ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.5, May 2010, pp.1055−1066
doi: 10.3724/SP.J.1001.2010.03744

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

网络检索用户行为可靠性分析*

岑荣伟+, 刘奕群, 张 敏, 茹立云, 马少平

(清华大学 智能技术与系统国家重点实验室,北京 100084)

Reliability Analysis for the Behavior of Web Retrieval Users

CEN Rong-Wei+, LIU Yi-Qun, ZHANG Min, RU Li-Yun, MA Shao-Ping

(State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China)

+ Corresponding author: E-mail: crw@mails.tsinghua.edu.cn

Cen RW, Liu YQ, Zhang M, Ru LY, Ma SP. Reliability analysis for the behavior of Web retrieval users. *Journal of Software*, 2010,21(5):1055−1066. http://www.jos.org.cn/1000-9825/3744.htm

Abstract: Based on large scale click-through data, this paper study the interactive process between user and search engine, and derive user decision process. A comparative study between clicks on relevant results and non-relevant ones analyzes the reliability of individual user click-through behavior on search results. Three types of features are proposed and estimated for separating reliable user clicks from other ones. Experimental results show that the proposed method evaluates the reliability of user behaviors effectively based on click context features of Web search users.

Key words: user behavior; click reliability; Web search engine system

摘 要: 基于大规模真实网络用户的行为日志,对用户与网络搜索引擎系统的交互过程和用户决策过程展开研究.通过比较具有相关信息的用户点击和普通点击的分布,对用户点击的 3 类上下文背景特征进行分析,从而实现对用户点击的可靠性评估.实验结果表明,通过对用户点击的上下文背景的特征分析,能够发现用户检索行为中的思维决策过程,并进而对用户点击的可靠性进行有效的评估.

关键词: 用户行为;点击可靠性;网络搜索引擎系统

中图法分类号: TP311 **文献标识码**: A

现代社会网络信息极度丰富,网络信息检索工具已成为人们访问互联网资源的主要媒介.作为当前网络信息检索的主要工具,搜索引擎已成为人们访问互联网资源的有效途径.因此,如何改进搜索引擎系统的检索算法,提高检索性能,已成为研究界和产业界关注的主要话题.其中,用户反馈是算法优化、系统维护和性能评估的重要手段,也是网络搜索和知识挖掘的重要研究领域之一,已越来越受到研究人员和系统开发者的关注.作为用户反馈的传统模式,手工评价需要耗费大量的人力和时间资源,难以大规模地实时开展.因此,如何有效挖掘和利用网络用户检索反馈的群体智慧信息已受到研究界的广泛关注(如文献[1,2]).然而,之前的相关研究表明,网

* Supported by the National Natural Science Foundation of China under Grant Nos.60736044, 60903107 (国家自然科学基金); the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant No.20090002120005 (高等学校博士学科点专项科研基金)

Received 2009-04-07; Revised 2009-06-09; Accepted 2009-10-10

络用户并不愿意主动向搜索引擎提供显示的反馈信息[3].最近,由于搜索引擎的广泛流行和用户检索量的大规模增加,基于用户点击信息的研究已成为信息检索和知识挖掘领域的主要研究方向[4].

然而,真实网络检索环境下的用户点击行为信息往往含有大量噪音,其中掺杂了包括网络爬虫等非正常的网络用户[4].2005 年,Joachims[5]展开了一项称为眼睛跟踪(eye-tracking)的研究,结果表明个体用户的点击信息由于搜索引擎结果排序、内容展示等多方面原因而具有偏向性,搜索查询和点击文档之间没有明显的绝对相关性.上述相关研究表明,有必要对网络用户的行为日志进行分析,进而提炼网络用户点击的有效信息,过滤噪音.当前用户行为信息的研究方法主要基于大规模用户点击行为的宏观统计分析,此类分析方法适用于处理用户访问频度高的热门词查询,不适合处理用户访问量较小但数量众多的长尾词查询,也不适合应用于用户的个性化搜索,针对不同兴趣的用户有区别地返回搜索结果.为解决此问题,本文在大规模真实网络用户日志的基础上,对个体用户点击过程中的思维决策过程和点击行为的上下文环境展开研究分析,对用户点击的可靠性进行评估,进而挖掘用户日志中的有效点击.

下面就本研究的相关工作展开讨论,阐明用户行为信息方面已有的研究成果及存在的问题.通过对用户和搜索引擎之间交互过程的分析,本文定义了用户行为可靠性概念并给出特征评价的方式.然后,本文进一步研究用户点击行为的上下文环境特征,分析用户的点击过程和决策过程.最后,根据所得到的特征分析结果,对用户的点击进行评估和筛选.

1 相关研究工作概述

在有效挖掘网络用户群体智慧信息的相关研究中,诸多研究者提出多种用户行为的模型和方法,并将用户反馈信息应用于信息检索和知识挖掘的各子领域中,如搜索结果重排序、检索函数学习、搜索引擎自动性能评价等[6−8].

Tan 等人将 Web 日志中的机器行为模型化,从而识别和过滤非正常网络用户的行为信息[9,10].Yates, Kammenhuber 等人基于马尔可夫过程假设来模拟网络搜索用户的查询修改和页面点击过程,进而对用户的行为进行相关解释[11,12].Sadagopan 在文献[13]中定义了一般用户行为和非一般用户行为的概念,以用户搜索 Session 为单位,利用马尔可夫过程模拟用户行为的各种状态变化,从而判断用户搜索 Session 的一般性.现有这方面研究的目的在于解释正常搜索用户的行为,识别机器行为或非正常用户的行为,而对同一用户的不同行为缺少分析和对比.

2005 年,Joachims[5]对搜索用户点击的有效性开展了一项基础性研究工作,称为眼睛跟踪.它通过分析用户点击行为的决策过程,挖掘用户点击过程中蕴含的智慧信息.基于 Joachims 的工作,Agichitein[4]也提出返回结果的排序位置对用户点击造成的偏置问题,并提出了统计背景模型和多种用户行为模型解释群体用户行为,挖掘同一查询内各点击文档之间的相关关系的强弱.最近,Craswell[14]和 Guo[15]提出瀑布模型模拟用户的点击行为,解释并预测了用户的点击过程.已有的这类研究工作或是基于可控制的实验环境条件,与真实的 Web 用户交互行为存在较大的差距,从中挖掘的相关规则和方法也难以应用于真实的环境;或是从宏观的角度出发,着眼于群体用户的点击行为分析,缺少对个体用户行为的有效性分析,尤其是对同一查询,这类方法需要大量的用户点击信息,难以处理用户访问频度低的长尾查询词.

本文针对已有研究工作的不足,基于个体搜索用户在查询点击行为过程中的上下文环境特征,分析用户的点击行为和思维决策过程,进而给出用户在点击过程中的行为偏好,对用户点击的可靠性展开评估.

2 用户行为及行为可靠性

在分析用户行为日志之前,本文先对用户和搜索引擎之间可能的交互过程以及用户在此交互过程中可能的思维决策进行假设分析,并定义搜索用户行为的点击可靠性概念.

2.1 用户检索过程

网络搜索用户在信息查询时通常与搜索引擎系统之间有一个交互过程.首先,用户有一个查询需求主题或者查询目的.根据该主题或目的,用户基于已有的搜索经验和知识构造相关查询关键词,并将其提交给在线搜索引擎系统,如 Baidu,Google,Sogou 等.搜索引擎系统根据用户的查询关键词采用一定的算法和检索策略返回可能相关的结果文档列表.用户通过对比返回结果文档的相关信息,如标题,摘要,URL,前、后文档等,点击可能相关的能够满足其搜索目的的结果文档.如果该结果文档满足搜索需求,用户则可能离开该查询主题的搜索.反之,如果不满足搜索需求,用户则会返回搜索结果页,继续查找其他可能相关的结果文档并进行点击;或者修改查询关键词,与搜索引擎系统进一步交互.当用户对点击结果文档满意或者认为无法找到相关结果文档时,它会选择离开当前的查询主题,或者换一个搜索引擎系统继续搜索.表 1 是来自真实网络搜索用户的交互过程实例,从中我们可以推测该用户想找汽车的相关信息,并可能存在一定的购车意愿,其主要意向为“丰田卡罗拉”.在搜索过程中,该用户还参考和对比其他品牌的汽车,如“上海大众”和“广州本田”.最终,该用户查看丰田卡罗拉汽车的相关配置信息,可以推测其最终目的是想了解该车的细节参数信息,为购车做准备.

Table 1 A case of an interaction between a user and a search engine

表 1 真实网络搜索用户的交互过程实例

Time	Query	Rank	Page clicked
20:58:58	丰田	6	www.autohome.com.cn/526/
21:02:34	丰田	5	www.autohome.com.cn/110/
21:03:23	丰田	6	www.autohome.com.cn/526/
21:04:11	上海大众	5	www.che168.com/che168/cardb/brand/brand_58.html
21:06:14	广州本田	3	car.autohome.com.cn/brand/32/
21:09:23	丰田	2	car.autohome.com.cn/brand/63/
21:10:20	丰田	4	price.pcauto.com.cn/brand.jsp?bid=31
21:11:20	丰田	10	www.che168.com/che168/cardb/brand/brand_24.html
21:12:43	丰田卡罗拉	1	www.autohome.com.cn/526/
21:19:12	丰田卡罗拉	11	www.autohome.com.cn/526/options.html

通过上面的分析我们知道,在用户和搜索引擎的不断交互过程中,用户的每次点击行为都在一定的上下文背景环境之中.因此,通过对当前点击的上下文背景分析,我们可以对用户在交互行为过程中的思维决策过程有所认识,进而可以对当前用户的满意程度和点击行为的可靠性作出估计和判断,找出点击文档和查询词之间的相关信息.

2.2 用户点击可靠度

2005 年,Joachims 等[5]人指出,用户点击不能用于判断查询和文档之间的绝对相关性,只能用于评估相对相关性.本文使用点击可靠性来衡量查询和文档之间相关性的程度.为判断哪类点击对应的查询和文档是相关的,先给出用户点击可靠性的概念.我们形式化地以 c 表示当前点击,是用户点击对应的“查询-文档”对,$R(c)$表示点击 c 对应的“查询-文档”对的相关性.若 $R(c)=1$ 表示该“查询-文档”对是相关的,则点击 c 称为相关点击;反之,若 $R(c)=0$ 表示该“查询-文档”对是不相关的,则 c 称为不相关点击.

用户点击可靠性:对于给定的用户点击 c,用户点击可靠性$\Re(c)$是点击 c 对应的页面和查询之间存在相关性的概率估计.若点击文档与查询是相关的,则该点击可靠性强;否则,可靠性弱.用概率公式表达如下:

$$\Re(c)=P(R(c)=1\mid c) \tag{1}$$

从上面的定义可以看出,本文定义的用户点击可靠性与已有的相关研究有所区别,如判断是否为正常用户的点击、是否为机器行为等方面,也与以查询主题 Session 为粒度的用户行为判断有所不同[9,10,12].本文从概率的角度,以独立的用户点击作为分析粒度,找出用户日志中的相关点击.

为判断用户点击在特定上下文背景中的可靠性,我们需要评估在给定上下文背景特征状态 F 时,其为相关点击的可能性,用概率表示为 $P(R(c)=1|F)$.利用贝叶斯法则,有如下表达式:

$$P(R(c)=1|F)=\frac{P(F|R(c)=1)}{P(F)}P(R(c)=1) \tag{2}$$

上述公式包含 3 个部分.$P(R(c)=1)$是用户点击为相关点击的整体概率,可以用所有点击中相关点击的比例表示,反映用户点击行为的一个整体特性,是一个未知的常量.我们将上述公式转化为

$$P(R(c)=1|F)\propto\frac{P(F|R(c)=1)}{P(F)} \tag{3}$$

公式(3)中,$R(F)$是用户点击日志中具有上下文背景特征 F 的概率,可通过全体点击数据中含有 F 特征的点击比例进行估计;$P(F|R(c)=1)$是相关点击集合中含有上下文背景特征 F 的概率,可通过采样标注的相关点击集合中含有 F 特征的点击比例进行估计.

由公式(3)可知,若一个点击具有上下文背景特征 F,则其可靠性与特征 F 在相关点击集合中的分布成正比,与特征 F 在全体点击数据集合中的分布成反比.若某上下文背景特征在这两集合中的分布差异较大,则该特征可以用于评估点击可靠性.我们将 $P(F|R(c)=1)/P(F)$ 定义为点击的可靠性度 CRV(click reliability value).其值越大,表明该特征用于评估点击可靠性的效果越好.当 CRV 大于 1 时,即 $P(F|R(c)=1)/P(F)>1$ 时,由式(2)可知,$P(R(c)=1|F)>P(R(c)=1)$,即用该特征筛选得到的点击可靠性优于全集上的点击可靠性.

3 用户点击行为上下文背景分析

传统利用用户点击信息的方法假设点击对应的查询与文档具有一定的相关性.为了减少点击噪音的影响,传统方法采用统计分析的方式来保证点击的可靠性,要求所处理的用户查询拥有大量点击.然而,这种模式不针对单个用户点击进行独立分析,很难判断用户每次查询和点击行为的可靠性,因此无法处理访问频率较稀疏但数目较多的长尾词查询.本节将利用真实网络搜索引擎的用户访问日志,分析用户在查询和点击过程中的上下文背景环境和可能的思维决策过程,评估用户点击的可靠性.

3.1 数据准备

之前的相关研究表明[4],实验环境下的用户行为和真实网络环境下的用户行为是有所差异的.为了研究真实网络用户的搜索行为,我们在一家著名商用搜索引擎的帮助下,收集到一段时间内的真实网络环境下的用户搜索点击日志.这些日志包含 2008 年 9 月 19 日~2008 年 10 月 24 日共 38 天网络用户与该搜索引擎进行交互的日志记录.

表 2 列出了这些日志所包含的信息项.其中,User ID 是系统分配给用户的唯一标识,通过该 ID 可以对同一用户的日志信息进行聚集,实现对同一个用户搜索行为的分析研究.由于该 ID 只针对用户进行区分,而无法区分不同时间段下不同的查询主题或查询目的.因此,为了有效区分同一 ID 下不同的查询主题,我们对日志进行 Session 切分.已有研究表明,使用时间信息对日志进行切分是非常有效的方法[16],因此在我们的日志数据预处理过程中,主要使用时间间隔的方式对用户日志进行 Session 切分,得到合理的独立用户查询主题.经预处理后,我们共有 9 139 万次用户查询,1.94 亿条独立的用户点击记录和 5 805 万个用户 Session.这里,当以 Session 为分析研究对象时,构成以 Session 为粒度的上下文背景.同时,在以同一 Session 内同一查询词为对象时,构成以查询词为粒度的上下文背景.

Table 2 Information items in the click-through logs collected

表 2 收集的点击日志的相关信息项

Item	Record content
Query	The user query submitted
URL	URL of the result clicked by the user
Rank	The rank of the result clicked by the user
Order	The order of the result in the click sequence
User ID	Automatically assigned user's identification code according to his query session
Time	Time of the clicking or querying event

为了比较各个特征在相关点击集合和全体点击集合上的不同分布差异,在同一家搜索引擎公司的帮助下,我们人工标注一部分具有相关性的“查询-文档”对.标注过程使用 Pooling 的方法,将日志中随机抽样的 3 000 个查询关键词提交给多个网络搜索引擎系统,把返回的前 20 位结果放入 Pooling 池中,使用人工标注的方法对其进行相关性标注,最后得到 8.9 万相关的“查询-文档”对.利用该标注结果,我们对用户点击日志进行筛选.如果用户的查询和点击页面在标注的相关“查询-文档”对集合中,则认为该点击是可靠相关的,将这些可靠相关点击集合记为 Rel-Set.我们共筛选出 128.5 万条可靠的用户点击日志.同时,全体用户点击日志记为 Whole-Set.本文将分析和比较这两个集合中各点击上下文背景特征的分布差异,进而实现对点击可靠性的评估.

3.2 信息熵特征

在网络用户和搜索引擎系统的交互过程中,用户的查询和点击过程存在着不确定性.信息熵是 Shannon[17] 于 1948 年提出来的,用于描述事物出现的不确定性.事物出现的不确定性越大,其熵就越大.信息熵通常用如下公式表示:

$$Entropy = -\sum_{p_i} p_i \log(p_i) \tag{4}$$

其中,p_i 是事物 i 出现的概率,$-\log(p_i)$为事物 i 的自信息(self-information).

借用熵的概念,我们对用户和搜索引擎系统交互过程中的行为进行确定性度量.本文主要分析用户在交互过程中可能存在的两类不确定性.其一是查询的不确定性:在用户查询过程中,若查询关键词能够准确地表述用户的查询意图,并被搜索引擎系统所理解,返回用户想要的搜索目标结果,则用户可能只提交 1 次查询,不再提交新的查询;反之,用户可能通过不断修改查询关键词来得到满意的结果.其二是点击结果文档的不确定性:如果用户点击结果文档后,该文档是用户想要的目标页面,则用户可能停止与搜索引擎的交互;否则,该用户可能继续点击认为相关的结果文档,直到找到满意的结果页面.我们将这两类上下文背景特征称为用户点击的信息熵特征.下面就这两个特征进行分析和比较.

查询数(QueryNum):在同一用户查询主题内,即同一 Session 过程中,用户提交给搜索引擎系统的独立查询词个数.

图 1 展示了该特征在 Whole-Set 和 Rel-Set 上的特征分布.从图中可以看出,62%相关点击的用户只提交 1 个查询,该比例大于所有点击日志集合中 24%的比例.同时可以看出,大部分相关点击(90.3%)的用户提交的查询次数小于等于 3 次.从该上下文背景特征可以推测对于用户提交较多查询词的 Session 中的点击,其平均的点击可靠性较低,点击对应的查询与文档之间的平均相关性也较低.使用点击可靠性度进行衡量有 $CRV(QueryNum=1)=2.55$,说明其效果明显.该特征分析的结论也说明在用户和搜索引擎系统交互过程中存在着查询词的不确定性.

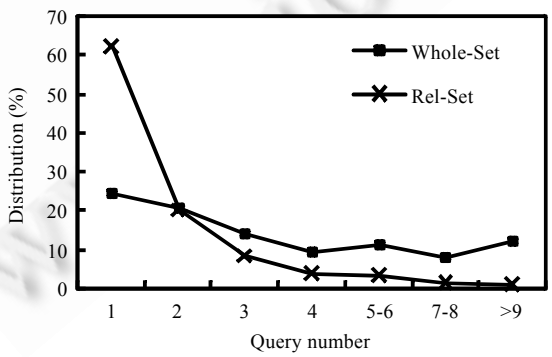


Fig.1 QueryNum distributions of Rel-Set and Whole-Set

图 1 Rel-Set 和 Whole-Set 集合上查询数特征的分布

点击熵(ClickEntropy):在当前用户查询过程中,点击分布的信息熵.点击熵按如下公式计算得到:

$$ClickEntropy = -\sum_{p_i} p_i \log(p_i) \tag{5}$$

其中,p_i 为该用户当前点击页面 i 的点击分布,其计算方法为 $p_i = \frac{\text{当前Session中点击页面 } i \text{ 的次数}}{\text{当前Session的总点击次数}}$.该点击分布 p_i 与传统研究中的点击率 CTR(click-through rate)[4,7]不同,传统研究中的 CTR 是基于所有提交该查询的用户和对应的点击页面统计得到的点击分布,而这里的点击分布 p_i 只针对当前用户 Session 计算得到,因此能够体现当前点击的上下文背景信息.

图 2 展示了该特征在 Whole-Set 和 Rel-Set 上的特征分布.由图 2 可以看出,在 Rel-Set 集合中点击信息熵特征值小的点击比例较大.如当 *ClickEntropy*=0 时,即用户在一个 Session 中,只点击 1 个页面(一次或多次,用户可能对同一页面点击多次),这样的点击有 60%,而在 Whole-Set 集合上只有 32.2%.对于 Rel-Set 集合,一次查询过程中点击页面数小于等于 2(*ClickEntropy* 小于等于 1.0)的比例为 81%,而 Whole-Set 集合该比例只有 52.4%.同样,从该上下文背景特征可以推测出,对于用户的比较离散的点击,其可靠性较低,点击页面和查询词之间的相关性也较差.使用点击可靠性度进行衡量有 $CRV(ClickEntropy = 0) = 1.85$,说明该特征有效.该结论也说明用户与搜索引擎的交互过程中存在着点击的不确定性.

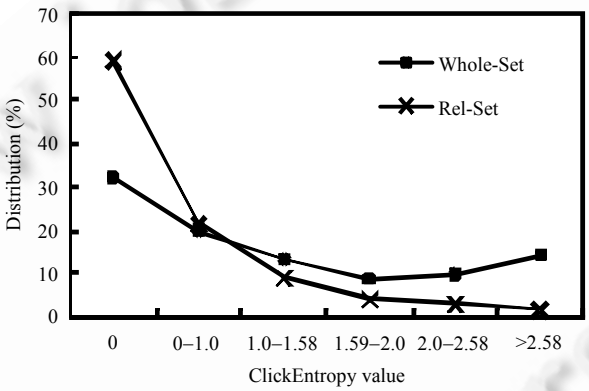


Fig.2 ClickEntropy distributions of Rel-Set and Whole-Set

图 2 Rel-Set 和 Whole-Set 集合上点击熵特征的分布

类似地,我们分析了其他相关特征,如同 Session 过程中用户点击的页面数,如同一查询词下用户点击的页面量等.分析表明,与一般的用户点击相比,可靠性高的用户点击其在用户查询和点击页面上具有更强的确定性.如果用户对结果的满意度较高,则用户提交的查询和点击结果都较为确定.根据式(2)和式(3)的贝叶斯公式可以得出,如果用户提交的查询和点击结果都比较确定,则用户对点击结果可能比较满意,点击可靠性性较高.反之,如果用户提交较多的查询或者点击较多的页面,则这些点击的整体可靠性相对也较低.

3.3 点击顺序特征

用户在一次搜索过程中,各点击会按时间先后顺序形成点击序列.对不同序列位置下的点击,其点击可靠性是所有差异的.Jochims 等人在文献[5]中提到类似的因素,规则“Click>Earlier Click”说明用户后续点击文档的相关性比之前点击的文档相关性要高,虽然在文献[18]中验证这种与搜索结果位置序列相逆的规则其标注性能较差.这一节将根据用户点击所在的序列位置特性进行分析,同时比较在用户 Session 和查询词不同粒度大小下的上下文背景特征的性能.

是否为 Session/Query 中首次点击(FirstClickInSession/FirstClickInQuery):当前用户点击是否为所在 Session/查询词点击序列中的首次点击.这里有两个不同粒度的序列范围,即所在 Session 和所在的查询词,构成两个不同观察粒度的上下文背景.

图 3 展示了这两个特征在 Whole-Set 和 Rel-Set 上的特征分布.从图中可以看出,对于 Rel-Set 集合,Session

粒度下首次点击的比例高达 55.4%(对于 Whole-Set 集合该比例为 25.8%).类似地,在 Query 粒度下,这两个集合中是首次点击的比例分别为 67.4%和 32.57%.经计算,对应的点击可靠性度分别为 $CRV(FirstClickInSession=Yes)$=2.15 和 $CRV(FirstClickInQuery=Yes)=1.42$.可见,若当前的点击为 Session 或 Query 序列中的首次点击,其为相关点击的概率较大.这主要是由于用户在开始一个查询主题或新开始一个查询时,在点击之前,通常会对搜索引擎返回的结果中的标题、摘要、URL 等信息进行认真比较和判断,经过一定考虑后选择点击认为最可能相关的结果页面.对于后续的点击,用户可能不对结果页面进行比较或比较得不充分,查询和点击文档之间的相关性程度也较低.

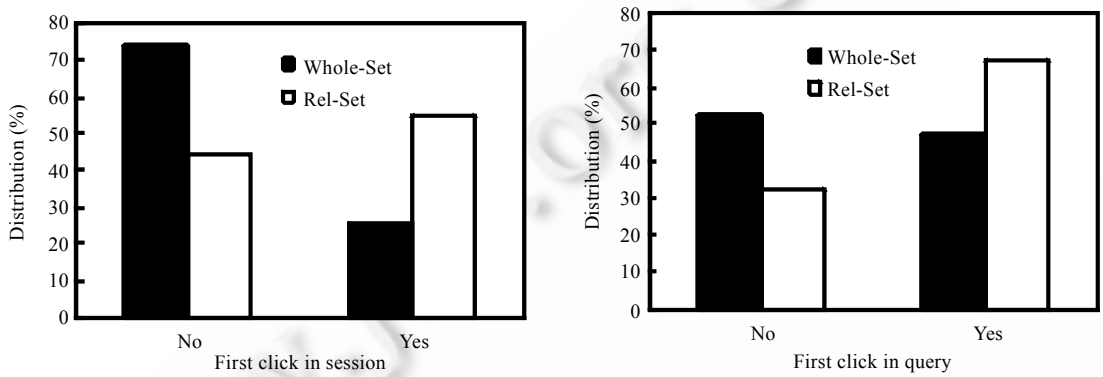


Fig.3 FirstClickInSession/FirstClickInQuery distributions of Rel-Set and Whole-Set

图 3 Rel-Set 和 Whole-Set 集合上是否为 Session/Query 中首次点击的点击分布

是否为 Session/Query 最后一次点击(LastClickInSession/LastClickInQuery):当前用户点击是否为所在 Session/查询词点击序列中的最后一次点击.类似于首次点击的上下文背景特征,该特征也分两个粒度进行分析.

图 4 给出了该特征在不同粒度下的点击分布.可以看出,该特征具有与首次点击类似的分布.对于 Rel-Set 集合,Session 粒度下最后一次点击的比例为 50%(对于 Whole-Set 集合该比例为 26.7%).类似地,Query 粒度下这两个集合中最后一次点击的比例分别为 63.6%和 36.4%.其对应的点击可靠性度 $CRV(LastClickInSession=Yes)=1.87$ 和 $CRV(LastClickInQuery=Yes)=1.35$,可见,这两个特征都有一定的区分度.这种现象可以解释为,如果用户对当前点击的结果满意,找到了需要的页面,则该用户可能不再继续和搜索引擎系统进行交互.这两个特征可以认为是文献[5]中规则“Click>Earlier Click”的特例.

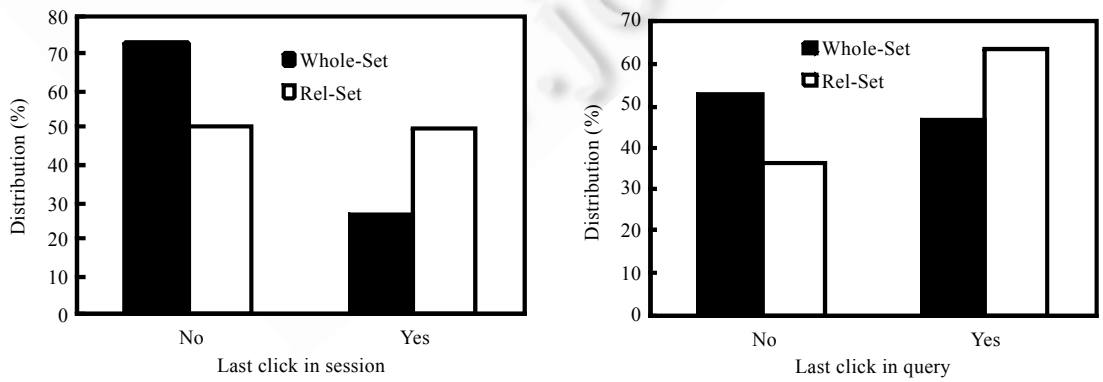


Fig.4 LastClickInSession/LastClickInQuery distributions of Rel-Set and Whole-Set

图 4 Rel-Set 和 Whole-Set 集合上是否为 Session/Query 中最后一次点击的点击分布

上述 4 个特征是点击序列中特殊位置的点击,分别以 Session 和 Query 为考察粒度,得到了类似的结论.同时也可以看出,以 Session 为粒度的特征要优于以 Query 为粒度的特征.通过对日志的统计分析,我们发现其中有 71.01%的 Session,用户只提交一次查询(该比例与余慧佳等人[19]统计的 66.46%结果基本一致,由于我们对 Session 做了进一步的切分,所以比例相对更大).对于用户只提交一次查询的 Session,如果某个点击为 Session 中的首次或最后一次点击,其同时也必为查询中的首次点击或最后一次点击,因此这两组不同粒度的特征具有一定的相关性.经分析,特征 FirstClickInSession 与 FirstClickInQuery 的相关性为 0.654,特征 LastClickInSession 与 LastClickInQuery 的相关性为 0.625.

我们将查询中为首次点击或最后一次点击,且同时为 Session 的首次点击或最后一次点击的点击给予过滤,考察查询粒度下,首次点击和最后一次点击的性能.图 5 给出了这两个特征经过过滤后的点击分布,从中可以看出,查询为粒度的首次或最后一次点击,其上下文背景特征是失效的.用户在查询词更换过程中的首次和最后一次点击的可靠性与一般点击的可靠性基本是相同的.上述分析表明,用户的首次和最后一次点击的特征在 Session 粒度下是有效的,在 Query 粒度下其有效部分也为 Session 中对应的首次和最后一次点击.可推测用户在同 Session 中的行为和思维决策过程是连续的.用户在查询过程中,当其改变查询词时,并不满意前一个查询的查询结果,也没有对后一个查询的首次点击作更多细致的比较.

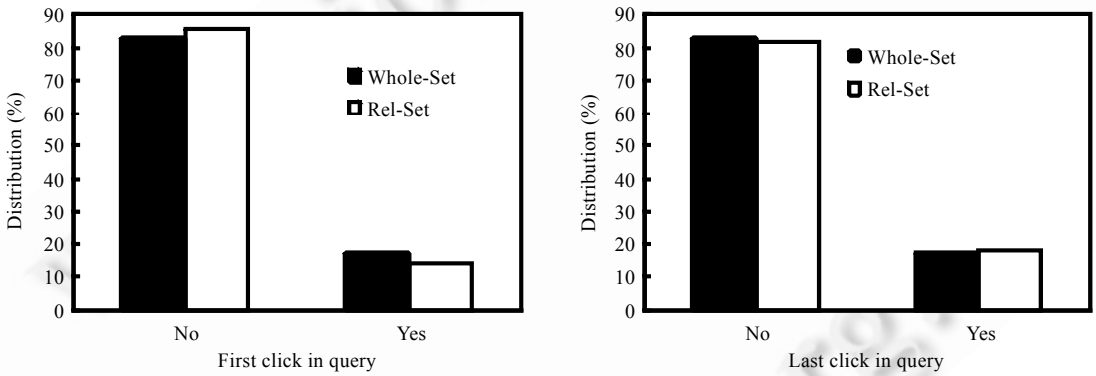


Fig.5 Distributions of first click and last click in query not in session

图 5 查询中非 Session 的首次和最后一次点击的点击分布

在日志中存在大量只有 1 次点击的 Session,对于这部分点击,其首次点击和最后一次点击两个特征将同时生效.通过对日志的统计,我们发现只有 1 次点击的 Session 占 Whole-Set 集合的 16.1%,占 Rel-Set 集合的 31.4%.因此,有必要在这部分 Session 过滤后重新考察首次点击和最后一次点击两个特征的性能和效果.

图 6 显示,过滤一次点击的 Session 后,Whole-Set 集合中首次点击和最后一次点击各占 18.5%,而在 Rel-Set 集合中首次点击和最后一次点击分别占 34.6%和 26.56%(在 Whole-Set 集合中,首次点击和最后一次点击总是成对出现,因此比例分布一致.而在 Rel-Set 集合中,要求点击对应的查询和文档必须相关,因此首次点击和最后一次点击并非成对出现).经计算,首次点击和最后一次点击特征的点击可靠性值分别为 1.87 和 1.44,说明去除一次点击的 Session 后,首次点击和最后一次点击这两个特征仍然有效.

根据上述点击序列特性的分析结果,我们可以看出,用户的点击过程受搜索引擎返回结果和展现的影响,其决策过程随着交互过程的不断变化而变化,并影响着最终的点击可靠性.在日志挖掘过程中,我们可以通过对用户点击序列的上下文背景特征来挖掘用户在点击过程的决策思维,进而发现日志中的有效信息.

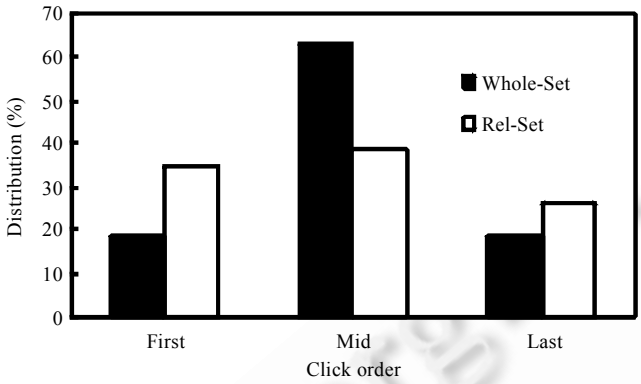


Fig.6 Distributions of first click and last click in sessions with more than one click

图 6 点击大于 1 次的 Session 中首次点击和最后一次点击的点击分布

3.4 其他行为特征

返回结果页面的位置通常被认为对用户点击产生较大的影响,存在着点击偏置(rank bias)[4,5,14,15],增加了用户点击信息挖掘的难度.对于该特征,通常有两种观点:一种是排在搜索引擎结果前列的页面,或是用户更容易看到并点击,或是用户对搜索引擎比较信赖,会造成多于当前文档理应得到的点击量;另一种观点是搜索引擎本身有着较好的智能,因此,排序结果前列的相关性本身就比较高,其得到的点击理应较多.在这里,我们分析普通用户点击和相关点击的位置特征的分布差异,考察搜索结果位置对用户点击造成的影响.

图 7 给出了 Rel-Set 和 Whole-Set 集合上用户点击随点击位置变化的分布情况.从图 7 中可以看出,对于 Whole-Set 集合,用户的点击分布随着结果位置的增加而减小,第一位的点击约占 30.3%的用户点击量.对于 Rel-Set 集合,我们发现其有类似的特性,但其递减幅度更大,第一位的点击比例更大,达到 47.8%.其第一位的点击可靠性度 *CRV* 值为 1.58.这说明虽然用户点击存在点击偏置,但搜索引擎系统的智能对点击的可靠性影响更大,主要体现在第一位返回结果的点击上.从两者的分布差异可以看出,搜索引擎有着较好的智能,检索算法本身也有较好的效果.我们可以推测搜索引擎企业对第一位结果非常重视,可能通过各种策略和算法来优化搜索的第一位结果.

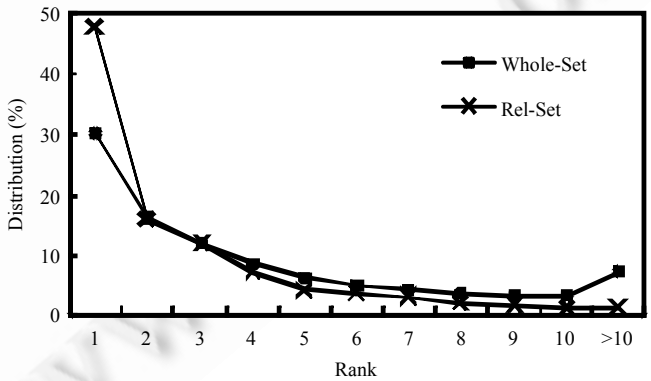


Fig.7 Distributions of clicks at rank positions

图 7 点击结果位置的点击分布

3.5 点击筛选实验与结果分析

3.5.1 实验设置和方法简述

本节将利用上文分析得到的点击上下文背景特征对用户点击进行评估,筛选出可靠性强的点击.根据第 2.2

节的公式(3),我们可以利用单个特征对点击的可靠性进行评估.基于朴素贝叶斯假设,我们可以综合多个特征对用户点击进行评价.实验中使用贝叶斯学习器对用户点击日志进行学习和测试,输出相关点击的概率值,用于评价点击可靠性,判断点击质量.选择使用连续的概率值作为结果输出,而不是二值化的分类结果,有利于根据实际的应用背景选择不同比例的高质量点击.

本文将第 3.1 节中描述的用户日志 Whole-Set 和标注数据 Rel-Set 作为学习和测试的样例,其中 2/3 样例用于训练,1/3 样例用于测试.根据学习器输出的可靠性评分对用户点击进行筛选.在性能评估上,同时考虑召回率和精度之间的关系,本文使用 ROC 曲线的 AUC(area under the roc curve)评价指标.该指标在质量评估的相关研究中被广泛采用[20].ROC 曲线能够评价当选择不同阈值时,不同召回率和精度之间难以比较的问题,以 FPR(false positive rate)为横坐标、TPR(true positive rate)为纵坐标绘制曲线.AUC 值为 ROC 曲线下的面积,可用于评判方法的有效性,其含义为将正确结果排在错误结果之前的概率.

3.5.2 实验结果与分析

通过贝叶斯学习器的学习和测试之后,我们根据得到的点击可靠性评分对日志中的点击在 Whole-Set 和 Rel-Set 上进行排序,并绘制成对应的 ROC 曲线(如图 8 所示).通过计算 ROC 曲线下的面积,我们得到 AUC 值为 0.792,意味着基于我们的算法有 79.2%的概率将一个相关点击排在不相关点击之前.作为对比,我们同时绘制了从 Whole-Set 集合中随机筛选点击的 ROC 曲线(图 8 中的虚线),其 AUC 值为 0.5.

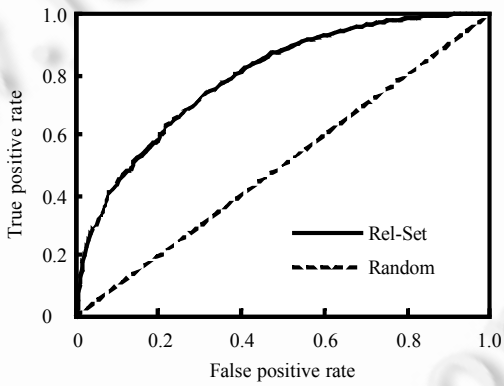


Fig.8 ROC curves to evaluate the performance of reliable click selection method, compared with random selection method

图 8 可靠性点击选择算法和随机选择算法的 ROC 曲线性能比较

从 ROC 曲线我们可以看出,通过对点击进行评估,能够筛选出可靠性高的用户点击,过滤不可靠的点击,降低可能存在的点击噪音.表 3 列出了过滤掉 80%点击日志时,算法能够保留 60%的相关点击;当过滤掉 40%点击日志时,算法能够保留 92.8%的相关点击,明显优于随机筛选用户点击日志的方式.因此,经过算法过滤的点击其有较大的可靠性可用于相关性判断.

Table 3 Data filter ratio on Rel-Set and Whole-Set

表 3 Whole-Set 和 Rel-Set 集合上的不同数据过滤比例

Data set	Filtering ratio (%)			
Whole-Set	20.0	40.0	60.0	80.0
Rel-Set	60.0	81.4	92.8	98.4

上述实验结果表明,通过分析用户在点击过程中的上下文背景特征,推测用户行为过程中的思维决策过程,能够有效挖掘用户点击过程中的偏好,筛选出可靠性强的相关点击.

4 结论和未来的工作

网络用户行为蕴含大量有价值的信息,对于信息检索和知识挖掘都具有重要的作用.本文在结合相关已有研究成果的基础上,基于大规模真实网络用户的访问日志,对用户在检索过程中的点击行为和思维决策过程展开分析和研究.研究结果表明,通过对个体用户点击行为的上下文背景环境分析,我们能够对用户在点击过程中的思维决策过程有所判断和定位,进而可以对点击的可靠性给予有效评估.本文的主要结论有:

(1) 基于信息熵概念,用户提交查询和点击的确定性程度和用户点击的可靠性相关.

(2) 通过用户点击序列的分析,Session 粒度下的用户首次和最后一次点击有着较高的可靠性,而查询粒度的用户首次点击和最后一次点击对判断点击可靠性基本无效.

(3) 搜索结果位置对结果的相关性有一定的影响,返回的首位结果具有较高的可靠性.

(4) 用户的决策过程和搜索引擎的结果影响着用户的行为,通过用户行为的上下文背景可以猜测用户的思维决策.

(5) 利用用户点击的上下文背景特征,能够从日志中筛选出具有较高可靠性的点击,提高点击质量.

通过对用户点击的准确评估,允许我们在更广的查询空间里处理用户的查询请求,如可以更加准确地处理长尾词查询、添加个性化搜索等,而不再局限于已有的基于统计分析的热门用户查询词.未来的工作也主要将沿着这些新的应用进行展开,改进搜索引擎系统的性能.

References:

[1] Baeza-Yates R, Tiberi A. Extracting semantic relations from query logs. In: Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2007). New York: ACM Press, 2007. 76−85.

[2] Fuxman A, Tsaparas P, Achan K, Agrawal R. Using the wisdom of the crowds for keyword generation. In: Proc. of the 17th Int'l Conf. on World Wide Web (WWW 2008). New York: ACM Press, 2008. 61−70.

[3] Joachims T, Freitag D, Mitchell T. WebWatcher: A tour guide for the world wide Web. In: Proc. of the 15th Int'l Joint Conf. on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1997. 770−777.

[4] Agichtein E, Brill E, Dumais S, Ragno R. Learning user interaction models for predicting Web search result preferences. In: Proc. of the 29th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2006). New York: ACM Press, 2006. 3−10.

[5] Joachims T, Granka L, Pan B, Hembrooke H, Gay G. Accurately interpreting clickthrough data as implicit feedback. In: Proc. of the 28th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2005). New York: ACM Press, 2005. 154−161.

[6] Agichtein E, Brill E, Dumais S. Improving Web search ranking by incorporating user behavior information. In: Proc. of the 29th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2006). New York: ACM Press, 2006. 19−26.

[7] Dou ZC, Song RH, Yuan XJ, Wen JR. Are click-through data adequate for learning Web search rankings? In: Proc. of the 17th ACM Conf. on Information and Knowledge Management (CIKM 2008). New York: ACM Press, 2008. 73−8.

[8] Liu YQ, Cen RW, Zhang M, Ru LY, Ma SP. Automatic search engine evaluation based on user behavior analysis. Journal of Software, 2008,19(11):3023−3032 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/3023.htm [doi: 10.3724/SP.J.1001.2008.03023]

[9] Tan PN, Kumar V. Modeling of Web robot navigational patterns. In: Proc.of the ACM WebKDD Workshop. New York: ACM Press, 2000.

[10] Tan PN, Kumar V. Discovery of Web robot sessions based on their navigational patterns. Data Mining and Knowledge Discovery, 2002,6(1):9−35. [doi: 10.1023/A:1013228602957]

[11] Baeza-Yates R, Hurtado C, Mendoza M, Dupret G. Modeling user search behavior. In: Proc. of the 3rd Latin American Web Congress (LA-WEB 2005). Washington: IEEE Computer Society Press, 2005. 242−251.

[12] Kammenhuber N, Luxenburger J, Feldmann A, Weikum G. Web search clickstreams. In: Proc. of the 6th ACM SIGCOMM Conf. on Internet Measurement (IMC 2006). New York: ACM Press, 2006. 245−250.

[13] Sadagopan N, Li J. Characterizing typical and atypical user sessions in clickstreams. In: Proc. of the 17th Int'l Conf. on World Wide Web (WWW 2008). New York: ACM Press, 2008. 885−894.

[14] Craswell N, Zoeter O, Taylor M, Ramsey B. An experimental comparison of click position-bias models. In: Baeza-Yates R, Boldi P, Ribeiro-Neto B, Cambazoglu BB, eds. Proc. of the Int'l Conf. on Web Search and Web Data Mining (WWW 2008). New York: ACM Press, 2008. 87−94.

[15] Guo F, Liu C, Wang YM. Efficient multiple-click models in Web search. In: Baeza-Yates R, Boldi P, Ribeiro-Neto B, Cambazoglu BB, eds. Proc. of the 2nd ACM Int'l Conf. on Web Search and Data Mining (WSDM 2009). New York: ACM Press, 2009. 124−131.

[16] Zhang L, Li YN, Wang B, Li P, Jiang ZF. Session segmentation based on query logs of Web search. Journal of Chinese Information Processing, 2009,23(2):54−61. (in Chinese with English abstract).

[17] Shannon CE. A mathematical theory of communication. Bell System Technical Journal, 1948,27:379−423, 623−656.

[18] Agrawal R, Halverson A, Kenthapadi K, Mishra N, Tsaparas P. Generating labels from clicks. In: Baeza-Yates R, Boldi P, Ribeiro-Neto B, Cambazoglu BB, eds. Proc. of the 2nd ACM Int'l Conf. on Web Search and Data Mining (WSDM 2009). New York: ACM Press, 2009. 172−181.

[19] Yu HJ, Liu YQ, Zhang M, Ru LY, Ma SP. Research in search engine user behavior based on log analysis. Journal of Chinese Information Processing, 2007,21(1):109−114 (in Chinese with English abstract).

[20] Svore K, Wu Q, Burges CJC, Raman A. Improving Web spam classification using rank-time features. In: Proc. of the 3rd Int'l Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2007), Vol.215. New York: ACM Press, 2007. 9−16.

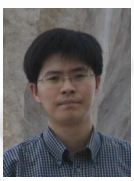
附中文参考文献:

[8] 刘奕群,岑荣伟,张敏,茹立云,马少平.基于用户行为分析的搜索引擎自动性能评价.软件学报,2008,19(11):3023−3032. http://www.jos.org.cn/1000-9825/19/3023.htm [doi: 10.3724/SP.J.1001.2008.03023]

[16] 张磊,李亚楠,王斌,李鹏,蒋在帆.网页搜索引擎查询日志的 session 划分研究.中文信息学报,2009,23(2):54−61.

[19] 余慧佳,刘奕群,张敏,茹立云,马少平.基于大规模日志分析的网络搜索引擎用户行为研究.中文信息学报,2009,21(1):109−114.

岑荣伟(1982－),男,浙江慈溪人,博士生,主要研究领域为信息检索,机器学习.



茹立云(1979－),男,博士生,主要研究领域为自然语言处理,网络信息检索.

刘奕群(1981－),男,博士,助理研究员,主要研究领域为信息检索.

马少平(1961－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为知识工程,信息检索,汉字识别与后处理,中文古籍数字化.

张敏(1977－),女,博士,副教授,主要研究领域为机器学习,信息检索.