长度、问题长度、网页平均点击数、从问题提出到解决的时间间隔等特征。

实验以人工标记的用户满意程度为准，分析造成用户不满意的原因。随机选取400条Chiebukuro网页的浏览记录，以及这些记录所属的查询过程，作为实验样本集。3位研究人员（作者所在研究室的日本联合研究实验室成员）分别标记这400条记录的用户满意程度，最终的结果以3位研究人员给出的评价标记值的平均值为准，并采用舍入原则得到最后的结果。标记结果：满意网页的比例占总数的13%，一般满意的比例为32.5%，不太满意的占30.5%，不满意的占24.0%。从影响因素的3个不同方面对其中被标记为不满意的Chiebukuro浏览记录做进一步的分析，分析结果如表4所列。

表4　分析结果

可能的原因	不太满意	不满意	合计	百分比
查询相关性不高	15	10	25	11.47%
CQA信息质量不高	44	33	77	35.32%
CQA信息过时	22	25	47	21.56%
其他	41	28	69	31.65%

发现在样本集中这部分用户不满意多是查询相关度不高造成的影响，而“过时”信息造成用户不满意的情况只占其中小部分。出现这种情况的原因可能是数据样本较小，同时抽取样本集也并没有特别的针对性。因此在接下来的工作中，我们准备采用机器学习的方法来辅助判断用户满意度，抽取一个更大的样本集做相关的分析。

结束语　本文提出了针对用户网页浏览日志数据的查询过程划分方法，并在划分结果的基础上对2万多个用户的网页浏览日志做了一系列的统计分析。结果显示：用户在查询信息时对各网站的站内搜索依赖较高，30%的查询过程是多查询并发进行的，有90%的查询都在20分钟内完成，17%的查询会在中途更换查询关键字。对CQA网站Chiebukuro的信息时效性做了初步分析，但结果并没有发现最初预期的关于查询浏览页面与信息时效方面的典型特征。在未来的工作中需要更有针对性地对CQA信息时效性差异问题做深入的分析讨论，并通过收集一部分中文CQA使用人群的网页浏览日志做中文CQA信息时效性差异分析。

参考文献

[1] Silverstein C, Marais H, Henzinger M, et al. Analysis of a very large web search engine query log[C]// ACM SIGIR Forum. 1999, 33: 6-12

[2] He D, Göker A. Detecting session boundaries from web user logs

[C]//Proceedings of the BCS-IRSG 22nd annual colloquium on information retrieval research. 2000:57-66

[3] Radlinski F, Joachims T. Query chains: learning to rank from implicit feedback[C]//Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining. 2005:239-248

[4] Jansen B J, Spink A. How are we searching the World Wide Web? A comparison of nine search engine transaction logs[J]. Information Processing & Management, 2006, 42(1):248-263

[5] Spink A, Park M, Jansen B J, et al. Multitasking during Web search sessions[J]. Information Processing & Management, 2006, 42(1):264-275

[6] Lau T, Horvitz E. Patterns of search: analyzing and modeling Web query refinement[C]//Proceeding UM'99—Proceeding of the Seventh International Conference on User Modeling. Springer-Verlag New York, 1999:119-128

[7] He D, Göker A, Harper D J. Combining evidence for automatic web session identification[J]. Information Processing & Management, 2002, 38(5):727-742

[8] Ozmutlu H C, Cavdur F. Application of automatic topic identification on excite web search engine data logs[J]. Information Processing & Management, 2005, 41(5):1243-1262

[9] Shen X, Tan B, Zhai C. Implicit user modeling for personalized search[C]//Proceedings of the 14th ACM International Conference on Information and Knowledge Management. 2005: 824-831

[10] Jones R, Klinkner K L. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs[C]// CIKM'08. 2008:699-708

[11] 张磊,李亚楠,王斌,等. 网页搜索引擎查询日志的 Session 划分研究[J]. 中文信息学报,2009,23(2):54-61

[12] Lucchese C, Orlando S, Perego R, et al. Identifying task-based sessions in search engine query logs[C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. 2011:277-286

[13] Lui Y, Agichtein E. On the Evolution of the Yahoo! Answers QA Community[C]//the ACM SIGIR International Conference on Research and Development in Information Retrieval. Singapore, 2008:737-738

[14] Nam K K, Ackerman M S. Question in, Knowledge in?: a study of naver's question answering community[C]// Proceedings of CHI'09. Boston, MA, 2009:779-788

[15] Rodrigues E M, Frayling N M. Socializing or knowledge sharing?: characterizing social intent in community question answering[C]// Proceedings of CIKM 2009. Hong Kong, China, 2009:1127-1136

[16] Liu Q L, Agichtein E, Dror G, et al. Predicting web searcher satisfaction with existing community-based answers[C]// Proceedings of SIGIR'11. Beijing, China, 2011

[17] Jiang D, Pei J, Li H. Mining Search and Browse Logs for Web Search: A Survey[J]. ACM Transactions on Computational Logic, 2013, 4(4):1-42

[18] Wikipedia. Uniform resource locator[EB/OL]. http://en.wikipedia.org/wiki/Uniform_resource_locator

[19] Hassan A, Jones R, Klinkner K L. Beyond DCG: User behavior as a predictor of a successful search[C]// Proceedings of the third ACM international conference on Web search and data mining. 2010:221-230

[20] Liu Y, Bian J, Agichtein E. Predicting information seeker satisfaction in community question answering[C]// Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2008:483-490

(上接第 90 页)

参 考 文 献

[1] 姚天顺,朱靖波,张琍,等. 自然语言理解———种让机器懂得人类语言的研究(第 2 版)[M]. 北京:清华大学出版社,2002

[2] 俞鸿魁,张华平,刘群,等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报,2006,27(2):87-94

[3] 曹勇刚,曹羽中,金茂忠,等. 面向信息检索的自适应中文分词系统[J]. 软件学报,2006,17(3):356-363

[4] 刘挺,马金山. 汉语自动句法分析的理论与方法[J]. 当代语言学,2009,11(2):100-112

[5] Mark A. Paskin, Cubic-time Parsing and Learning Algorithms for Grammatical Bigram Models [R]. Technique report, 2001

[6] 熊德意,刘群,林守勋. 融合丰富语言知识的汉语统计句法分析[J]. 中文信息学报,2005,19(3):61-66

[7] 李正华,车万翔,刘挺. 基于柱状搜索的高阶依存句法分析[J]. 中文信息学报,2010,24(1):37-41

[8] Harper M P, Huang Zhong-qiang. Chinese Statistical Parsing [M]// Joseph Olive, John McCary, Caitlin Christianson, eds. Handbook of Natural Language Processing and Machine Translation. Reston Virginia, Defense Research Projects Agency, 2011:90-102

[9] 代印唐,吴承荣,马胜祥,等. 层级分类概率句法分析[J]. 软件学报,2011,22(2):245-257

[10] 陈功,罗森林,陈开江,等. 结合结构下文及词汇信息的汉语句法分析方法[J]. 中文信息学报,2012,26(1):9-15

[11] Bizera C, Lehmannb J, Kobilarova G, et al. DBpedia-A Crystallization Point for the Web of Data [C]// Proceedings of Web Semantics: Science, Services and Agents on the World Wide Web. 2009:154-165

[12] 罗军舟,金嘉晖,宋爱波,等. 云计算:体系架构与关键技术[J]. 通信学报,2011,32(7):3-21

[13] 于戈,谷峪,鲍玉斌,等. 云计算环境下的大规模图数据处理技术[J]. 计算机学报,2011,34(10):1753-1767

[14] Bahga A, Madisetti V K. Analyzing Massive Machine Maintenance Data in a Computing Cloud [J]. IEEE Transactions on Parallel and Distributed Systems, 2012, 23(10):1831-1843

[15] 李锐,王斌. 文本处理中的 MapReduce 技术 [J]. 中文信息学报,2012,26(4):9-20

[16] 宁可为,王炜,李园伟. 基于 Hadoop 的句群相似度计算 [J]. 计算机系统应用,2010,19(12):59-63

[17] http://www-nlp.stanford.edu/software/lex-parser.shtml[EB/OL]