

网络用户行为分析法和建模法研究综述*

吴清强

(厦门大学软件学院, 厦门 361000)

摘要: 随着当下互联网的发展与网络使用人数的增多, 用户的行为操作数据更加容易获取和利用。本文主要对网络用户行为研究中的用户行为分析法和建模法进行理论研究和实践现状的综述。通过对用户行为数据进行不同方式的分析和不同方向的建模, 并对行为建模进一步分析, 可以发掘用户行为模式和联系, 为下一步互联网的发展提供依据和指导。

关键词: 用户行为分析; 数据挖掘; 用户行为模型

中图分类号: TP391

DOI: 10.3772/j.issn.1673-2286.2015.11.006

网络用户行为是指用户与互联网环境及服务交互产生的一系列活动^[1]。网络用户行为分析则是指在获得网络访问基本数据的情况下, 对有关数据进行统计、分析, 从中发现用户访问网站的规律, 并将这些规律与网络资源整合、网络营销策略等相结合, 从而发现目前网络资源整合和营销活动中可能存在的问题, 并为进一步修正或重新制定包括数字图书馆在内的网络资源和营销策略提供依据^[2]。通过分析网络用户行为监测数据, 可以让互联网服务的提供者更加详细、清楚地了解用户的行为习惯, 更有针对性地解决相关问题, 提高用户体验, 更好满足用户需求, 最终提升互联网服务提供者的效益。鉴于此, 研究和实践领域近年都产生了较多的研究成果, 本文对这些研究成果, 尤其是其中的用户行为分析方法、用户行为模型构建方法的相关内容进行系统梳理, 有助于更好把握网络用户行为分析的研究和实践现状, 为下一步进行网络用户可靠性行为识别打下坚实的基础。

1 网络用户行为分析方法

在利用各种用户行为记录工具获取用户行为数据之后, 首先要选择合适的方法对这些数据进行分析。现

有的分析方法主要包括统计分析法、聚类分析法、关联分析法、决策树法、神经网络法以及时序数据挖掘法等, 下面分别进行阐述。

1.1 统计分析法

统计分析法是最基本的行为分析方法, 主要对用户行为分解、归类后进行数量统计, 得出某个类型行为的数据总量, 并借助于这些数据总量分析用户行为的规律^[3]。国内外近年都有一些研究采用了统计分析方法, 如Ye Wu等^[4]和Jiefei Yu等^[5]以统计分析方法研究了论坛(天涯和新浪)上用户行为的差异, 得出用户帖子的问答分布遵循某一幂率分布, 通过函数的转化和拟合实现对网民的一些回复及观点的定量分析, 并对比了中国网民与西方网民的异同; 程鹏^[6]通过收集用户的点击流和访问日志等数据, 利用数据统计分析技术进行用户行为分析; 尤婷^[7]和邓夏玮^[8]同样也以统计分析的方法对社交网站中用户行为的特征进行了研究。

1.2 聚类分析法

与分类统计分析不同的是, 聚类分析是一种探索

* 本研究得到国家社科基金项目“面向检索的网络用户行为可靠性度量研究”(编号: 13CTQ011)资助。

性的分析,在分类聚类过程中,并没有事先设定分类标准,而是基于样本数据自动进行分类。鉴于用户行为数据的复杂性,聚类分析比分类统计法具有更广的使用范围,也是用户行为分析中最为常用的分析方法。Sang Hyun Oh等^[9]通过用户行为建模,提出了一种应用于用户异常行为监测的聚类分析方法,重点讨论了用于聚类的特征值选择问题,实验证明该聚类分析比统计分析更加精确地描述用户行为。Marcelo Maia等^[10]通过抓取大量的Youtube用户数据,采用聚类算法聚集相同行为的用户,从而更好地刻画不同类别的用户行为。孙燕花^[11]则基于网络Netflow流信息进行了IP为源和目的统计建模,改进了快速聚类K-Medoids算法,设计了一种异常反选择标准,即密度评估标准,对聚类后的异常行为进行识别。刘鹏^[12]提出了基于熵对数据分组和基于数据特点一次合并多个数据样本的快速层次聚类算法,对用户业务偏好变化随时间变化的规律进行研究,并利用非齐次泊松过程对网络用户上下线行为进行建模分析。蔡岳等^[13]提出一种基于用户行为聚类的搜索引擎算法,从用户行为日志中挖掘用户意图,并根据用户的反馈信息定位用户意图信息。陈敏等^[14]结合Web用户浏览行为的特点,引入了粗糙度概念,提出了一种新的路径相似度计算方法,在计算相似度时不仅把用户的浏览模式作为一种序列模式来考虑,还充分考虑了用户在网上浏览的时间因素,以实现Web用户浏览行为的聚类。延皓^[15]基于流量监测提出了一种基于熵的访问状态转移矩阵(CWSTM)聚类算法,用于用户会话行为统计分析、上网时段喜好分析、Web喜好分析和用户Web访问状态转移研究。张霞等^[16]提出了一种基于用户查询意图的聚类分析算法,通过检索用户关键词,将数据信息采用树状的形式存储,从而获取和树相邻接的上下信息,通过关键信息来分析用户的查询意图。李磊等^[17]将微博用户分为一般关注、主动参与和信息传播三种类型,提出了一种基于微博主题的用户聚类算法,用于网络舆情的检测。张万山等^[18]在Web资源个性化推荐中,跟踪用户浏览和检索Web资源的行为数据,提出了基于主题聚类的个性化推荐算法,实现了Web资源的动态推荐。

1.3 关联分析法

在用户行为分析中,常将用户行为习惯和其他行为习惯借助于Apriori等关联规则算法进行关联分析,以

发现不同行为习惯之间的关系和规律,从而达到对用户行为预测的目的。王爱平等^[19]从宽度优先、深度优先、数据集划分、采样、增量式更新、约束关联、多层多维关联等方面对数据挖掘中常用的关联规则挖掘算法进行详细系统的介绍。王永利^[20]在Web用户行为挖掘过程中发现了关联规则经典算法Apriori的不足,并针对这些不足,提出了一种基于Web挖掘的分类关联规则挖掘算法。一些研究成果应用于信息安全检测中的用户行为分析,如:潘蕾等^[21]针对网络监听系统需要解决的提取用户访问模式信息中的多维多值关联规则问题,对传统的关联规则算法进行了扩充和改进,提取有效的关联规则,反映用户的行为模式。戴臻等^[22]针对入侵检测系统中快速获取用户使用模式的需求,在Apriori算法基础上提出了一种基于特定模式树的用户行为关联规则挖掘算法,通过递归挖掘模式树获得最大频繁集。周云霞等^[23]在数据库入侵检测系统用户行为挖掘中改进了FP-Growth算法,提高挖掘效率。罗强^[24]利用Apriori算法提出了社交网络用户行为关联分析算法以划分虚拟社团,提高针对虚拟社团用户信息推送的针对性。数字图书馆领域近年也有研究成果^[25-26]将关联分析应用于图书馆用户借阅行为分析,从而为用户提供更加个性化的图书推荐。

1.4 决策树法

决策树利用信息和数据的树状图形向决策人提供后续问题决策辅助。在网络用户信息行为分析过程中,因为涉及到用户后续信息行为的预测,决策树也被较多地运用在网络用户信息行为分析中。徐孝娟等^[27]利用Ethnographic决策树分析了科学网博客用户的行为,并利用问卷调查验证预测效果。贺露^[28]利用决策树法并结合线性回归法分析预测社交网用户的性格特征。李贤鹏等^[29]对分类预测中广泛使用的ID3决策树算法进行分析,指出了该算法的取值偏向性以及运算效率不高等优点,在此基础上提出了一种改进的ID3算法并将其应用于某移动通信公司的客户流失预测。曾雪等^[30]引入代价敏感学习理论,将不同的错分代价纳入建模过程,以构建一个基于代价敏感的决策树的电信客户离网分析模型,有效地提高对流失客户的预测性能。邹竞等^[31]阐明了决策树算法是电信行业提高客户忠诚度、防止客户流失的重要手段,介绍了决策树算法应用于电信行业客户流失分析中的方法、步骤及具体实现过程。

1.5 神经网络

神经网络,也称连接模型,是一种模仿动物神经网络行为特征,进行分布式并行信息处理的算法数据模型。神经网络在网络用户信息行为分析中主要用于预测用户接下来的信息行为,动态调整相应策略,提供更加个性化的服务。刘蓉等^[32]提出手工定制和系统自动建模相结合的自适应神经网络建模方法,可动态调整神经网络的参数,修正用户模型,使神经网络的输出能随用户的兴趣而改变。左琳^[33]则将稳定转换法应用于神经网络的混沌控制中,基于改进的LPCA神经网络方法,提出了一种新的网络用户行为分析模型,揭示了用户区域属性对使用模式和应用习惯的影响。段隆振等^[34]根据Kohonen自组织特征映射神经网络中学习阶段的性质,运用双Kohonen神经网络组合成新的自组织训练挖掘模型,改善了聚类效果,为Web用户访问模式挖掘提供了一种可行的方法。李宇华^[35]将神经网络应用于网络安全入侵行为监测方面,将神经网络、模糊系统和减法聚类三者融合,开发出可有效监测攻击行为的模糊神经网络用户行为分析系统。

1.6 时序数据挖掘

时序数据是指与时间相关或含有时间信息的数据或用数字或符号表示的时间序列,由于相关数据随着时间连续变化,能够反映出某个待观察过程在一定时间内的状态或表现^[36]。鉴于网络用户信息行为有着明显的时序特征,采用这种方法可以学习网络用户过去的时间特征,并能够预测用户未来的行为。Ron Hutchins^[37]从一个美国拨号上网服务提供商RADIUS服务器上采集了2000年5月至9月共60多万不同用户登录信息的数据,首先以1分钟为间隔对节日、周末和工作日登录次数变化情况进行描述,然后又对用户不同日期的在线时长分布、用户的地理分布以及不同地区平均登录时间进行了分析,此外还对用户数目的变化做出了估计模型。Hutchins^[38]采用与Ron Hutchins相同的数据源,以5分钟为间隔对几个月的登录次数变化情况进行描述,以及对几个月的在线时长分布分别作了对比分析,还研究了单个用户的登录间隔时间分布。数字图书馆领域近年也有一些研究成果^[39-40]将时序数据挖掘方法应用在用户借阅行为分析方面,跟踪用户需求的调整 and 变化。

2 网络用户行为模型构建方法

模型是对真实世界中问题域内的事物的描述。网络用户行为分析大多需要对用户行为进行建模,即用图示、文字、符号等组成的流程图等形式对用户行为及其规律进行描述,是用户行为分析的重要基础。纵观网络用户行为模型的构建方法,主要有两大类:一类是图形化构建表征方法,一类是语言描述类表征方法。下面分别对相关情况进行阐述。

2.1 图形化网络行为模型构建方法

目前用于网络行为分析的模型主要有3个:(1)客户行为模型图(Customer Behavior Model Graph, CBMG)^[41]。客户行为模型图是一个基于转换概率矩阵的图形,主要用于描述网络用户如何从一个状态转化到另一个状态。客户行为模型图主要由代表状态名的椭圆和代表转换的箭头符号组成,转换箭头上标明概率,用于表征从用户进入网站或系统开始到离开网站或系统结束时的一系列状态和改变,可以直接转化为概率矩阵进一步计算^[42-43]。(2)客户访问模型(Customer Visit Model, CVM)^[44]。客户访问模型将多个会话(如检索、浏览、下载、复制等)表示成一个会话向量集合,每一个会话为该向量集合中的一个向量,用于表征用户在网站或系统上的所有行为^[43-45]。(3)客户/服务器交互图(Client/Server Interaction Diagrams, CSID)^[46]。主要用于描述用户与服务器之间为实现某一功能而可能产生的系列交互。交互图主要由代表用户的方形、代表服务器的椭圆以及交互箭头共同组成。通过客户/服务器交互图可以将用户与服务器之间所有交互动作表征出来为下一步的计算、规律分析提供直观依据^[42-43]。

2.2 语言描述类表征方法

网络用户行为的模型构建也可以利用规范的语言进行描述,其中较为典型的的就是GOMS (Goals, Operators, Methods and Selection rules)模型, GOMS即目标、操作、方法和选择规则。模型采用分而治之的思想,将一个任务进行多层次的细化,通过目标、操作、方法和选择规则四个元素来描述人机交互行为,以便进行网络用户行为特征和规律的分析^[47]。目前,包括GOMSED、Qgoms、CogTool在内的很多工具都已经支

持GOMS模型,也衍生出了很多改进的模型,如KLM (Keystroke Level Modeling)^[48]、NGOMSL(Natural GOMS Language)^[49]和CPM-GOMS(Cognitive Perceptual Motor GOMS)^[50]等。另一种语言类表征方法是时序关系说明语言(Language of Temporal Ordering Specification, LOTOS)^[51],该语言用一套形式化和严格的表示法刻画系统外部可见行为之间的时序关系,可保证描述不存在二义性,便于分析和一致性测试理论的研究。用户行为标注(User Action Notion, UAN)^[52]则是另外一种使用较多的,用于描述用户的行为序列以及在执行任务时所用界面的简单符号语言,主要着眼于用户和界面两个实体之间的交互。

3 总结

网络用户行为蕴含着大量有意义的信息和有价值的研究方向。搜集用户行为数据,进行数据统计分析、用户行为建模、分析行为模型,对挖掘网络用户行为有着重要意义。根据不同的目的,采用不同的方式,将会发掘出不同的用户行为习惯。本文通过梳理用户行为分析方法以及用户行为建模方法,对网络用户行为分析研究进行归纳。未来的用户行为分析将主要沿着这些方向展开以及更新改进。

参考文献

- [1] 何苾菲.基于微博用户行为的兴趣模型构建和可视化方法研究[D].哈尔滨工业大学,2013.
- [2] 秦超.电子商务网站访客行为分析系统[D].上海:上海交通大学,2006.
- [3] 网站用户分析的五大常规方法[EB/OL]. [2014-12-03]. <http://www.admin5.com/article/20100726/254679.shtml>.
- [4] Wu, Y., Zhou C., Chen M., et al. Human comment dynamics in on-line social systems [J]. Physica A: Statistical Mechanics and its Applications, 2010, 389(24): 5832-5837.
- [5] Yu, J., Hu Y., Yu M., et al. Analyzing netizens' view and reply behaviors on the forum [J]. Physica A: Statistical Mechanics and its Applications, 2010, 389(16): 3267-3273.
- [6] 程鹏.中大型互联网网站用户行为分析数据平台[D]. 上海:复旦大学,2012.
- [7] 尤婷.社交网站用户行为特征及其内在机制研究[D]. 北京:北京邮电大学,2012.
- [8] 邓夏玮.基于社交网络的用户行为研究[D]. 北京:北京交通大学,2012.
- [9] Oh S.H., Lee W.S., An anomaly intrusion detection method by clustering normal user behavior[J]. Computers & Security, 2003. 22(7): 596-612.
- [10] Maia, M., J. Almeida, V.I. Almeida. Identifying user behavior in online social networks[C]// Association for Computing Machinery. Proceedings of the 1st workshop on Social network systems. New York:2008.
- [11] 孙燕花.基于聚类的网络用户行为分析[D].长沙:中南大学,2011.
- [12] 刘鹏.网络用户行为分析的若干问题研究[D].北京:北京邮电大学,2010.
- [13] 蔡岳,袁津生.用户行为聚类的搜索引擎算法与实现[J].计算机系统应用,2010,19(4):94-97.
- [14] 陈敏,苗夺谦,段其国.基于用户浏览行为聚类Web用户[J].计算机科学, 2008,35(3):186-187.
- [15] 延皓.基于流量监测的网络用户行为分析[D].北京:北京邮电大学,2011.
- [16] 张霞,马一宁,陈静汝.一种基于用户查询意图的聚类分析算法[J]. 电脑知识与技术,2012,8(14).
- [17] 李磊,刘继.面向舆情主题的微博用户行为聚类实证分析[J].情报杂志,2014(3):118-121.
- [18] 张万山,肖瑶,梁俊杰,等.基于主题聚类的Web资源个性化推荐研究[J].微电子学与计算机, 2015(4):35-39.
- [19] 王爱平,王占凤,陶嗣干,等.数据挖掘中常用关联规则挖掘算法[J].计算机技术与发展,2010(4):105-108.
- [20] 王永利.关联规则挖掘算法及其在Web挖掘上应用的研究[D].哈尔滨:哈尔滨工程大学,2013.
- [21] 潘蕾,苏晶,徐汀荣.网络访问行为关联规则提取的研究与设计[J].计算机应用与软件, 2008,25(3): 189-191.
- [22] 戴臻,费洪晓,李俊,等.基于特定模式树的用户行为关联规则挖掘算法[J].计算机系统应用,2007,5(5): 56-59
- [23] 周云霞,栗磊.基于数据库用户行为分析的改进FP-Growth算法[J].科学技术与工程,2011(18):4380-4383.
- [24] 罗强.社交网络用户行为关联分析的关键技术研究[D].成都:电子科技大学,2013.
- [25] 陈定权,朱维凤.关联规则与图书馆书目推荐[J].情报理论与实践,2009(6):81-84.
- [26] 李欣.基于关联规则的图书馆图书智能查询系统研究[D].长春:东北师范大学,2012.
- [27] 徐孝娟,赵翔翔,朱庆华.民族志决策树方法在学术博客用户行为中的研究[J].现代图书情报技术, 2014(1):79-86.
- [28] 贺露.基于社交网络的用户性格与行为分析[D].北京:北京邮电大学,2014.
- [29] 李贤鹏,何松华,赵孝敏,等.改进的ID3算法在客户流失预测中的应用[J].计算机工程与应用,2009, 45(10): 242-244.
- [30] 曾雪,胡建华,王清心.基于代价敏感的决策树的电信离网分析模型[J].计算机与现代化,2009(4): 62-64.
- [31] 邹竟,谢鲲. C4.5算法在移动通信行业客户流失分析中的应用[J].计

- 算技术与自动化,2009, 28(3): 98-101.
- [32] 刘蓉,陈鹏,张兴艳.个性化网页推荐中基于神经网络的自适应用户模型研究[J].电子测量技术, 2007, 30(4):165-168.
- [33] 左琳.神经网络及在网络用户行为分析中的应用研究[D].成都:电子科技大学,2011.
- [34] 段隆振,朱敏,王靓明.基于双Kohonen神经网络的Web用户访问模式挖掘算法[J].计算机工程与科学, 2009(9):95-98.
- [35] 李宇华.基于模糊神经网络的用户行为分析[D].武汉:华中科技大学,2013.
- [36] 黄书剑.时序数据上的数据挖掘[J].软件学报,2004(1):1-8.
- [37] Hutchins Ron, Zegura Ellen W., Kolesnikov Oleg, et al. Usage characteristics of dial-in Internet users: a national study[J/OL].[2015-02-21]. http://wenku.baidu.com/link?url=mxHKst01J9UkKxuzAhEGXS-IQM4n_TpBh7ZcfetgswcdlETqL8AIL_8WP99Fm3AVWXgEQDaOeN7_DNEFgOY4T993jqTnHE3cU2z7KLQm4G.
- [38] Hutchins, R., Zegura EW., Liashenko A, et al. Internet user access via dial-up networks-traffic characterization and statistics[C]//IEEE. Ninth International Conference. 2001:11-14.
- [39] 付沙.基于序列模式挖掘的图书馆用户借阅行为分析[J].情报理论与实践,2014(6):103-106.
- [40] 陈春颖,熊拥军.基于序列模式挖掘的读者借阅行为分析[J].图书情报知识,2011(2):92-96.
- [41] Xu Xiaolin, Jin Hai, Wu Song, et al., URMG: Enhanced CBMG-Based Method for Automatically Testing Web Applications in the Cloud [J]. Tsinghua Science and Technology, 2014,19(1):65-75.
- [42] User Behavior Modeling: CBMG Creation[EB/OL]. [2015-02-24]. <http://www.isot.ece.uvic.ca/projects/Performance-Analyzer/manual/bm-cbm.html>.
- [43] 余肖生,马费成.网络用户行为模型的构建方法研究[J].情报科学, 2011, 29(4): 605-608.
- [44] Menasce, D., Abraho, B., Barbará, D., et al. Fractal Characterization of Web Workloads[J/OL]. [2015-02-21]. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.3087&rep=rep1&type=pdf>.
- [45] Wang Xiajun, Huang Song, Fu Song. Characterizing Workload of Web Applications on Virtualized Servers[M]// Big Data Benchmarks, Performance Optimization, and Emerging Hardware. Springer International Publishing, 2014:98-108.
- [46] Peiffer, C., I. L'heureux. System and method for maintaining statefulness during client-server interactions:US, 8,346,848[P/OL][2015-02-21]. <http://www.freepatentsonline.com/8346848.html>.
- [47] Stanton, N.A., Salmon P.M, Walker G.H.. Human factors methods: a practical guide for engineering and design[M].Farnham: Ashgate Publishing, Ltd,2013.
- [48] Karousos, N., et al. Effortless tool-based evaluation of web form filling tasks using keystroke level model and fitts law[C]. CHI'13 Extended Abstracts on Human Factors in Computing Systems. New York: ACM,2013.
- [49] Ritter, F.E., Gordon D., Baxter, et al. Foundations for Designing User-Centered Systems[M]// Methodology I: Task Analysis., Springer, 2014:309-333.
- [50] Erazo, O., Pino, J.A., Pino R., et al. Magic Mirror for Neurorehabilitation of People with Upper Limb Dysfunction Using Kinect[J]. Hawaii International Conference on System Sciences,2014:2607-2615.
- [51] Sebrechts, M Marc. The psychology of human-computer interaction(Book)[J]. American Scientist, 1983,71(6):656.
- [52] Hartson, H.R., Gray, P.D.. Temporal aspects of tasks in the User Action Notation[J]. Human-Computer Interaction, 1992,7(1):1-45.

作者简介

吴清强, 男, 博士, 厦门大学软件学院副教授, E-mail: 553401552@qq.com。

Overview of Network User Behavior Analysis and Modeling

WU QingQiang

(Software School of Xiamen University, Xiamen 361000, China)

Abstract: With the development of the Internet and number increase of Internet users, the access and utilization of the user's behavior of operating data becomes easier. This paper mainly summarizes the research and practice of user behavior analysis and user modeling in the behavior analysis of network users. Through the modeling analysis of different ways and in different directions of user behavior data, the behavior modeling further analysis so as to explore user behavior patterns and associations, to provide basis and direction for the next step of the Internet development.

Keywords: User Behavior Analysis; Data Mining; User Behavior Model

(收稿日期: 2015-10-13)