山西大学学报(自然科学版)42(1):1－11,2019
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.cnki.shanxi.univ(nat.sci.).2018.11.05.001

图像搜索环境下用户满意度预测方法研究

陈雪松[1],张帆[2],刘奕群[2]*,罗成[2],张敏[2],马少平[2]

(1.青海大学 计算机技术与应用系,青海 西宁 810016;
2.清华大学 计算机科学与技术系,北京信息科学与技术国家研究中心,北京 100084)

摘　要:搜索用户的满意度是搜索引擎性能的重要体现。在网页搜索中,通过用户行为,即用户和搜索引擎之间的交互信息,可以比较准确地预测用户满意度。根据图像和网页搜索过程中的相似和差异,整合和设计出用户交互行为中存在的特征,用来训练模型,并对用户满意度进行预测。实验结果表明,文章所提出的组合模型对图像搜索环境下用户满意度的预测有着较高的准确率。

关键词:图像搜索;用户行为;满意度预测

中图分类号:TP391　　**文献标志码**:A　　**文章编号**:0253－2395(2019)01－0001－11

Prediction of User Satisfaction in Image Search Environment

CHEN Xuesong[1],ZHANG Fan[2],LIU Yiqun[2]*,LUO Cheng[2],ZHANG Min[2],MA Shaoping[2]

(1. *Department of Computer Technology and Application, Qinghai University, Xining* 810016, *China*;
2. *Department of Computer Science and Technology, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing* 100084, *China*)

Abstract: User satisfaction is a prime means to measure search engine performance. User behavior or the interaction data between user and search engine can predict user satisfaction well in traditional web search. Our work concentrate on the utility of the interaction data in image search. According to the similarities and differences between web search and image search, we collect and design features existing in interaction data to train model and predict user satisfaction. The experimental results show that combined model performs well in image search environment.

Key words: image search; user behavior; satisfaction prediction

0　引言

这是一个数据的时代,人们每天都被大量的数据包围着,网络上的数据资源更是不可胜数,如何从海量的数据中尽快地且高质量地寻找出所需数据的需求,催生了信息检索学科[1-2]的发展。信息检索的一个重要载体是搜索引擎,当人们遇到问题的时候,便会通过搜索引擎寻找答案。

目前,广泛应用的信息检索方式是网页搜索,随着日益增长的物质文化需求,用户对网页搜索形式的文本检索有了更高的要求,同时,也希望能搜索到更多类型的信息,比如图像、音乐、视频等等。图像搜索便是

收稿日期:2018-11-05;**接受日期**:2018-11-26

基金项目:国家自然科学基金(61622208;61732008;61532011);国家 973 计划(2015CB358700)

作者简介:陈雪松(1996－),男,山东济南人,研究方向为网络信息检索。E-mail:chenxuesong1128@163.com

* **通信作者**:刘奕群(LIU Yiqun),E-mail:yiqunliu@tsinghua.edu.cn

引文格式:陈雪松,张帆,刘奕群,等.图像搜索环境下用户满意度预测方法研究[J].山西大学学报(自然科学版),2019,**42**(1):1-11. DOI:10.13451/j.cnki.shanxi.univ(nat.sci.).2018.11.05.001

一个应运而生的信息检索场景。在图像搜索环境下,用户有着多样的搜索意图[3-4]。演讲者作汇报展示时,会寻找恰当的图片辅助表达自己的主题;行人走在路边遇到不认识的植物,希望图像搜索能够满足自己的知识需求;办公人员在烦躁的时刻也希望能通过搞笑图片调节自己的情绪。获取用户在不同搜索意图下的满意度是提高搜索引擎性能和竞争力的重要方式。实际搜索环境下,收集每次查询会话后的用户满意度将会是一种花费很大并且难以实施的方法,也会对用户的搜索体验产生负面影响。最近的研究表明,网页搜索场景下,用户在浏览过程中与搜索引擎的交互行为如鼠标的移动、点击,滑轮的滚动等,都是预测满意度的强信号[5-8]。借助于用户和搜索引擎的交互行为对用户满意度进行预测的方法可大致总结为两种:一是从交互行为信息中设计特征来预测满意度[3,9],二是对用户的动作序列进行建模来预测用户满意度[10-11]。

Fig. 1 Difference between web search and image search

图 1 网页搜索和图片搜索区别示意图

从网页搜索到图像搜索,如图 1 所示,整体搜索结果(Search Engine Result Pages,SERPs)的展示形式由一维变成了二维;每个搜索结果的展示内容由标题加摘要变成了缩略图加关键字;同时,翻页的控制方式也由点击按钮变成了滚动滑轮。图像搜索和网页搜索在搜索结果的展示方式,用户与搜索引擎的交互方式等方面的改变,势必影响了用户与搜索引擎的交互行为。在图像搜索环境中,用户与搜索引擎的交互行为的变化决定了用户满意度的预测方法需要重新考量。

本文收集了图像搜索环境下,1 500 多个用户查询会话中的交互行为和用户满意度的反馈,分析了衡量搜索引擎性能的评价指标在图像搜索环境下的表现情况,根据指标表现来设计用户在浏览过程中的动作特征,将其作为梯度提升决策树算法(GBDT)的特征来训练模型从而预测用户满意度,同时将用户在浏览过程中,存在的动作作为马尔可夫模型(Markov Model)的状态,根据动作序列,生成用户在满意和不满意查询下的状态转移概率图,对用户满意度进行预测。最后,本文设计了 GBDT 和马尔可夫模型的组合模型来预测用户满意度,准确率达到了 78%。

在 20 世纪 90 年代,用户满意度首先被 Su 引入到信息检索领域[12],用来表示当用户拥有一个查询需求或者目标的时候,他对于搜索引擎返回结果的满意程度。Jones 等人[13]强调了用户满意度的重要性并且将其作为信息检索评估的基础。用户满意度在信息检索评估中占有了极其重要的地位,因此有了很多相关的工作。Al-Maskari[14]调研了在信息检索中,影响用户满意度的一些因素。Wang 等人[15]论证了用户的满意程度在搜索结果相关性评估和查询建议条目中的重要作用。Hassan 的工作表明[16],用户满意度在衡量搜索引擎性能时的价值比查询和结果的相关性更重要。用户满意度的重要性引导着预测用户满意度工作的开展。他们通过三种不同的方式来衡量点击到访页面(Landing Page)停留时间,进而对单次点击的用户满意度进行预测。也有学者[8]创新性地建立起鼠标的移动轨迹中存在的模式来预测用户满意度。同时,Mehrotra 等人[17]通过用户在查询过程中点击、滚动等动作的次序对用户满意度进行预测。随着移动端搜索流量的增加,很多学者也开始关注移动搜索场景下,用户满意度的预测方法。

本文在选择 GBDT 模型特征时,参考了搜索引擎性能的评价指标[18-20]。搜索引擎性能评价的方式主要包括离线评价方式和在线评价方式。其中离线评价方式考虑了查询文档对的相关性、结果的位置、用户的执

着程度等因素,该评价方法需要外部评估人员进行标注,成本较高,因此本文不采用离线指标作为特征对用户满意度进行预测。对于网页搜索中被广泛使用的在线指标,如点击率、点击结果排名、UCTR、PLC 等,本文通过基于 Concordance 的区分度等指数来衡量这些指标在预测用户满意度中的效用,将效用高的指标选为特征,用来训练模型,预测满意度。同时,本文针对图像搜索场景,提出了一些新的特征。

本文所使用的马尔可夫模型主要考虑了在查询过程中,连续的动作所存在的潜在的关系对用户满意度的影响。Hassan[6]比较了成功的查询和不成功的查询中动作的状态转移概率问题,并通过马尔可夫模型预测用户的查询是否是一次成功的查询。Wu 等人[21]提出了一些预测用户满意度时,可以作为马尔可夫状态的用户动作。本文考虑了用户在查询过程中,不同的动作转移应占有不同权重的问题,并且总结了满意和不满意查询中存在的典型的动作转移模式。

1 实验数据收集

实验采用在校内有偿招募被试者的方式收集数据,被试者依次来到实验室,在指定的机器上进行指定任务的图像搜索,在每次查询会话完成后,被试者需进行满意度打分。

1.1 实验环境

在该实验数据收集过程中,被试者通过 Google Chrome 浏览器,在 17 英寸,分辨率为 1 366×768 像素的 LCD 显示器上进行图像搜索任务。用户在搜索过程中所有的查询内容、鼠标移动、点击、划入(划出)元素、滑轮的滚动、标签的切换等信息都会被记录下来。

1.2 实验过程

本实验在高校招募了 36 名本科生(13 名女生和 22 名男生),年龄分布在 18 到 25 周岁,来自于工科、人文、社会科学和艺术院系等。所有的被试者在实验之前均有图像搜索的经历。

实验开始前,被试者首先要完成一个热身性质的图像搜索任务,从而熟悉整个用户实验。然后被试者按照网页提示的信息依次进行 12 个图像搜索任务。对于每一个任务,被试者会首先看到任务的描述信息,该描述信息用来模拟真实搜索中的用户需求,比如说通过图像搜索引擎找到一张哈利波特的海报用来做 PPT。实验的具体流程如下:

被试者先读任务描述,然后用通俗的语言把任务重述一遍,保证彻底理解了模拟的用户需求。然后,被试者点击“开始任务”按钮进行图像搜索。当被试者认为任务完成或者找不到满意的结果时,便可以点击“结束任务”按钮来结束。任务结束后,被试者在该任务下的每次查询内容将再次展示出来,实验会要求被试者对每次查询进行 5 个等级的用户满意度打分。

2 GBDT 模型

本文通过评测在线指标与用户满意度(5 级标注)的 Pearson 相关系数和 Concordance 一致性指数及基于 Concordance 的区分度指数进行 GDBT 模型特征的选择,利用 sklearn 中 GradientBoostingClassifier 分类器进行训练,采用十折交叉验证的方式进行模型评价。

2.1 特征显著性评测方法

Pearson 相关系数和 Concordance 一致性指数为常用的相关性评价指标,本文不再具体介绍。基于 Concordance 的区分度指数算法设计如下:

算法 1 基于 Concordance 的区分度指数算法

INPUT:在线指标特征向量 *OnlineMetricVector*,
满意度向量 *SatisfactionVector*
OUTPUT:区分度指数 *distinction*
1:function CONCORDANCE(*OnlineMetricVector*,*Satis factionVector*)
2: *PairNum*←0
3: *PostiveCorrectionScore*←0
4: *NegativeCORRECTION Score*←0
5: $avg \leftarrow \overline{OnlineMetricVector}$
6: for $i=0$ $j\rightarrow$ *LengthOfOnlineMetricVector* d0
7: for $j=i\rightarrow$*LengthOfSatisfactionVector* do
8: *PairNum*←*PairNum*+1
9: *product*←(*OnlineMetric*[*i*]− *OnlineMetric*[*j*])/*avg* * (*Satisfaction*[*i*]−*Satusfactuib*[*j*])
10: if *product*>α then
11: *PostiveCorrectionScore*←*PostiveCorrectionScore*+1
12: end if
13: if *product*<α then
14: *NegativeCorrectionScore*←*NegativeCorrectionScore*+1
15: end if
16: if α<=*product*<=α then
17: *PostiveCorrectionScore*←*PostiveCorrectionScore*+1
18: *NegativeCorrectionScore*←*NegativeCorrectionScore*+1
19: end if
20: end for
21: end for
22: *Normalization*(*PostiveCorrectionScore*,*NegativeCorrectionScore*)
23: *ConcordanceValue*=*Max*(*PostiveCorrectionScore*,*NegativeCorrectionScore*)
24: *distinction*=*abs*(*PostiveCorrectionScore*-*NegativeCorrectionScore*)
25: return *distinction*
26:end function

在算法 1 中,在线指标特征向量 *OnlineMetricVector* 的定义为

$$OnlineMetricVector = \begin{bmatrix} OnlineMetric_1^1 & \cdots & OnlineMetric_1^m \\ \cdots & \ddots & \cdots \\ OnlineMetric_n^1 & \cdots & OnlineMetric_n^m \end{bmatrix} \tag{1}$$

其中 $OnlineMetric_n^m$ 表示第 n 次查询中,第 m 个特征的值。满意度向量 *SatisfactionVector* 的定义为

$$SatisfactionVector = \begin{bmatrix} Satisfaction_1 \\ \vdots \\ Satisfaction_n \end{bmatrix} \tag{2}$$

其中 $Satisfaction_n$ 表示第 n 次查询中,用户的满意度情况。

在 Concordance 算法中,由于在线指标特征数值较多,并且存在一定的计算误差(鼠标移动距离,滑轮滚动等都不是绝对精确的数值)对于任意两个对应位置上归一化后的在线指标特征和用户满意度数值相差不大时(即绝对值小于 α),则认为该对位置上的值既支持两个向量呈正相关,又支持两个向量呈负相关。如果该种位置对数量较多,就会导致两个向量的正相关和负相关 Concordance 一致性指数都比较大,但是该在线指标特征并不能很好地体现出用户的满意度,为了解决此类问题,本文提出了基于 Concordance 的区分度指数,其数值大小体现了支持两个向量正相关和负相关的差值。该区分度指数与 Concordance 一致性指数在衡量在线指标是否可以作为 GBDT 模型特征时各有所长,本文在选取作为模型特征的在线指标时综合考虑了 Pearson 相关系数和上述两个指数。

2.2 模型特征

在传统的网页搜索中,常用的在线指标汇总如表 1 所示。

表 1 在线指标及其描述

Table 1 Online metrics and description

在线指标	描述	指标来源
UCTR	一次查询中是否存在点击	点击
QCTR	一次查询过程中点击的次数	点击
MaxRR/MinRR/MeanRR	点击了的所有链接的排序值倒数的最大/最小/平均值	点击
PLC	点击的总次数除以点击位置最低的那次点击在 y 方向上的位置	点击
MaxScroll	滑轮滚动过的最长距离	滑轮滚动
QueryDwellTime	在一次查询会话内所用的时间	查询
SumClickDwell/AvgClickDwell	一次查询下,所有点击后打开图片详情页面后,在图片详情页面停留的总时间/平均时间	点击
TimeToFirstClick	从查询开始到第一次点击的时间长度	点击
TimeToLastClick	从查询开始到最后一次点击的时间长度	点击
DsatClickCount/DsatClickRatio	在一次查询中,不满意点击的次数或比例	点击

本文充分考虑了图像搜索场景下的应用环境,提出了如下特征:

Query id:query id 中的 id 表示在同一查询任务下,当前查询属于该任务中的第几次查询。该特征与用户满意度呈一般负相关,也就是说,在同一任务下,用户的查询次数越多,越容易出现不满意的查询。

LastClickToEnd:用户在当前查询下,最后一次点击的时间点与查询结束的时间点之间的时间长度。该指标与用户满意度呈现较强的负相关性。也就是说,用户在点击图片后,与结束查询的时间越短,满意程度越高。如果用户在查询结束和最后一次点击之间存在较多的动作,比如说 hover(鼠标悬浮),scroll(滑轮滚动)等,就意味着用户还在寻找着更合适的图片,容易感到不满意。现有的结论表示用户的最后一次点击一般是得到了满意的结果来结束查询,用户满意后就会停止查询,用户最后一次点击发生后,结束查询的时间越短,用户满意的可能性就越大,该指标能体现用户的这种行为。

QueryTermNum:用户使用图像搜索引擎时,输入的查询内容不同,对图像搜索引擎返回的结果的期望不同。比如说用户输入"衬衫"和"宽松款女士白色衬衫"时,前者表示用户对搜索引擎有一个宽泛的要求,只要是衬衫即可,后者表示用户对图像搜索引擎返回的属性有了"宽松款"、"女士"、"白色"的要求,期待搜索引擎的返回的结果能够满足所有属性,因此,需要有一个对查询内容复杂度衡量的指标。借助自然语言处理(NLP)中的 jieba 分词工具,对用户提交的所有 query 进行分词处理。去掉了查询内容中的停用词、连词等,对查询内容中剩下的以形容词、名词为主的单词进行加权、计数得到一个简单的用于估计查询内容复杂度的数值。结果表明,该数值与用户满意度存在较弱的负相关,即查询内容中包含单词越多,数值越大,用户满意度越低。也就是说用户提交的查询内容中包含的查询词越多,搜索引擎越难以让用户满意。

Distribution:在查询中,将鼠标悬停时间、鼠标的移动距离和鼠标的移动速度划分为不同的区间,统计不同区间中特征在特征总数的占比,用来作为在线指标。其中鼠标移动速度特征表现较好。对于鼠标移动速度来说,鼠标移动速度非常快([0,0.5]px/ms)的比重越大用户在查询中越容易不满意,鼠标移动速度在[0.5,∞px/ms]的比重越大,用户越容易满意。换句话说,用户鼠标移动速度快的比例大,代表着用户没有在搜索结果页面中检查到满意的结果,是一种不耐烦的表现。

NonMoveTimeRatio:已有研究表明[22],在以文字为主的网页中,鼠标移动和人的注意力有很高的相关性,因此,用户在鼠标移动和鼠标不移动的两个状态下,对搜索结果页面的检查方式不一样。在鼠标不移动的状态下,搜索结果中很有可能有用户感兴趣的内容,用户的注意力集中,注意力切换较慢;在鼠标移动的状态下,用户在查找内容,注意力切换快。在一次查询会话中,用户注意力集中的时间的比例可通过如下公式计算($T_{[start,end]}$ 表示该次查询会话的总时间,$T_{mouse\_move_i}$ 表示第 i 次鼠标移动的时间),将该比例作为指标,表现较好。

$$p_{Non\_Move}=\frac{T_{[start,end]}-\sum_{i=0}^{n}T_{mouse\_move_i}}{T_{[start,end]}} \tag{3}$$

2.3 特征筛选

指标 TTFC,TTLC,LCTE 的应用场景是在查询会话中存在至少一次点击的情况,但在用户的实际搜索中,有些查询会话并不存在点击行为,对于该类会话,假定点击行为发生的时间距离标记时刻(TTFC 和 TTLC 的标记时刻是查询会话开始的时刻,LCTE 的标记时刻是查询会话结束的时刻)无穷远。根据 2.1 节中所提出的特征显著性评测方法,对 2.2 节中所有特征进行筛选,筛选后的特征及其在 3 种评测方法下的表现如表 2 所示。

表 2 GBDT 特征显著性评测结果

Table 2 Significant results of GBDT features

指标	Pearson 相关系数	Concordance 一致性指数/%	区分度指数/%
Query id	−0.294*	−76.18	19.84
UCTR	0.535*	93.58	29.11
QCTR	0.364*	85.04	30.63
MaxRR	0.268*	84.44	30.65
MinRR	0.128*	83.61	29.83
MeanRR	0.223*	83.57	30.95
MaxRRow	0.434*	85.89	31.40
MinRRow	0.297*	83.39	30.28
MeanRRow	0.397*	83.40	31.22
PLC	0.329*	83.54	31.53
QDT	0.271*	75.34	24.52
SCD	0.247*	83.91	32.03
ACD	0.212*	83.59	30.85
TTFC	−0.535*	−93.10	29.19
TTLC	−0.534*	−90.65	28.67
DsatCC	0.328*	84.53	25.47
DsatCR	0.415*	88.25	23.06
LCTE	−0.532*	−90.54	28.59
NonMoveTimeRatio	0.307*	74.63	20.35
distance_0	−0.08	−72.27	13.19
distance_1	0.138*	83.56	15.07
speed_1	−0.258*	−73.47	15.27
speed_2	0.168*	68.55	11.43

Pearson 相关系数列中含有 * 代表 t 检验显著性水平在 0.001。distance_n,speed_n 分别代表鼠标移动距离、速度的分布情况。

2.4 模型效果

本文首先把用户满意度分成两类,用户满意度为 4,5 的查询看作是用户满意的查询,用户满意度为 1,2,3 的查询看作是用户不满意的查询。除本文提出的 5 个特征外,剩余的所有在线指标为特征训练的模型作为 baseline。图 2 展示了本文所设计的特征在预测用户满意度时的表现,所有特征对预测用户满意度均有不同程度的贡献。

将上述所有的在线指标作为特征用于常见分类器的训练,如 k-近邻(KNN)、支持向量机(SVM)、朴素贝叶斯(Naive Bayesian)、GBDT 等模型,同时把 Wu 等人[17]工作中所设计的特征训练生成 GBDT 模型的表现作为 Baseline,所有模型的性能对比如表 3 所示。

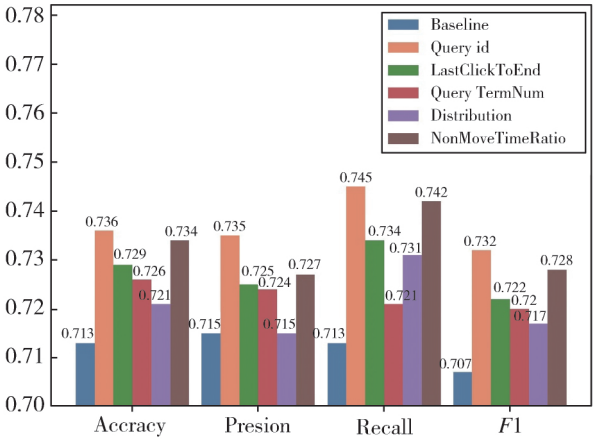

Fig. 2 Performance to predict user satisfaction by our designed features

图 2 本文所设计特征在预测用户满意度时的表现

表 3 模型性能对比图

Table 3 Comparison of performance in different models

	Accuracy	Precision	Recall	*F*1
SVM	0.506	0.259	0.506	0.342
KNN	0.708	0.706	0.701	0.700
Naive Bayesian	0.730	0.692	0.816	0.744
Wu	0.742	0.800	0.823	0.810
Our	0.770(3.77%)	0.817(2.13%)	0.828(0.61%)	0.821(1.36%)

通过上表可以看出,在所有的分类器中,GBDT 模型表现最好。两个 GBDT 模型作为对比,本文所采用的特征训练所得的模型表现较好,精度提高了 3.77%,同时在 Wu 的工作中,用到的特征数量是 33 个,在本文的模型中,用于 GBDT 模型的特征数量有 23 个。

3 马尔可夫模型

3.1 马尔可夫基准模型

在马尔可夫模型中,本文考虑了用户在整个查询会话中,动作转移概率的问题。比如说从查询开始到点击动作,从点击动作到滑轮滚动动作等动作间的转移概率。本文将所有的动作划分为了六类,具体动作及其描述如表 4 所示。

表 4 马尔可夫模型中动作状态及其描述

Table 4 States and descriptions of Markov Model

动作	描述
Start	用户开始查询的标记
Hover	在 SERPs 页面中,用户在一张图片结果上鼠标停留
Click	用户点击了 SERPs 中的图片链接
Down	鼠标滑轮向下滚动
Up	鼠标滑轮向上滚动
End	用户结束了查询

本文首先将数据集划分为训练集和测试集,在训练集中,将其划分为用户满意的数据集部分和用户不满意的数据集部分,为两部分数据集生成两个状态转移矩阵,也就意味着生成了用户满意情况下的状态转移图和用户不满意情况下的状态转移图。用户满意情况下的状态转移图如图 3 所示,用户不满意情况下的状态转移图如图 4 所示。

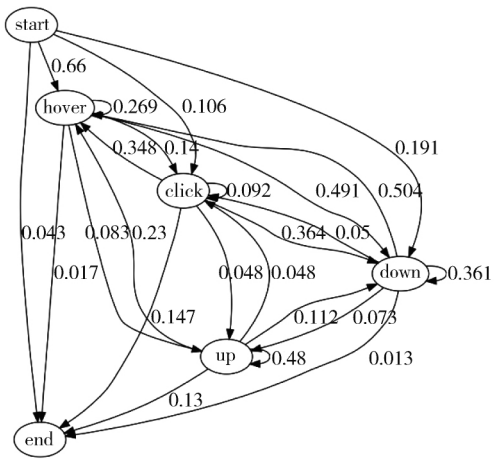

Fig. 3 Transition diagram under satisfaction

图 3 用户满意情况下的状态转移图

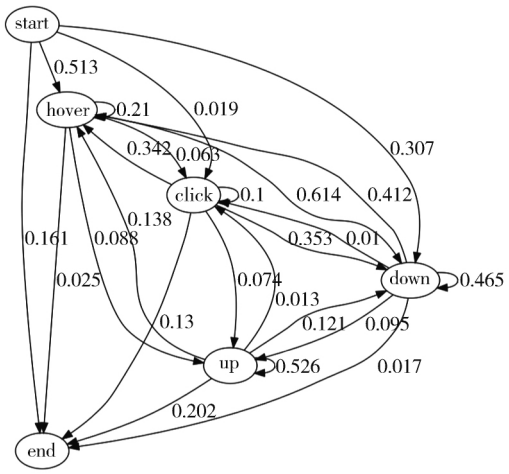

Fig. 4 Transition diagram under dissatisfaction

图 4 用户不满意情况下的状态转移图

根据测试集中一次查询会话中的动作序列来预测用户满意度时,两个状态转移图均可计算当前查询所包含的动作序列的得分。该查询在哪个状态转移图的得分越高,就可以说明该查询中的动作序列更符合其

状态转移图的预测结果。将此马尔可夫模型作为 baseline。

3.2 典型动作模式

对于马尔可夫模型中存在的任意两个动作转移,它支持所在查询是满意或者不满意的程度不同,通过比较用户满意和不满意情况下,状态转移概率的比值,可以将两种情况下的典型动作模式筛选出来。典型动作模式生成算法如下。

算法 2 用户在满意和不满意查询中典型动作模式得分算法

```
INPUT:用户在满意查询中的状态转移矩阵 SatTransitionMatrix,用户在不满意查询中的状态转移矩阵 DSatTransition-
      Mtrix
OUTPUT:满意查询中的动作模式 SatActionPattern,不满意查询中的动作模式 DSatActionPattern
1:function GETACTIONPATTERN(SatTM = SatTransitionMatrix, DSatTM = D SatTransitionMtrixl
2:  states←[start,hover,click,down,up,end]
3:  for OriginaIState=start→end do
4:    for DestinationState=start→end do
5:      SatOvetDSatValue←
      SatTM[OriginalState][DestinationState]/DSatTM[OriginaIState][DestinationState]
6:      DSatOverSatValue←
      DSatTM[OriginaIState][DestinationState]/SatTM[OriginalState][DestinationState]
7:      if Sat0verDSatValue>1+α then
8:        SatActionPattern[OriginalStatel][DestinationState]←SatOverDSatValue
9:      end if
10:     if DSatOvetSatValue>1+α then
11:       DSatActionPattern[OriginalState][DestinationState]←DSatOverDSatValue
12:     end if
13:   end for
14:  end for
15:end function
```

在算法 2 中,输入量所需的用户在满意查询中的转移矩阵和不满意查询中的状态转移矩阵可由对应的状态转移图得到。在用户满意情况下的状态转移图中,$P(\text{pos})_{\text{start}\to\text{end}}=0.043$,在用户不满意情况下的状态转移图中,$P(\text{neg})_{\text{start}\to\text{end}}=0.161$,根据这两个概率,可以计算比值 $Ratio_{\text{pos/neg}}=\dfrac{P(\text{pos})_{\text{start}\to\text{end}}}{P(\text{neg})_{\text{start}\to\text{end}}}$ 和 $Ratio_{\text{neg/pos}}=\dfrac{P(\text{neg})_{\text{start}\to\text{end}}}{P(\text{pos})_{\text{start}\to\text{end}}}$,得到前者比值为 0.267,后者比值为 3.744,后者比值较大,说明该状态转移是用户不满意查询中的典型动作模式,即如果在一次查询中,存在用户开始查询后,没有点击,滑动等动作,直接结束掉查询的动作模式,则说明该次查询极有可能是一次用户不满意的查询。同时,该值越大,说明该状态转移所代表的动作模式越典型。

用户在满意或者不满意情况下,存在的典型动作模式如图 5 和图 6 所示。

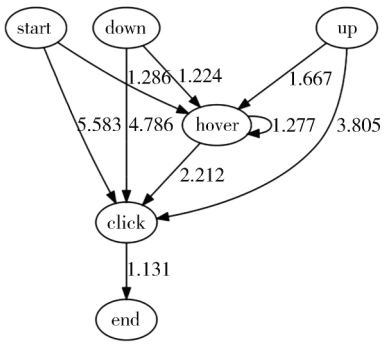

Fig. 5 Typical action pattern under satisfaction search

图 5 用户在满意查询中的典型动作模式

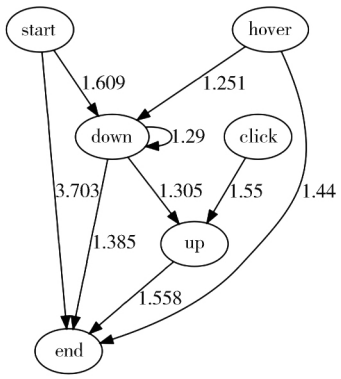

Fig. 6 Typical action pattern under dissatisfaction search

图 6 用户在不满意查询中的典型动作模式

如果用户在一次查询会话中的动作序列为{start, click, end},该序列是用户在满意查询下典型动作模式中所存在的序列,在该典型动作模式中得分较高,故而用户该次查询满意的概率较大。

3.3 加权马尔可夫模型(weighted Markov Model)

通过对查询结束前一个动作内容(包括 start,scroll,hover,jump in, jump out)进行统计分析,得到不同满意度下的查询结束前动作分布图(见图 7),由动作分布图可以看出,用户满意度较低(用户满意度为 1,2,3)的查询中,用户一般以 scroll 结束查询,用户满意度高的查询中,用户会以 jump in(鼠标点击图片链接后返回搜索结果页面)结束查询。用户以 scroll 作为查询结束的最后一个动作,意味着用户在结束当前查询前,仍然用鼠标滚动滑轮,试图寻找着满意的答案,该行为是用户不满意的一个强信号;用户以 jump in 作为查询结束的最后一个动作,意味着用户在搜索结果页面中点击了一个图片链接,用户在检查 landing page(图片详情页面)后,对整个查询是满意的,回到搜索结果页面(jump in)后,就直接结束了当前查询。因此,加权马尔可夫模型提高了这两个动作预测用户满意度时的权重。

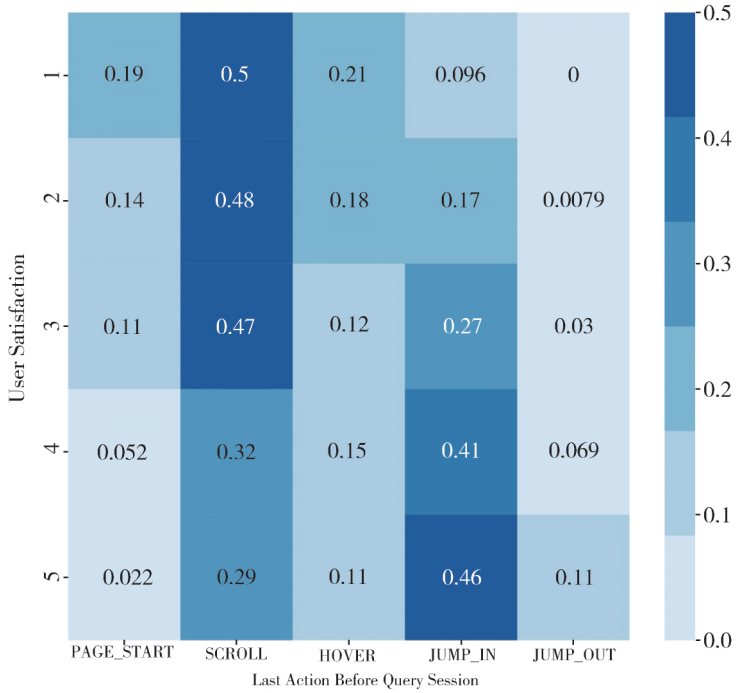

Fig. 7 Distribution of last action before query session under different user satisfaction

图 7 不同满意度下的查询结束前动作分布图

4 混合模型

最后,本文将 GBDT 模型和马尔可夫模型结合对用户满意度进行预测。首先,在马尔可夫模型中抽取出了两类特征,一类是马尔可夫模型的预测输出值,一类是马尔可夫模型中所存在的典型动作模式的得分情况。

(1) 马尔可夫模型预测结果作为特征:马尔可夫模型的输出值可以为 GBDT 模型提供 2 维的特征,一维是用户是否满意(0 或者 1,离散值),一维是用户满意或者不满意的可能性(1 左右的值,比 1 越大,代表模型认为用户满意的可能性越大,比 1 越小,代表模型认为用户不满意的可能性越大,连续值)。

(2) 马尔可夫典型动作模式作为特征:分别计算要预测查询会话中存在的动作转移在用户满意和不满意情况下典型动作模式图的总得分作为 GBDT 模型的特征,对用户满意度进行预测。

本文将抽取的马尔可夫模型的特征添加至 GBDT 模型已有的特征中,生成 GBDT+马尔可夫特征模型(GBDT+Markov's features Model),对用户满意度进行预测。同时,对于同一查询会话的用户满意度预测,GBDT 模型和马尔可夫模型都会有一个预测结果,本文提出的 GBDT 与马尔可夫置信度选择模型(Confidence Selection Model)是将两个模型中置信度高的结果作为对用户的满意度预测的最终结果。

5 模型表现

将 GBDT 模型和马尔可夫基准模型作为 baseline,本文提出的所有拓展模型的表现如表 5 所示(其中,括号中第一个值代表相对于马尔可夫基准模型的表现,第二个值代表相对于作为 baseline 的 GBDT 模型的表现)。

表 5 模型性能表

Table 5 Performance of models

	Accuracy	Precision	Recall	*F*1
Markov (baseline)	0.719	0.868	0.677	0.761
GBDT (baseline)	0.770	0.817	0.828	0.821
Weighted Markov	0.737(2.50%)	0.834(−3.92%)	0.749(10.64%)	0.789(1.97%)
GBDT+Markov's features	0.775	0.826	0.830	0.830
	(7.79%,+0.65%)	(−4.84%,1.10%)	(22.60%,0.24%)	(9.07%,1.10%)
Confidence Selection	0.781	0.828	0.831	0.834
	(8.62%,1.43%)	(−4.61%,1.35%)	(22.75%,0.36%)	(9.59%,1.58%)

在所有模型中,GBDT 与马尔可夫置信度选择模型的预测结果的效果最好,预测的精度达到了 78.1%。整体上来看,GBDT 相关的模型比单纯基于马尔可夫模型的相关模型(马尔可夫基准模型和加权马尔可夫模型)表现要好一些。

6 结论

用户满意度是衡量搜索引擎性能的关键因素。准确地预测用户满意度可以辅助搜索引擎不断改良,从而具有更高的行业竞争力。在传统的网页搜索中,根据用户与搜索引擎交互过程中存在的特征和用户使用搜索引擎时的动作序列能够准确地预测用户的满意度。相较于网页搜索,图像搜索引擎提供了不同的结果展示方式,改变了用户与搜索引擎的交互行为。本文围绕着图像搜索环境下的用户满意度预测方法进行设计、研究。

本文首先提出了基于 Concordance 的区分度指数,用来衡量用户与搜索引擎之间的交互信息中存在的一些特征在预测用户满意度时的效用。其次,针对图像搜索环境下,提出了的新的特征来描述用户与图像搜索引擎间的交互行为,进而预测用户满意度。并设计算法总结出了用户在满意查询和不满意查询中存在的典型动作模式。最后,本文整合了用户的行为特征,动作模式,动作状态转移情况等,设计出的模型在预测用户满意度时的准确率达到了 78%左右。本工作对在线指标的设计,用户满意度的预测等相关领域的研究都有着一定的参考价值。

本文中所使用的数据集包含了约 1 500 次本科生进行图像搜索的信息,是一个较小的数据集,该数据集在被试者的职业、年龄上存在局限性,因此所提出用户满意度预测模型的泛化能力有待评价。由于数据集较小,对于用户在搜索过程中的动作类型区分较少,用户的行为特征提取较为宽泛,也是导致用户满意度的预测精度不是很高的原因之一。在实际的图像搜索环境中,用户可以同时看到多个搜索结果,在对比图像结果后才进行点击查看,图像本身的吸引性等内容特征也对用户的满意度影响较大,因此用户与图像搜索引擎存在着更多、更复杂的交互行为有待研究。同时,如何更好地解释用于预测用户满意度的特征的含义也是今后的研究方向之一。

参考文献:

[1] Schütze H, Manning C D, Raghavan P. Introduction to Information Retrieval[M]. Cambridge University Press, 2008. DOI: 10.1017/CBO9780511809071.

[2] Larson R R. Introduction to Information Retrieval[J]. *Journal of the American Society for Information Science and Technology*, 2010, **61**(4): 852-853. DOI: 10.1002/asi.v61:4.

[3] Xie X, Liu Y, Wang X, *et al*. Investigating Examination Behavior of Image Search Users[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2017: 275-284. DOI: 10.1145/3077136.3080799.

[4] Wang X, Wen J R, Dou Z, *et al*. Search Result Diversity Evaluation based on Intent Hierarchies[J]. *IEEE Transactions on Knowledge & Data Engineering*, 2018, **30**(1): 156-169. DOI: 10.1109/TKDE.2017.2729559.

[5] Guo Q, Lagun D, Agichtein E. Predicting Web Search Success with Fine-grained Interaction Data[C]// Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012: 2050-2054. DOI: 10.1145/2396761.2398570.

[6] 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. ACM, 2010: 221-230. DOI: 10.1145/1718487.1718515.

[7] Lagun D, Hsieh C H, Webster D, *et al*. Towards Better Measurement of Attention and Satisfaction in Mobile Search[C]// Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 2014: 113-122. DOI: 10.1145/2600428.2609631.

[8] Liu Y, Chen Y, Tang J, *et al*. Different Users, Different Opinions: Predicting Search Satisfaction with Mouse Movement Information[C]// Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2015: 493-502. DOI: 10.1145/2766462.2767721.

[9] Dan O, Davison B D. Measuring and Predicting Search Engine Users' Satisfaction[J]. *ACM Computing Surveys (CSUR)*, 2016, **49**(1): 1-35. DOI: 10.1145/2893486.

[10] Ageev M, Guo Q, Lagun D, *et al*. Find It If You Can: a Game for Modeling Different Types of Web Search Success Using Interaction Data[C]// Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011: 345-354. DOI: 10.1145/2009916.2009965.

[11] Hassan A. A Semi-supervised Approach to Modeling Web Search Satisfaction[C]// Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012: 275-284. DOI: 10.1145/2348283.2348323.

[12] Su L T. Evaluation Measures for Interactive Information Retrieval[J]. *Information Processing & Management*, 1992, **28**(4): 503-516. DOI: 10.1016/0306-4573(92)90007-M.

[13] Jones K S. Information Retrieval Experiment[M]. Butterworth-Heinemann, 1981. ISBN: 0408106484.

[14] Al-Maskari A, Sanderson M. A Review of Factors Influencing User Satisfaction in Information Retrieval[J]. *Journal of the Association for Information Science and Technology*, 2010, **61**(5): 859-868. DOI: 10.1002/asi.21300.

[15] Wang H, Song Y, Chang M W, *et al*. Modeling Action-level Satisfaction for Search Task Satisfaction Prediction[C]// Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 2014: 123-132. DOI: 10.1145/2600428.2609607.

[16] Kim Y, Hassan A, White R W, et al. Comparing Client and Server Dwell Time Estimates for Click-level Satisfaction Prediction[C]// Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 2014: 895-898. DOI: 10.1145/2600428.2609468.

[17] Mehrotra R, Zitouni I, Hassan Awadallah A, et al. User Interaction Sequences for Search Satisfaction Prediction[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2017: 165-174. DOI: 10.1145/3077136.3080833.

[18] Hofmann K, Li L, Radlinski F. Online Evaluation for Information Retrieval[J]. *Foundations and Trends® in Information Retrieval*, 2016, **10**(1): 1-117. DOI: 10.1561/1500000051.

[19] Wu P, Hoi S C H, Zhao P, *et al*. Online Multi-modal Distance Metric Learning with Application to Image Retrieval[J]. *Ieee transactions on knowledge and data engineering*, 2016, **28**(2): 454-467. DOI: 10.1109/TKDE.2015.2477296.

[20] Isinkaye F O, Folajimi Y O, Ojokoh B A. Recommendation Systems: Principles, Methods and Evaluation[J]. *Egyptian Informatics Journal*, 2015, **16**(3): 261-273. DOI: 10.1109/ICISC.2017.8068649.

[21] Wu Z, Liu Y, Zhang M, *et al*. Understanding and Predicting User Satisfaction in Image Search[C]// WSDM'18 Workshop on Learning from User Interactions (Learn-IR'18), February 2018, Los Angeles, California, USA.

[22] Chen M C, Anderson J R, Sohn M H. What Can a Mouse Cursor Tell Us More?: Correlation of Eye/Mouse Movements on Web Browsing[C]// CHI'01 extended abstracts on Human factors in computing systems. ACM, 2001: 281-282. DOI: 10.1145/634067.634234.