计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2017.20160804
54(6):1133-1143,2017

搜索引擎用户满意度评估

刘奕群
(清华大学计算机科学与技术系 北京 100084)
(yiqunliu@tsinghua.edu.cn)

Satisfaction Prediction of Web Search Users

Liu Yiqun
(*Department of Computer Science & Technology, Tsinghua University, Beijing* 100084)

Abstract User satisfaction is one of the prime concerns for Web search related studies. It is a non-trivial task for three major reasons: 1) Traditional approaches for search performance evaluation mainly rely on editorial judgments of the relevance of search results. The relationship between search satisfaction and relevance-based evaluation still remains under-investigated. 2) Most existing researches are based on the hypothesis that all results on search result pages (SERPs) are homogeneous while a variety of heterogeneous components have been aggregated into modern SERPs to improve search performance. 3) Most existing studies on satisfaction prediction primarily rely on users' click-through and query reformulation behaviors but there are plenty of search sessions without such information. In this paper, we summarize our recent efforts to shed light on these research questions. Firstly, we perform a laboratory study to investigate the relationship between relevance and users' perceived usefulness and satisfaction. After that, we also investigate the impact of vertical results with different qualities, presentation styles and positions on search satisfaction with specifically designed SERPs. Finally, inspired by recent studies in predicting result relevance based on mouse movement patterns, we propose novel strategies to extract high quality mouse movement patterns from SERPs for satisfaction prediction. Experimental results show that our proposed method outperforms existing approaches in heterogeneous search environment.

Key words search satisfaction; relevance; aggregated search; mouse movement; Web search engine

摘 要 用户满意度评估一直是互联网搜索领域的研究热点,并具有3方面的挑战:1)传统的搜索性能评估方法大多基于对检索结果相关性的标注,但大多数基于相关性标注的评价指标并非针对互联网搜索环境而设计,其结果与搜索用户主观满意度之间的关系缺乏相应研究;2)大多数已有的工作都基于搜索结果同质化的假设,但随着搜索引擎的发展,异质化的搜索结果元素开始频繁地出现在搜索结果列表中;3)已有的关于搜索满意度评估的工作主要基于用户的点击和查询修改行为开展,但实际搜索中会有大量的用户会话中缺失此类信息.总结了近期为解决这些研究问题开展的实验研究工作:1)构建了用户行为实验系统,分析了结果相关性与用户所感知到的结果效用和满意度之间的关系;2)基于仔细设计的异质化搜索结果页面,定量地分析了垂直搜索结果的质量、展现形式、位置等因素对用户满意度的影响;3)受现有的采用鼠标移动信息进行搜索结果相关性预测的工作启发,提出了在搜索结果页面上抽取用户

收稿日期:2016-11-10;**修回日期**:2017-02-17
基金项目:国家自然科学基金优秀青年科学基金项目(61622208)
This work was supported by the National Natural Science Foundation of China for Excellent Young Scientists (61622208).

鼠标移动行为模式并进行满意度评估的方法.实验结果表明:在真实搜索环境下,所提出的方法优于现有的模型.

关键词 搜索满意度;相关性;垂直搜索;鼠标移动信息;网络搜索引擎

中图法分类号 TP391

性能评价作为检索系统改进排序算法、检测困难查询、优化查询系统的重要依据和保障,一直是信息检索领域关注的研究热点. Cleverdon 提出的 Cranfield 评价体系[1]是检索系统(包括搜索引擎)性能评价方面的经典研究框架,该体系通过标准查询输入下系统输出与标准输出的差异来衡量检索系统的性能优劣.然而,随着网络搜索相关技术的发展与普及,传统的以"文档—查询"相关性标注为主要依据的 Cranfield 体系在实际应用中体现出了越来越多的局限性[2].对搜索引擎性能的评价模式因而变得越来越多样化,而对搜索引擎用户的满意度评估就是其中的一个重要的研究方向. Su[3]在 20 世纪 70 年代首次将满意度评估引入到信息检索领域.随后,Kelly 完善了对满意度的概念定义,将其定义为"用户满足特定信息需求或达成特定信息获取目的的程度"[4].对搜索用户的满意度评估结果能够为搜索引擎商业运营带来最直观的性能描述,也因此受到搜索产品研发人员和搜索广告商的高度重视.

虽然已有学者进行了大量的搜索满意度评估相关研究,但当前关于如何合理量化估计用户的搜索满意度方面,仍然存在 3 方面的研究挑战:1)基于搜索满意度的评价结果和传统的基于"文档-查询"相关性标注的 Cranfield 评价结果之间的关系缺乏定量研究,搜索用户满意度和搜索结果质量之间的关联关系也需进一步研究. 2)现有的大多数搜索满意度研究都基于同质化的搜索结果页面(search engine result pages, SERPs)的假设进行,即搜索结果页面上的所有结果都具有相同的展现形式:一个带超链接的标题和一个短摘要.近年来,随着商业搜索引擎的快速发展,越来越多的以异质化形式呈现的垂直结果(视频、图片、知识图谱等)出现在搜索结果页面,用户的检验和点击行为也随之发生了显著的改变[5-6],但用户异质搜索环境下的满意度感知过程仍然鲜有研究. 3)与采用点击信息即可达到较好评估效果的搜索结果相关性不同,用户的搜索满意度可能与大量的交互行为细节有关,大多数现有满意度评估方法往往仅基于点击行为或查询修改行为开展[7-8],但在相当多的真实搜索会话中点击和查询修改往往均不存在[9-10],这就导致了传统预测方法的失效.

基于以上这 3 方面研究难点,在本文中,我们尝试提出并回答 3 个问题:

1) 搜索满意度评估与传统的基于 Cranfield 体系的搜索引擎相关性评价有什么联系与差别?

2) 异质搜索环境下的用户满意度感知会受到哪些因素的影响?

3) 有哪些用户交互行为特征可以协助我们更好地预测搜索满意度?

对于问题 1,我们设计用户行为实验,从相关性标注可能与用户真实感受到的文档效用(usefulness)存在差异出发,分析了文档效用和满意度标注之间的差别,并进一步研究说明了效用、相关性和搜索满意度之间的关系,对搜索满意度和基于相关性的评价指标之间存在的差异进行了解释.

对于问题 2,我们定量研究了不同展现形式、位置与质量的垂直搜索结果对用户满意度感知造成的影响,以深入分析用户在真实异质搜索环境下的满意度感知过程.

对于问题 3,考虑到鼠标移动信息包含大量丰富的用户与搜索引擎的交互细节[11],并可以被低成本地大规模收集,我们尝试从鼠标移动数据中挖掘不同满意度用户会话之间的行为模式差异,并将其运用到搜索满意度评估的任务中.

为了研究以上 3 个问题,我们搭建了一个实验性的搜索引擎系统并用以收集用户在完成搜索任务时的交互行为信息,同时也收集用户的满意度反馈以及标注人员对搜索结果的相关性与效用标注信息.本文的主要贡献包括:

1) 系统分析了相关性、效用和搜索满意度之间的关系,并提出了 2 种能够有效估计用户实际感受的文档效用的方法;

2) 对异质环境对搜索满意度的影响进行了系统的研究,并定量分析了垂直结果的质量、展现形式、展现位置等因素对搜索满意度的影响;

3) 建立了采用鼠标移动模式在搜索环境下进行搜索满意度评估的方法框架,并提出了基于距离

差异和基于分布差异的2种鼠标移动模式筛选方法,相比传统方法而言获得了显著的效果提升.

1 网络搜索用户实验

为了研究搜索满意度,我们组织了一系列用户实验(user study),以收集包含用户显式满意度反馈的搜索行为记录.在用户实验过程中,我们通过招募参与者在实验室环境下使用实验搜索引擎系统完成一系列搜索任务,出于模拟真实的搜索引擎使用场景的需要,所有的查询任务均从搜索引擎查询日志中筛选随机筛选中频查询得到.之所以选取中频查询,一方面是由于低频查询中包含着大量低质量乃至拼写错误、内容上非法的查询,使用这些查询作为用户实验的对象会给实验对象造成很大不便;另一方面,搜索引擎对于高频查询的处理技术已经相当成熟,甚至有不少高频查询结果是经过手工编辑校订而非自动生成的,这对于我们考察普遍情况下搜索用户满意度情况的实验目的构成了障碍.为了确保所有的用户在进行同样的查询任务实验时信息查找需求的一致性,我们针对每个查询都编写了简短的查询任务说明,明确说明了该查询的搜索场景、何种资源是用户需要的,以避免可能的歧义影响.

对于不同的研究问题,我们开展3方面工作:1)设置了不同的搜索任务;2)利用实验搜索引擎实现了控制返回给用户的结果、对用户进行问卷调查、收集包括搜索满意度反馈在内的显式反馈、自动记录搜索行为信息等功能;3)对收集到的用户搜索行为数据进行了不同类型的人工标注.图1展示了用户实验的一般流程.本节接下来将简要介绍我们为了解决前述3个研究问题所组织的3次用户实验.

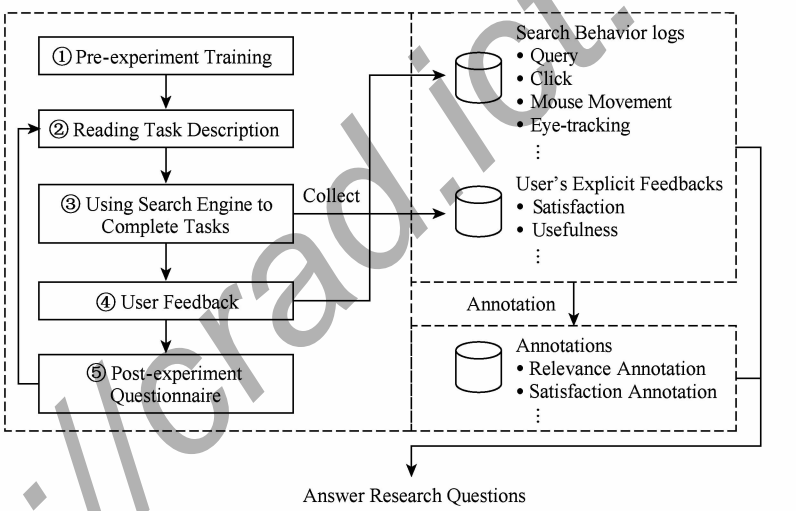

Fig. 1 A general protocol of Web search user studies

图1 网络搜索用户实验的一般流程

1) 用户实验1

为了分析相关性、效用和搜索满意度之间的关系,在用户实验1中,我们招募了29位参与者,在实验室环境下使用我们搭建的实验搜索引擎分别完成了12个搜索任务.为了模拟真实的搜索环境,我们所搭建的实验搜索引擎界面与目前主流的商业搜索引擎类似,支持用户自由地进行查询改写和翻页等操作.

在参与者使用实验搜索引擎完成每一个搜索任务之后,实验系统会要求参与者回顾整个搜索过程,对每个点击过的文档提交4级(1~4分)的效用反馈,对每个提交的查询提交5级(1~5分)的查询级别满意度反馈,最后对整个任务提交5级的任务级别满意度反馈.图2展示了参与者进行效用和满意度反馈时的操作页面.

在进行用户实验收集了包含用户显式的效用和满意度反馈的搜索行为记录之后,我们请第三方标注者对其进行了相关性、效用和满意度标注.我们对所有参与者点击过的结果进行了4级(1~4分)相关性和效用标注;对每个参与者提交的查询进行了查询级别的5级搜索满意度标注;对每个搜索任务进行了任务级别的5级搜索满意度标注;最后,为了分析搜索满意度与传统搜索性能评价指标之间的关系,我们还对每个查询返回的前5条结果进行了4级相关性标注.在进行相关性标注时,我们会将查询连同结果标题和摘要展示给标注人员,并要求标注

人员在点击标题查看结果内容后进行相关性标注,每个查询-结果对会被至少 3 位标注人员独立地标注.效用和搜索满意度标注是同时进行的,为了重现搜索上下文环境,我们会将参与者在完成该搜索任务时完整的行为记录展示给标注人员.这些行为记录包括在整个搜索会话中参与者提交的查询、点击的结果,以及在每个查询和点击的结果上的停留时间等信息.图 3 展示了标注人员进行效用和搜索满意度标注的标注页面.我们同样要求每个搜索任务记录被至少 3 个标注人员独立地标注.

Fig. 2 User feedback interface

图 2 用户反馈页面

Fig. 3 Usefulness and satisfaction annotation interface

图 3 文档效用和搜索满意度标注页面

2) 用户实验 2

为了定量研究异质化结果对搜索满意度的影响,在用户实验 2 中我们对搜索结果页面的垂直结果的质量、展现形式、展现位置 3 个因素分别进行控制.

在用户实验 2 中,我们招募了 35 位参与者,每个参与者需要完成 30 个查询任务,这些查询任务都是日常搜索中常见的中频查询,从商业搜索引擎的大规模搜索日志中采样得到.该实验等概率地在搜索结果页面的第 1,3,5 位插入了文本、图文、图片、下载、新闻这 5 种展现形式的垂直结果,垂直结果的质量(相关/不相关)也随机地进行了控制.在完成每个查询任务时,参与者会首先阅读事先给定的查询词和详细的查询任务描述,随后他将被导引到一个搜索结果页面,该页面的查询词固定,并有 10 个事先从商业搜索引擎抓取的查询结果.参与者可以按照自己的习惯随意浏览、点击系统所提供的搜索结果,如果参与者完成了查询任务,或者认为系统所提供的结果不足以满足需求,就可以结束该查询.

参与者每完成一个查询,都会被要求给一个 1~5 分的 5 级满意度反馈,其中 5 表示对刚刚完成的搜索体验最满意,1 表示对该搜索体验最不满意,

随后用户就可以开始下一个查询任务.每个参与者在真正开始任务之前,会先做2个不记录数据的查询任务用以熟悉实验流程.

在所有的参与者完成实验后,我们还邀请了专业的标注人员参照用户的浏览日志进行满意度标注.每个参与者的实验过程均被完整地录制,并被提供给标注人员作为参考,以确保标注人员可以最大程度复现用户当时的搜索过程.标注人员需要给出和用户同标准的5级满意度反馈,每个用户搜索日志会有2个标注人员进行标注,标注人员在整个数据集上的标注一致性为0.48[12].

3) 用户实验3

为了研究鼠标移动模式在真实搜索环境下的预测效果,在用户实验3中我们采用与真实搜索环境完全一致的搜索结果页面.

在用户实验3中,我们招募了30位参与者,每人仍需完成与用户实验2中相同的30个查询任务,与用户实验2中所不同的是,用户实验3中的搜索结果页面是从商业搜索引擎直接抓取获得,没有做任何的结果筛选或变量控制,因而与真实环境完全一致,实际得到的搜索结果列表中,平均一个查询任务的10个搜索结果中包含7.4个垂直结果.

2 相关性、效用和搜索满意度

相关性(relevance)是信息检索领域内一个非常重要的概念.根据概率排序原则(probability ranking principle[13]),对于用户提交的查询,搜索引擎应该尽可能返回一个按照相关性从高到低排序的结果列表.而在传统的Cranfield检索评价方法[1]中,为了比较不同搜索引擎的有效性(effectiveness),我们需要构建一个包含待查询语料库、用于测试的信息需求和查询集合以及相应查询-结果对的相关性标注的测试集合.基于查询-结果对的相关性标注信息,我们可以使用一系列搜索性能评价指标(如MAP,$nDCG$[14],ERR[15]等),对搜索引擎针对测试查询返回的结果列表进行评价.

在理想状况下,相关性标注能反映一个结果文档是否能满足用户的信息需求,那么基于其计算的搜索引擎评价指标就能较好地反映用户的搜索满意度.但在实际中,我们往往无法从真实的搜索用户那里获得相关性反馈信息,而只能依赖第三方标注人员进行相关性标注.在这种情况下,第三方标注人员只能根据提交的查询猜测和估计真实用户的信息需求.并且,他们很少能获知真实用户提交查询时的搜索上下文信息,而只能独立地对每一个查询-结果对进行相关性标注.这些限制使得相关性标注往往只能基于查询和结果文档是否在主题层面上相关,进而可能与用户实际感受到的结果文档的效用(usefulness)存在较大差异.

针对以上问题,我们通过设计和组织用户实验,尝试对相关性(relevance)、效用(usefulness)和搜索满意度(satisfaction)三者之间的关系进行研究和分析[16].基于用户实验收集到的数据,我们提出并尝试回答4个子研究问题:

1) 用户感受到的效用和第三方标注人员的相关性标注之间是否存在差异?

2) 来自用户的效用反馈和来自标注人员的相关性标注与用户的搜索满意度之间存在怎样的联系?

3) 由于在实际应用中,我们不能获得来自用户的效用反馈信息.那么,我们能否依赖第三方标注者来获得文档级别的效用标注?

4) 我们能否基于搜索日志中的用户行为和搜索上下文特征,自动地生成可靠的效用标签?

2.1 相关性和效用之间的区别

基于第1节用户实验1中获得的数据,我们首先分析了标注者提供的相关性标注和用户提供的文档效用反馈之间的差异和联系.图4(a)展示了4级相关性标注R和4级效用反馈U_u之间的联合概率分布情况.

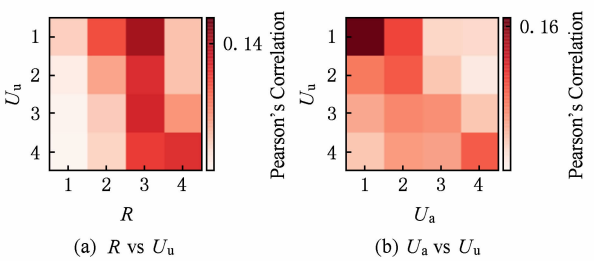

Fig. 4 Joint distribution of usefulness feedbacks and annotations

图4 用户效用反馈与标注者之间的联合概率分布

从图4(a)中我们发现,尽管相关性标注和效用反馈之间存在一定正相关关系(Pearson相关系数,Pearson's),它们之间存在系统性的误差.一方面,在被点击的结果中,只有一小部分相关性级别为1或2的结果,被用户认为是有用的(效用级别为3或4);但在另一方面,有大量相关性级别为3的结果被用户认为是对完成任务完全没有帮助(效用反馈级别为1).这说明,结果文档的相关性和其效用并不

完全一致,结果的相关性高是结果能给用户带来高效用的必要非充分条件.

2.2 相关性、效用和搜索满意度的联系

由于我们发现相关性标注和用户反馈的结果效用并不一致,那么从搜索引擎评价和搜索满意度分析的角度我们进一步分析了相关性和效用2种文档级别的指标与查询级别的搜索满意度之间的关系.

由于参与者只对被点击结果提供了效用反馈,所以,我们分别使用被点击结果的效用反馈和相关性标注计算了4种基于点击序列的在线评价指标:点击累积增益(cCG)、点击衰减累积增益($cDCG$)、最大点击增益($cMAX$)和平均点击增益($cCG/\#clicks$).为了进行对比,我们也基于前5位搜索结果,用相关性标注计算了包括DCG,MAP和ERR在内的传统评价指标.我们通过计算它们与查询级别的搜索满意度反馈之间的Pearson相关系数r来衡量不同指标与搜索满意度之间联系的密切程度.

从结果中我们发现:

1) 基于效用反馈计算的4种在线评价指标与搜索满意度之间的相关系数均显著地比基于相关性标注计算的同一种评价指标对应的相关系数高.例如,基于效用反馈U_u计算的$cDCG(U_u)$与搜索满意度的Pearson相关系数$r=0.724$,而基于相关性标注R计算的$cDCG(R)$的相关系数$r=0.498$.其中与搜索满意度相关系数最高的在线评价指标为利用效用反馈U_u计算的$cMAX(U_u)$(Pearson相关系数$r=0.751$).

2) 4种基于点击序列的在线评价指标与搜索满意度之间的相关系数明显高于传统的基于搜索结果列表的评价指标.传统评价指标中与满意度相关系数最高的是$DCG@5$,相关系数$r=0.295$,显著地低于基于点击序列计算的$cDCG(R)$和$cDCG(U_u)$与搜索满意度之间的相关系数($r=0.498$和$r=0.724$).

以上发现说明,效用反馈和在线评价指标相较于相关性标注和基于结果排序的传统评价指标,与用户实际的搜索满意度联系更为密切.未来我们可以基于它们来构建一个更为贴近用户的搜索引擎评价方法.

2.3 效用标注和效用预测

由于在实际中,我们无法从真实用户那里获得效用反馈,所以,我们尝试了效用标注和效用预测2种方法来获得可靠而有效的效用标签.

首先,由于我们认为导致相关性标注不能很好地反映结果的真实效用的一个重要原因,是在进行相关性标注时标注人员无法获知真实的搜索上下文信息和用户行为信息.所以,在2.1节介绍的效用标注过程中,我们将这些信息提供给标注者,并要求标注者给出与搜索上下文相关的效用标注.通常可以用来自多个独立的标注者之间的标注一致程度(常用Cohen's κ统计量衡量)来衡量标注数据的可靠性(reliability).我们收集到的4级效用标注的κ值为0.530,达到了中等(moderate)的一致性水平,并且比4级相关性标注的κ值(0.413,达到合理(fair)的一致水平)更高.这说明,以传统的相关性标注为基准,我们提出的效用标注方法是可靠的.提供搜索上下文信息和用户行为信息实际上能帮助标注者做出更加可靠的判断.

其次,由于效用标注仍然需要标注人员参与,存在费时费力的问题.我们还尝试使用机器学习方法,利用用户行为和搜索上下文特征,自动地进行效用预测.我们使用的特征主要包括:查询级别(Q)的结果位置、查询长度、点击数量和点击停留时间;任务级别(S)的查询数量、无点击查询数量、任务完成时间和查询改写策略;用户级别(U)的点击、查询数量和停留时间的最小、最大和平均值.我们将效用预测当作一个利用上述特征预测用户实际的效用反馈的回归问题,并使用Gradient Boosting Regression Tree(GBRT)模型[17]作为我们的回归模型.

衡量效用标注和效用预测2种不依赖用户反馈生成效用标签的方法是否有效(valid),最直接的方法就是比较它们和用户实际的效用反馈是否一致.图4(b)中展示了效用标注U_a和作为最终标准的效用反馈U_u之间的联合概率分布.从图4(b)中可以发现,颜色较深的块均分布在对角线上,和图4(a)中的相关性R相比,效用标注U_a和真实的效用反馈U_u更为一致.

而对于效用预测,我们在表1中展示了采用不同特征组合预测得到的效用预测结果(U_Q,U_{Q+S}和U_{Q+S+U})与真实用户反馈之间的Pearson相关系数(r,越大越好)、平均平方误差(MSE,越小越好)和平均绝对误差(MAE,越小越好).我们同时列出了相关性标注和效用标注的结果作为参照.从表1中我们可以发现,利用用户行为和搜索上下文信息得到的效用预测,在与用户真实的效用反馈的一致性方面,显著地好于相关性标注,同时达到甚至超过了效用标注的水平.这说明,我们能够使用机器学习方法,有效地节省人工标注成本,自动生成可用的效用标签.

Table 1 Results for Usefulness Prediction

表 1 效用预测结果

Annotation	r	MSE	MAE
U_Q	0.398 *	1.198 **	0.894 **
U_{Q+S}	0.410 **	1.186 **	0.889 **
U_{Q+S+U}	0.461 **	1.103 **	0.851 **
U_a	0.413	1.512	0.852
R	0.332	1.786	1.020

Notes: The prediction performance is measured in Pearson's with usefulness feedback U_a. * indicates the performance is significantly different with relevance annotation R at $p<0.05$ level, and ** indicates the performance is significantly different with relevance annotation R at $p<0.01$ level. Darker shade indicates the performance is significantly different with usefulness annotation U_a at $p<0.05$ level, and lighter shade indicates the performance is significantly different with usefulness annotation U_a at $p<0.01$ level.

2.4 总 结

针对用户实际感受到的结果效用和传统相关性标注可能存在差异这一问题,我们通过用户实验收集了一个包含用户搜索行为记录、用户效用和满意度反馈,以及相应的相关性、效用和满意度标注的完整数据集.基于该数据集,我们系统地分析了效用、相关性和搜索满意度之间的关系,发现:1)结果相关性和结果效用并不完全一致,结果相关是结果能给用户带来效用的必要非充分条件;2)基于效用计算的在线评价指标与用户的搜索满意度存在较强的正相关(Pearson 相关系数 $r>0.7$);3)我们能够通过依赖第三方标注人员的效用标注和基于搜索行为记录的效用预测来有效地估计用户真实感受到的效用.结合以上 3 点发现,我们认为在未来可以基于文档级别效用标签和基于点击序列在线评价指标,设计一种更接近用户搜索满意度的搜索引擎评价方法.

3 异质化环境下的搜索满意度

基于第 1 节中用户实验 2 中获得的数据,我们可以研究垂直结果的质量、展现形式和展现位置等因素对搜索满意度的影响[18].

由于不同参与者对于满意度的感知标准可能会有所差别,所以在进行实验分析之前,我们首先将每个用户所给出的满意度反馈按照 Z-score[18] 进行了归一化,以从一定程度上去除不同参与者的主观性因素影响.

表 2 反映了不同展现形式的垂直结果对搜索满意度的影响.表格中的第 2~4 列的数值是相应类型垂直结果所对应的搜索日志的平均满意度,括号里的数值反映的是在带有对应类型垂直结果的情况下的满意度与不带垂直结果的情况下搜索满意度的差异.

Table 2 Effect of Verticals with Different Presentation Styles on Satisfaction

表 2 不同展现形式的垂直结果对搜索满意度的影响

Vertical Type	No vertical	on-topic vertical	off-topic vertical
Textual	5.15	5.10 (−0.05)	4.95 (−0.20 **)
Image & Textual	4.46	4.99 (+0.53 **)	4.67 (+0.21)
Image	5.17	5.07 (−0.10)	4.58 (−0.59 **)
Download	4.75	5.25 (+0.50 **)	4.60 (−0.15)
News	4.43	4.34 (−0.09)	4.38 (−0.05)

Notes: * indicates statistical significance at $p<0.1$ level, and ** indicates statistical significance at $p<0.05$ level. The values in the parentheses indicate the difference with no vertical values.

从表 2 中可以看到,与页面中没有垂直结果的情况相比,在页面中插入相关的图文类和下载类垂直结果可以使用户和标注者显著地感到更加满意.在页面中插入相关的图片类垂直结果并不会使用户更满意,这可能是因为能从图片中获取的信息往往也能够相对容易地在普通文本结果中获得.而如果在页面中插入了不相关的图片,则会显著地使用户的满意度降低,这是因为图片结果容易引人注意,而不相关的内容就会引起用户的不悦.新闻类的垂直结果对用户满意度没有显著的影响,并且除了新闻类垂直结果之外,在页面中插入相关的另外 4 种类型的垂直结果都会比插入不相关的相应垂直结果更容易让用户感到满意.

我们进一步研究了垂直结果在页面中不同位置时对用户满意度的影响,相关结果如表 3 所示.垂直结果被放在整个页面中的第 1、第 3、第 5 个位置进行效果的对比.从表 3 可以看到,当相关的垂直结果放在页面的高位时,会对用户的满意度带来显著提升;当有不相关的垂直结果放在页面的首位时,用户会明显地感觉到不满意;而当不相关的结果放在其他位置时,搜索满意度不会受到明显的影响.

基于以上在异质化搜索领域的研究,可以发现

异质化结果的存在确实会对用户的满意度感知带来显著的影响,可以总结为4点:1)相关的图文类和下载类垂直结果会带来显著的搜索满意度提升;2)相关的图片类垂直结果对提升用户满意度影响不大,但不相关的图片结果会显著降低搜索满意度;3)新闻类垂直结果对搜索满意度没有明显的影响;4)当垂直结果放在搜索结果列表中的位置越靠前,对用户满意度的影响越大.

Table 3 Effect of Ranking Positions of Verticals on Satisfaction

表3 垂直结果的位置对搜索满意度的影响

Vertical Position	No vertical	on-topic vertical	off-topic vertical
Rank 1	4.79	5.06 (+0.27 **)	4.43 (−0.36 **)
Rank 3	4.79	4.93 (+0.14)	4.63 (−0.16)
Rank 5	4.79	4.87 (+0.08 *)	4.85 (+0.06)

Notes: * indicates statistical significance at $p<0.1$ level, and ** indicates statistical significance at $p<0.05$ level. The values in the parentheses indicate the difference with no vertical values.

4 基于鼠标移动信息的搜索满意度评估

Lagun等人首次提出鼠标移动模式(mouse movement motif)的概念[19]. 他们将鼠标移动模式定义为频繁出现的鼠标位置序列,并实现了在搜索着陆页(landing page)上自动挖掘鼠标移动模式的算法来进行有效的搜索结果相关性预测. 在本节中,我们尝试进一步改进鼠标移动模式的抽取和筛选算法,在搜索结果页面上直接抽取鼠标移动模式,并将其运用到预测搜索满意度的任务中[20].

图5展示了用户在搜索结果页面上的鼠标移动轨迹的2个示例,图5(a)展示的是一个用户反馈为满意的例子,图5(b)是一个用户反馈为不满意的例子. 鼠标移动轨迹用带数字的圆表示,圆中数字由小到大表示了鼠标移动的顺序,图中红圈是我们的算法挖掘到的鼠标移动模式. 从图5(a)中可以看到,用户仔细地检验了第1个结果(能完成对应查询任务的关键结果),随后快速浏览了其他结果,然后就结束了查询. 鼠标移动轨迹显示他只花了相对小的成本就找到了完成任务所必需的信息. 作为对比,图5(b)中的大多数结果都难以满足用户的需求,鼠标移动轨迹显示该用户检验了页面上的很多结果,甚至检验了页面最下方的结果. 这意味着用户花了很大代价却只获得了很少的有用信息. 图5的例子说明,搜索结果页面上的鼠标移动轨迹包含了丰富的用户与搜索引擎交互的信息,可以帮助我们预测用户满意度.

(a) Example of a satisfied(SAT) search session (b) Example of a dissatisfied(DSAT) search session

Fig. 5 Examples of users' mouse movement trails on SERPs

图5 用户在搜索结果页面上的鼠标移动轨迹示例

4.1 鼠标移动模式抽取和筛选

在 Lagun 等人提出的算法的基础上,我们尝试从搜索结果页面直接挖掘具有高区分度的鼠标移动模式,并将其用于搜索满意度预测.

我们首先采用 Lagun 等人提出的算法[19]从整个数据集中挖掘出大量的备选鼠标移动模式,然后在此基础上进一步设计了 2 种鼠标移动模式的筛选算法.与 Lagun 等人所采用的基于频率的筛选方式不同,新提出的筛选方式充分利用了数据分布信息,能够筛选出对用户满意度具有高区分度的鼠标移动模式.为了叙述方便,我们用 SAT,$DSAT$ 分别表示被用户标注成满意和不满意的 2 类搜索会话,M_{SAT},M_{DSAT} 分别表示从 SAT,$DSAT$ 中挖掘出的鼠标移动模式.

4.1.1 基于距离差异的筛选策略

基于距离差异的筛选方法基于差异性假设:M_{SAT} 中的具有强区分度的鼠标移动模式应当与 M_{DSAT} 中的行为模式具有足够大的差异,反之亦然.参照这种准则,我们为每一个备选移动模式计算一个评分 S_{dist},对于 M_{SAT} 中的备选移动模式 C_{SAT_i},计算公式为

$$S_{\text{dist}}(C_{SAT_i})=\frac{\sum_{C_j\in M_{DSAT}} DTW(C_{SAT_i},C_j)}{|M_{DSAT}|}, \quad (1)$$

其中,$DTW(C_{SAT_i},C_j)$ 表示 2 个备选鼠标移动模式 C_{SAT_i} 和 C_j 的 DTW 距离.直观地,式(1)表示计算 M_{SAT} 中某一个移动模式与 M_{DSAT} 中所有移动模式的平均距离作为其评分.类似地,对于 M_{DSAT} 中的备选移动模式,有

$$S_{\text{dist}}(C_{DSAT_i})=\frac{\sum_{C_j\in M_{SAT}} DTW(C_{DSAT_i},C_j)}{|M_{SAT}|}, \quad (2)$$

在计算出所有备选移动模式的评分以后,我们按照该评分由大到小进行排序,并依次挑选一些具有强距离差异的鼠标移动模式.

4.1.2 基于分布差异的筛选策略

基于分布差异的筛选方法基于覆盖性假设:M_{SAT} 中的具有强区分度的移动模式应当覆盖足够多的 SAT 和足够少的 $DSAT$,反之亦然.在这种规则下,我们首先需要判断一个鼠标移动模式是否能覆盖某一个会话(表示为一个完整的光标位置的时间序列),因而,我们首先定义一个鼠标移动模式 C 与某个搜索日志 S 的距离:

$$Dist(C,S)=\min(DTW(C_i,C)\,|\,C_i\in S). \quad (3)$$

亦即,我们通过指定大小的滑动窗口在 S 中截取多个移动模式备选,而 C 与这些移动模式备选的距离中最小的一个即为 C 与会话 S 的距离.

此时我们再定义鼠标移动模式 C 在某一数据集合 D 上的覆盖率:

$$CR(C,D)=\frac{\left|\left\{\frac{|D|\,Dist(C,S_i)}{\sum_{S_i\in D}Dist(C,S_i)}<t\;|\;S_i\in D\right\}\right|}{|D|}, \quad (4)$$

其中,t 是一个定义覆盖的阈值,其作用是保证筛选出数量合适的鼠标移动模式,在我们的工作中,$t=\frac{1}{30}$.直观而言,即我们认为当 C 与某个 S 的距离足够小时,即为覆盖.

有了覆盖率的定义之后,我们就可以定义一个备选移动模式在 SAT 和 $DSAT$ 上的覆盖率的比值,作为该备选移动模式的分布差异得分:

$$S_{\text{distri}}(C_{SAT_i})=\frac{CR(C_{SAT_i},SAT)}{CR(C_{SAT_i},DSAT)}, \quad (5)$$

$$S_{\text{distri}}(C_{DSAT_i})=\frac{CR(C_{DSAT_i},DSAT)}{CR(C_{DSAT_i},SAT)}. \quad (6)$$

在我们计算出所有备选移动模式的评分以后,我们将其按照分布差异度评分由大到小进行排序,就可以挑选出具有强分布差异的鼠标移动模式.

4.2 满意度预测

在挖掘出鼠标移动模式后,按式(3)可以计算鼠标移动模式与搜索会话之间的距离,该距离就可以作为分类特征进行满意度预测.

基于第 1 节中用户实验 3 所获得的数据,我们验证鼠标移动模式对搜索满意度的预测效果.我们将数据中被用户标记为 3 的查询会话去除,因为给出 3 的评分表示用户没有明确的满意或不满意的倾向.被用户标记为 4 或 5 的查询会话被作为满意的数据样本,被用户标记为 1 或 2 的查询会话被作为不满意的数据样本.由于数据集的不平衡性,在进行训练的时候我们对满意的数据样本进行了降采样,以保证训练集的平衡性(测试集仍保持了原有的不平衡比例).我们采用 AUC 作为评价指标,因为其相比其他指标更不容易受到数据不平衡性的影响[21].所有的结果都基于 5 折交叉验证,预测所用的鼠标移动模式在每一折的训练集上都会重新计算.

图 6 比较了不同鼠标移动模式筛选方法的预测结果,横轴表示预测所采用的鼠标移动模式数量,纵轴表示五折交叉验证的 AUC 值,不同颜色的折线对应于不同的鼠标移动模式筛选方法,其中基于频率的筛选方法是 Lagun 在他们的工作中所采用的方法[19].从图 6 中可以看出,我们新提出的 2 种筛选

方式的预测效果显著优于基于频率的筛选方式,其中采用基于分布差异的筛选方法可以在只用了 50 个鼠标移动模式的时候就取得最优的预测效果,虽然基于距离的筛选方法在使用大量鼠标移动模式后也可以获得同等水平的效果,且基于分布的筛选方法在采用更多的鼠标移动模式之后,由于过拟合的原因会造成预测效果下降,但考虑到鼠标移动模式的抽取过程比较耗时,如果能使用较少的移动模式即可获得不错的效果,那将可以大大提升算法的运行效率,所以我们认为基于分布差异的筛选方式是最优的选择策略.

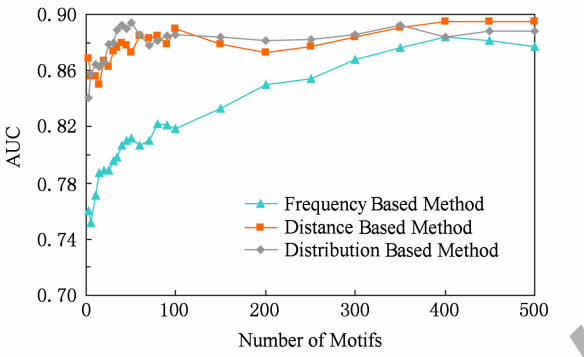

Fig. 6 Prediction performance with different motif selection strategies

图 6 不同鼠标移动模式筛选策略的预测效果

为了进一步验证鼠标移动模式对未知用户、查询的泛化能力,我们采用了 3 种不同的训练-测试集生成策略:1)随机采样. 训练集和测试集的数据划分是完全随机的. 2)按用户采样. 同一用户完成的搜索日志数据要么全在训练集中,要么全在测试集中. 3)按查询采样. 基于同一查询的搜索日志数据要么全在训练集中,要么全在测试集中.

我们实现文献[11]中的预测模型并将其作为基线方法,该方法中同时采用了点击行为等粗粒度特征和滚轮速度等细粒度指标,是当前采用鼠标行为数据进行搜索预测的最新方法之一[22]. 采用"成本-收益"预测框架对满意度进行预测,并取得了非常好的效果,因而我们也将其中的预测模型实现作为另一个基线方法. 不同预测方法在不同训练-测试集生成策略上的预测表现如表 4 所示,表 4 中的数值是 5 折交叉验证的 AUC 值,括号内的数值是基线方法和鼠标移动模式所提供的特征结合以后的方法相对于相应基线方法的效果提升,基于鼠标移动模式的方法采用了 50 个鼠标移动模式作为特征. 从表 4 中可以看到,采用鼠标移动模式可以获得与采用其他鼠标行为特征的方法相当的预测效果,当我们将鼠标移动信息整合到现有的模型中去时,在几乎所有的预测任务上都可以获得稳定的效果提升. 此外,表 4 中也体现出在不同的数据采样策略下,鼠标移动模式的预测效果基本稳定,这就表示通过小群体的查询日志提取的鼠标移动模式,可以对未知的用户及查询的搜索满意度进行很好的预测,该方法具有很强的泛化能力.

Table 4 Comparison of Different Methods for Predicting Search Satisfaction Across Different Users and Queries

表 4 不同方法对未知用户、查询的搜索满意度预测效果

Different Methods	Random Sample	Sample by User	Sample by Query
Ref[11]	0.892	0.890	0.923
Ref[22]	0.877	0.877	0.871
Motif	0.865	0.856	0.831
Motif+Ref[11]	0.932 (+4.5%)	0.936 (+5.2%)	0.925 (+0.2%)
Motif+Ref[22]	0.930 (+6.0%)	0.931 (+6.2%)	0.931 (+6.9%)

Notes: The values in the parentheses indicate the percent increase of the satisfaction with Motif compared with the original satisfaction.

5 总结与未来工作

随着网络搜索引擎的不断发展,搜索满意度这一贴近用户实际感受的评价指标日益受到研究者和搜索引擎公司的重视. 我们通过设计用户实验的方式对搜索满意度进行了全面系统的研究. 我们的研究发现:由于相关性标注与结果文档给用户带来的实际效用并不完全一致,传统的基于相关性的评价方式不能很好地估计用户实际感受到的搜索满意度. 同时,针对真实搜索环境下存在大量异质化搜索结果的现象,我们深入分析了垂直结果的质量、展现形式和展现位置对搜索满意度的影响. 最后,我们提出采用鼠标移动模式进行搜索满意度的预测,并提出了基于距离差异和基于分布差异的鼠标移动模式筛选方法,相比传统方法而言获得了显著的效果提升.

本文的研究结果可能在如下方面对商业搜索引擎的应用产生积极影响:1)用户满意度是搜索引擎性能评价的主要标准(gold standard),通过上述研究,我们成功揭示了满意度评价结果与已有的各种离线评价方法(Cranfield 方法)之间的关联关系,为更好地使用具有较强复用性和鲁棒性的离线策略拟合用户满意度评价结果、设计更合理的离线性能评价指标奠定了基础. 2)与传统用户满意度评价需要

借助真实用户反馈,耗费大量人力资源且反馈慢、结果稳定性差不同,本文尝试提出利用鼠标移动模式这一搜索引擎可以大规模采集的用户行为信号进行满意度预测,起到更高效的性能评价效果.3)本文提出的基于鼠标移动模式预测用户满意度的方法,客观上证实了这一反馈信息可以应用于搜索引擎的性能提升,我们在未来工作中将考虑应用这一思路对搜索引擎排序算法的设计(如点击模型设计)进行改进,试图在查询过程中主动利用用户反馈更好地满足用户信息需求.

参 考 文 献

[1] Cleverdon C. The Cranfield tests on index language devices [G] //Readings in Information Retrieval. San Francisco, CA: Morgan Kaufmann, 1997: 47-59

[2] Lang Hao, Wang Bin, Li Jintao, et al. Predicting query performance for text retrieval [J]. Journal of Software, 2008, 19(2): 291-300 (in Chinese)
(郎皓, 王斌, 李锦涛, 等. 文本检索的查询性能预测[J]. 软件学报, 2008, 19(2): 291-300)

[3] Su L T. Evaluation measures for interactive information retrieval [J]. Information Processing & Management, 1992, 28(4): 503-516

[4] Kelly D. Methods for evaluating interactive information retrieval systems with users [J]. Foundations and Trends in Information Retrieval, 2009, 3(1/2): 1-224

[5] Wang Chao, Liu Yiqun, Zhang Min, et al. Incorporating vertical results into search click models [C] //Proc of the 36th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2013: 503-512

[6] Liu Zeyang, Liu Yiqun, Zhou Ke, et al. Influence of vertical result in Web search examination [C] //Proc of the 38th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2015: 193-202

[7] Ageev M, Guo Qi, Lagun D, et al. Find it if you can: A game for modeling different types of Web search success using interaction data [C] //Proc of the 34th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2011: 345-354

[8] Field H A, Allan J, Jones R. Predicting searcher frustration [C] //Proc of the 33rd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2010: 34-41

[9] Li J, Huffman S, Tokuda A. Good abandonment in mobile and PC Internet search [C] //Proc of the 32nd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2009: 43-50

[10] Huang J, White R W, Dumais S. No clicks, no problem: Using cursor movements to understand and improve search [C] //Proc of the SIGCHI Conf on Human Factors in Computing Systems. New York: ACM, 2011: 1225-1234

[11] Guo Qi, Lagun D, Agichtein, E. Predicting Web search success with fine-grained interaction data [C] //Proc of the 21st ACM Int Conf on Information and Knowledge Management. New York: ACM, 2012: 2050-2054

[12] Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit [J]. Psychological Bulletin, 1968, 70(4): 213

[13] Robertson S E. The probability ranking principle in IR [J]. Journal of Documentation, 1977, 33(4): 294-304

[14] Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques [J]. ACM Trans on Information Systems, 2002, 20(4): 422-446

[15] Chapelle O, Metlzer D, Zhang Y, et al. Expected reciprocal rank for graded relevance [C] //Proc of the 18th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2009: 621-630

[16] Mao Jiaxin, Liu Yiqun, Zhou Ke, et al. When does relevance mean usefulness and user satisfaction in Web search? [C] //Proc of the 39th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2016: 463-472

[17] Friedman J H. Greedy function approximation: A gradient boosting machine [J]. Annals of Statistics, 2001, 29(5): 1189-1232

[18] Chen Ye, Liu Yiqun, Zhou Ke, et al. Does vertical bring more satisfaction? Predicting search satisfaction in a heterogeneous environment [C] //Proc of the 24th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2015: 1581-1590

[19] Lagun D, Ageev M, Guo Qi, et al. Discovering common motifs in cursor movement data for improving Web search [C] //Proc of the 7th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2014: 183-192

[20] Liu Yiqun, Chen Ye, Tang Jinhui, et al. Different users, different opinions: Predicting search satisfaction with mouse movement information [C] //Proc of the 38th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2015: 493-502

[21] He Haibo, Garcia E A. Learning from imbalanced data [J]. IEEE Trans on Knowledge and Data Engineering, 2009, 21(9): 1263-1284

[22] Jiang Jiepu, Hassan A A, Shi X, et al. Understanding and predicting graded search satisfaction [C] //Proc of the 8th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2015: 57-66

Liu Yiqun, born in 1981. PhD. Associate professor. Senior member of ACM and CCF, and council member of CAAI (China Association of Artificial Intelligence) and CIPSC (Chinese Information Processing Society of China). His main research interests include Web search, user behavior analysis, and natural language processing.