期刊菜单

基于会话聚类和马尔科夫链的动态用户行为模型改进研究

Improvement Research of Dynamic User Behavior Model Based on Session Clustering and Markov Chain

DOI: 10.12677/ECL.2019.81003, PDF, HTML, XML, 下载: 928 浏览: 2,060

科研立项经费支持

作者: 陈梅梅*, 茅金波: 东华大学旭日工商管理学院, 上海

关键词: 电子商务; 行为模型; 会话聚类; 马尔科夫链; E-Commerce; Behavioral Model; Session Cluster; Markov Chain

投稿

相关文章

- 基于隐马尔可夫模…
- 基于马尔科夫和…
- 基于加权马尔可夫…
- 基于灰色分数阶马…
- 基于隐马尔科夫模…

为你推荐

Contact us

摘要: 根据点击流建立用户行为模型并对用户特征进行分析，是企业制定精准的营销策略及提供个性化推荐的基础。首先，本文在传统的动态用户行为模型(CBMG)基础上提出了一种同时考虑页面类型和行为序列的改进的用户行为模型，以从用户路径偏好信息中充分反映其行为模式。其次，基于行为序列和页面类型对用户会话进行聚类，得到的不同行为模式的会话类别，针对不同会话类型基于马尔科夫链得到用户行为状态转移的动态模型。研究发现：基于改进的动态用户行为模型得到的不同类型用户的状态转移模式存在显著差别，且具有更高的可解释性。

Abstract: Establishing user behavior models based on clickstream data and analyzing user characteristics are the basis for companies to develop accurate marketing strategies and provide personalized recommendations. Firstly, based on the traditional dynamic user behavior model (CBMG), this paper proposes an improved user behavior model that considers both page type sequences and behavior sequences to fully reflect its behavior patterns from user path preference information. Secondly, the user session is clustered based on the behavior sequence and the page type, and the session categories of different behavior patterns are obtained. Based on the Markov chain, the dynamic model of user behavior state transition is obtained among different conversation types. The research shows that the state transition patterns of different types of users based on the improved dynamic user behavior model are significantly different and have higher interpretability.

文章引用： 陈梅梅, 茅金波. 基于会话聚类和马尔科夫链的动态用户行为模型改进研究[J]. 电子商务评论, 2019, 8(1): 14-21. https://doi.org/10.12677/ECL.2019.81003

1. 引言

截至2017年12月，中国网络购物用户规模已达5.33亿，较2016年底增加了6662万人，年增长率为14.27% [1] 。购物模式的转变使各类在线购物网站积累的数据越来越多。为了满足用户的个性化需求，基于点击流建立用户模型，制定精准的营销策略和个性化推荐策略是企业的必然选择。

用户模型分为兴趣模型和行为模型，两者的作用不同，兴趣模型体现了用户对内容的偏好，在构建推荐系统时利用兴趣模型可以决定向用户推荐的商品类型，行为模型体现了用户的行为信息的浏览特征，在推荐系统中决定了向用户推荐商品的方式 [2] ，本文侧重于对用户行为模型的研究。

用户行为模型的研究通常分为静态行为模型和动态行为模型。静态行为模型需要明确以哪些指标来反映用户的行为特征，而动态行为模型是通过计算静态行为模型中确定的指标来明确用户所处的状态以及各状态间的转移关系。

由于从点击流中难以直接提取用户行为信息，因此传统基于点击流的用户行为模型研究通过页面之间跳转的规律 [3] 或用户访问页面路径的规律 [4] 反映用户的行为特征，但这种方式对于行为信息的挖掘还不够充分，用户行为特征仅能得到部分解释。袁兴福等针对该问题，提出了一种根据点击流访问页面序列到用户行为的映射方案 [2] ，但其忽略了传统研究中页面类型对用户行为的细化解释作用，研究结果同样受到页面类型和行为类型标记体系的影响，存在一定的不合理性。基于上述研究，本文将在传统的动态用户行为模型基础上提出了一种同时考虑页面类型和行为序列的改进的用户行为模型，以从行为路径偏好信息中充分反映其行为模式。

2. 基于页面类型和行为序列的静态用户行为模型改进

静态行为模型的改进是优化动态行为模型的基础，因此，本文首先以真实的点击流数据为基础，提出一种基于页面类型和行为序列的静态改进用户行为模型。

2.1. 数据预处理

数据的预处理是建立静态行为模型的基础，主要包含数据描述、用户会话的切分、页面类型和行为类型的识别。

2.1.1. 数据描述

本文使用国内领先的某生鲜电商网站提供的数据集进行建模和分析，该数据集包含2017年9月至2018年1月300,000条用户的浏览日志数据，涉及到用户53,163位用户。该数据类似于点击流，包含的重要字段如表1所示：

字段名称	字段含义
User_id	用户ID
Time	创建时间
Action_page	页面名称
Action_code	行为名称
Commodity_name	查看的商品名称
Createtime	离开时间
OS	设备来源

Table 1. Meanings of important features for browser log

表1. 浏览日志重要字段及含义

排除因商品名称变更而导致的重复记录，得到299,233条日志记录。

2.1.2. 用户会话切分

会话是指在一次访问期间，用户从进入网站开始到离开网站所进行的一系列活动 [5] ，记录了用户在某段时间内的行为信息，能反映的行为特征更丰富。

研究中通常基于访问时间或者访问内容对会话的进行划分，由于电商网站的内容会根据用户进行个性化呈现，获取用户访问网页的原始内容难度很大，因此本文采用时间阈值来区分会话。主要借助表1中的“Time”字段对会话进行切分，尝试利用不同的时间阈值对会话分割时，发现不论是以分钟或是以小时为单位，所分割的会话平均长度很短且较为分散，将时间阈值放宽至天时，每个会话的记录数在1~166之间。鉴于会话建模分析的基本要求，本文以1天为阈值对用户会话进行切分，同时排除日志记录少于5条的无效会话以保证分析的有效性，最终筛选得到13,772个会话，涉及138,858条浏览记录。

2.2. 页面类型与行为类型的识别

识别页面类型和行为类型是建立静态行为模型的基础，在根据“Action_page”识别页面类型时，笔者从内容维度、结构维度、功能维度3个层面进行分析，对高频出现的页面以人工浏览的方式进行标记，最终得到的页面类型为：首页、商品详情页、商品列表页、个人资料页面、购物车页面、会员活动页、交易页面、未知页面，各页面类型的占比结果如图1所示。

在识别行为类型体系时，在页面类型的辅助解释下，根据“Action_code”字段进行标记，最终得到的行为类型为：浏览行为、检索行为、用户登录行为、个人资料管理行为、购物车管理行为、推荐选择行为、加购行为、结算行为、交易成功行为、系统跳转加载行为、未知行为，各行为类型的占比结果如图2所示。

2.3. 静态用户行为模型的改进

数学上，序列是被排成一列的对象(或事件)，每个元素之间的顺序非常重要，运用序列反映用户

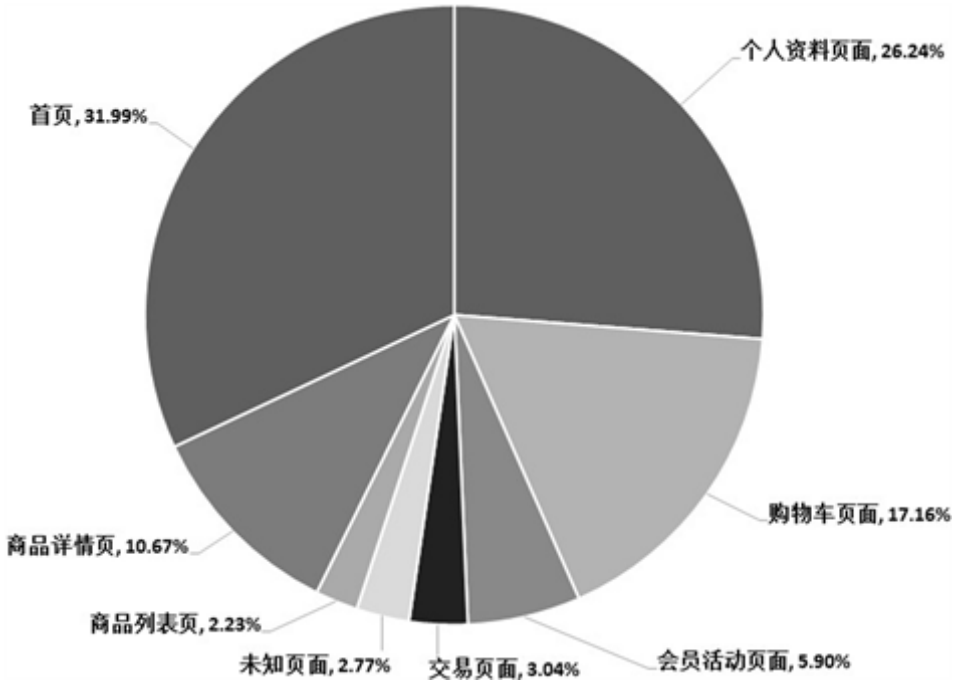

Figure 1. Page type distribution map

图1. 页面类型分布图

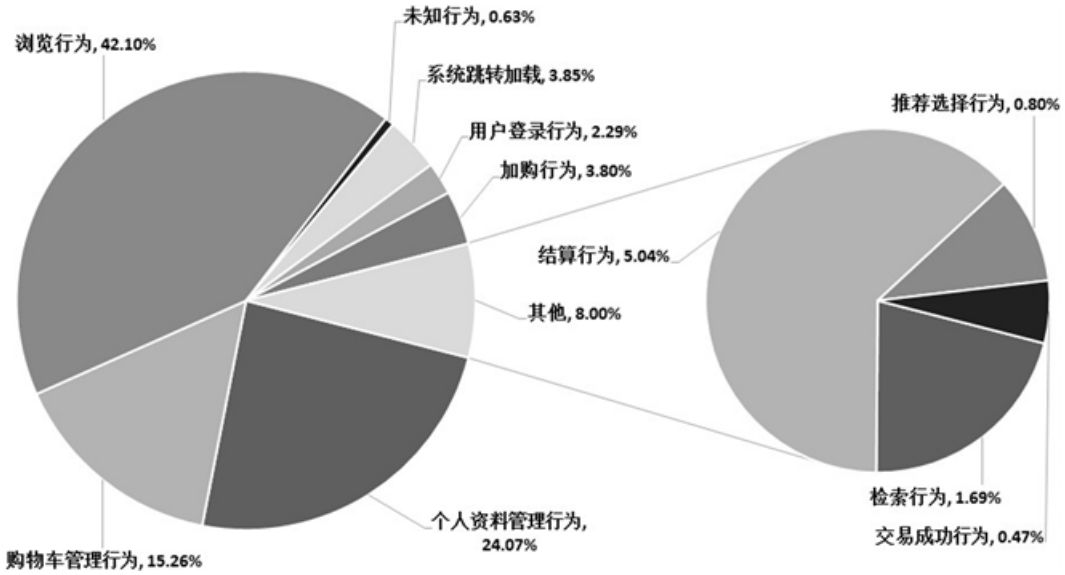

Figure 2. Behavior type distribution map

图2. 行为类型分布图

动态的行为特征是用户行为建模中的常用方法 [6] 。马尔科夫链是描述用户行为序列的经典方法，如朱志国在对用户浏览兴趣进行挖掘时，利用马尔科夫链建立用户的兴趣导航模型 [7] ；Montgomery A L等通过马尔科夫链对浏览路径的序列进行定义，预测用户在网上书店的行为变化 [8] 。因此本文在2.2节识别的页面类型和行为类型的基础上，利用马尔科夫链对会话中的页面序列和行为序列进行刻画。同时，考虑到单独考虑页面序列或者行为序列存在的对行为信息挖掘不充分的问题，本文对页面类型和行为类型进行组合，从两者结合的角度出发对用户的行为路径给予解释，如“在何种页面产生了何种点击行为”，充分反映用户的行为特征。

静态模型中包含的三种序列如图3所示，以椭圆表示各类序列的不同状态，其中 P_i 表示页面， A_i 表示行为， P_iA_i 表示“页面–行为”的组合，如 P_1A_1 表示用户“在 P_1 页面进行了 A_1 点击行为”。

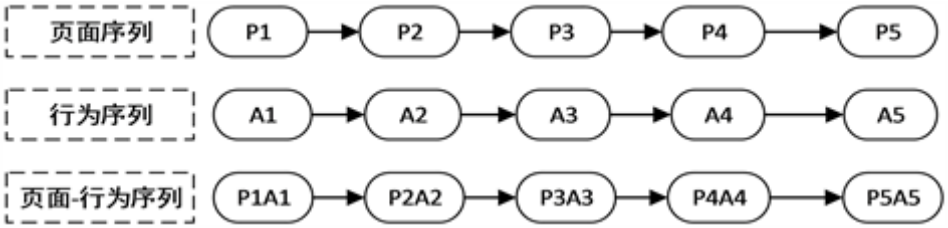

Figure 3. Diagram of session sequence

图3. 会话序列示意图

频繁项挖掘对关联变量的提取能为决策提供支持，在各个领域得到了广泛应用，如购物篮数据分析、网页预取、交叉购物、个性化网站等 [9] 。本文在序列特征的分析中引入频繁项的思想，提取会话中的频繁序列以在分析中充分反映用户的行为模式。

3. 基于会话聚类和马尔科夫链的动态用户行为模型

3.1. 基于行为模型的用户会话聚类分析

聚类分析法是网络数据挖掘的关键技术之一，在用户细分研究中得到了广泛的应用 [10] 。由于传统的动态行为模型并不对用户进行区分，无法给出精准营销的建议。因此，本文试图基于会话聚类的结果，得到不同类型会话的动态行为模式。

本文从页面类型和行为序列两类指标进行聚类分析，其中页面类型指标是指单个会话内用户浏览各页面类型的数量，采用普适性最高的K-means聚类算法进行聚类分析，从结果的可解释性、组内距离、类间距离来评价聚类效果，其中组内距离、类间距离的判断指标为“组内平方”和“Calinski-Harabaz Index”，前者越小聚类效果越好，后者则相反。不同聚类个数的判断系数结果如表2所示。

类别数量	组内平方(SE)	Calinski-Harabaz Index

2	890,571.91	3269.72
3	703,357.36	3429.47
4	626,145.39	3703.37
5	566,718.18	4189.85
6	517,192.62	3723.90

Table 2. Coefficient of cluster evaluation index

表2. 聚类评价指标系数表

由表2可以初步确定划分类别为5类，笔者进一步从聚类结果的可解释和分布实例的合理性最终确定分类个数为5，汇总的分类特征如表3所示。

聚类得到了以下5类会话：营销推动型会话、交易导向型会话、商品浏览型会话、购物车管理型会话、个人资料管理型会话，并通过典型行为和频繁行为序列等信息增强分类会话的可解释性。

3.2. 基于马尔科夫链的转移概率

考虑到客户行为模型图(CBMG)在捕捉用户浏览行为模式、揭示用户行为的有效性，本文利用CBMG对不同类别会话的动态行为模式进行描述。一个CBMG通常由状态和转换构成，一个状态使用带有状态名的椭圆表示，一个转换可以用有概率的箭头符号表示 [11] 。本文中的状态是指各行为类型，转换是指各行为类型之间的概率转换。Montgomery A L等认为用户在电商网站的浏览、点击等行为仅受到当前页面功能的影响，和先前浏览的页面并无直接关系，近似满足马尔科夫的无后效性 [8] 。因此本文将利用马尔

会话特征	分类1	分类2	分类3	分类4	分类5
	营销推动型	交易导向	商品浏览型	购物车管	个人资料管

		型		理型	理型
会话初始 行为比例	浏览行为, 67.8%	管理个人资料, 64.2%	浏览行为, 54.0%	购物车管理行为, 37.1%	管理个人资料， 56.4%
典型行为	浏览行为	个人资料管理行为	浏览行为	购物车管理行为	个人资料管理行为
频繁行为序列	会员活动页浏览→ 会员活动页浏览	个人资料管理→ 首页浏览	首页浏览→ 首页浏览	购物车管理→ 购物车管理	个人资料管理→ 个人资料管理
	会员活动页浏览→ 商品详情页浏览	个人资料管理→ 结算行为	首页浏览→ 商品详情页浏览	购物车管理→ 首页浏览	首页浏览→ 个人资料管理
会话平均记录数(条)	6	19	27	9	6

Table 3. Feature summary of classification

表3. 分类特征汇总表

科夫链对不同会话类型中各行为类型之间的相互转移概率进行计算。本文借助SQL2008进行会话的分类以形成基础数据，再通过python完成转移概率的计算以用于后续对动态行为模式的描述。

3.3. 基于会话聚类的动态行为模型分析

通过会话聚类得到了5种不同浏览行为模式的会话类型，代表着5种不同类型的用户，本节将3.2节计算所得的动态客户行为模型图进行整体会话和分类会话的对比，探讨整

体用户的浏览行为模式是否与区分用户类别所得的结果存在显著差别。考虑到文章篇幅，仅以营销推动型会话为例与整体进行对比分析。

图4和图5分别为整体会话的客户行为模型图和营销推动型会话的客户行为模型图，反映用户动态的浏览行为模式。可以发现，两者的共同之处为无论整体还是营销推动型的用户，在网购过程中均以“浏览行为”、“购物车管理行为”和“个人资料管理行为”为中心，其他行为类型以一定的概率向这几种行为转化，这种共通性是由电商网站本身的架构导致的从具体两个行为的转移概率大小来看，两者之间存在显著的差异，这也表明整体用户的浏览行为模式是一种平均化的结果，不同的类型的用户浏览行为模式具有各自的特征，这也证明了分类探讨用户行为模式的必要性和重要性。下面以图4和图5为例，探讨整体与分类之间的差别。

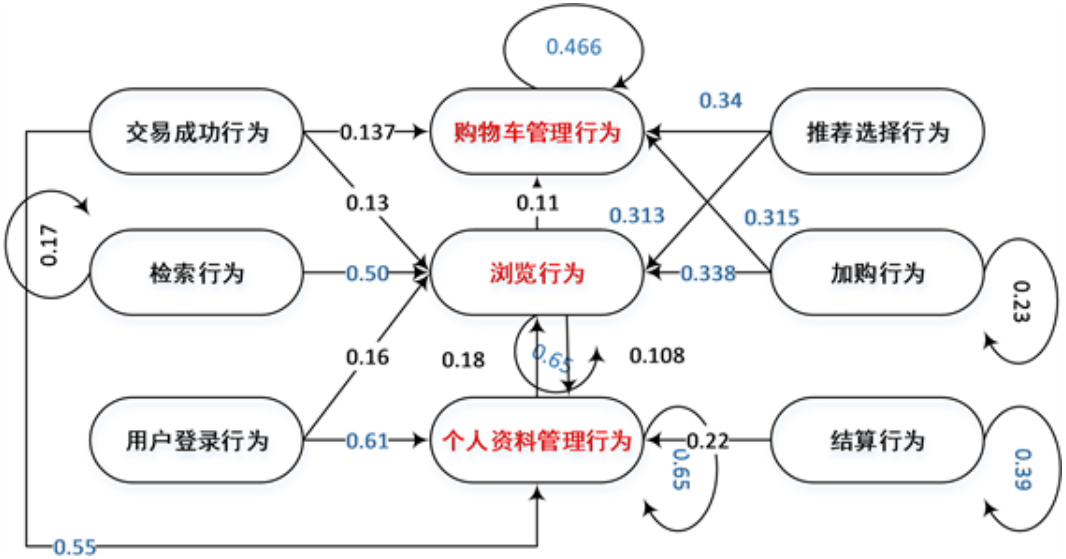

Figure 4. CBMG for the whole session

图4. 整体会话的CBMG

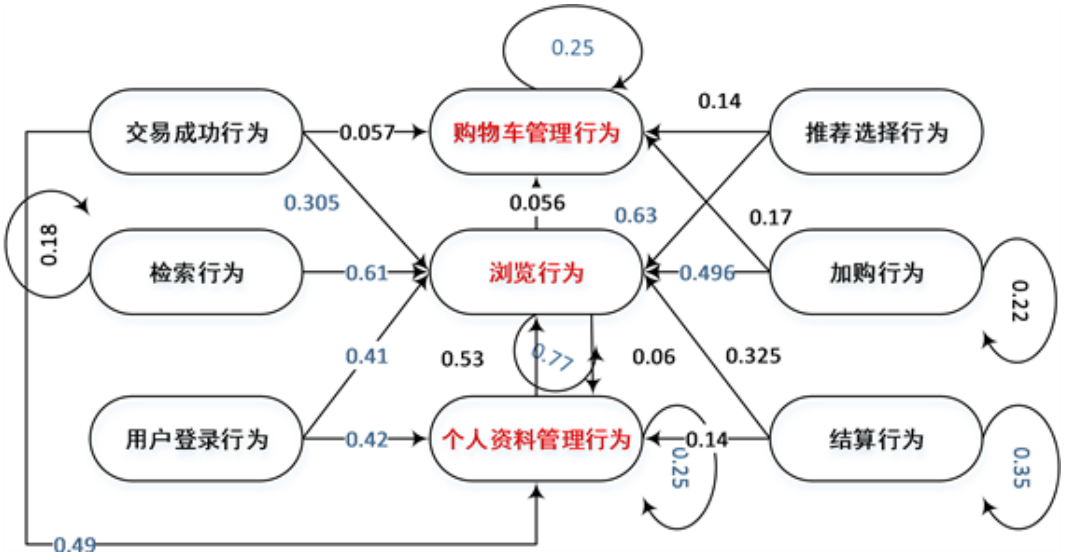

Figure 5. CBMG for the marketing-driven session

图5. 营销推动型会话CBMG

图4和图5对比分析来看，可以发现共同的典型行为路径为“用户登录行为–浏览行为”、“浏览行为–浏览行为”，表明登录后浏览以及反复进行商品浏览是用户在网购时的共同特征。但两者之间也存在较大的差别，如营销推动型用户的典型路径还包含“推荐选择行为–浏览行为”、“加购行为–浏览行为”，相对于整体而言，浏览行为是其典型而突出的行为特征，并且这类用户参考个性化推荐的概率较高，针对该类用户可以为企业提供如下的营销策略：在制定精准营销推荐时，可以将该类用户作为首批受众，通过观察该类用户的行为路径，进一步明确可能的营销增长点；通过分析该类用户的检索浏览历史，在搜索框推荐主题词；考虑在个人中心页面如“历史订单”等页面进行活动推荐以增加用户浏览和购买的概率。综上，从分类的角度探讨用户的浏览行为模式具有更高的现实意义，为企业的精准营销和推荐提供参考。

4. 结论与管理启示

本文在分析点击流参数的基础上，提出了一种同时考虑页面类型和行为序列的改进用户行为模型，并通过聚类分析得到了5种不同类别的会话，分别为营销推动型会话、交易导向型会话、商品浏览型会话、购物车管理型会话和个人资料管理型会话，反映出的会话特征具有较高的可解释性。基于会话聚类结果进行动态用户行为模型的分析，发现整体用户的浏览行为模式与单个分类用户的动态行为模式存在明显的差异，因此从点击流出发分析用户的浏览行为模式对电商企业尤为重要。企业不仅需要把控整体用户的特征，更要利用不同用户群的特征制定差异化的营销策略，例如针对营销推动型的用户，可以利用该用户群对个性化推荐、促销活动的高敏感性，将其作为活动的首批受众测试活动的有效性；而针对交易导向型的用户，由于该类用户购物偏好较为明确，企业需要根据其浏览行为记录在适当的时机做出高精确度的推荐。本文仅以某电商平台的数据为例展开分析，结论的普适性有待进一步检验，希望本文提出的改进模型及分析思路能为电商企业的精准营销和推荐策略制定提供参考，以期提高用户对营销和推荐活动的接受度和整体满意度，进而提高电商平台的购买转化率。

基金项目

国家社会科学基金资助项目(10BGL027)，东华大学人文社会科学预研究重大项目。

参考文献

[1] 中国互联网络信息中心. 第41次中国互联网络发展状况统计报告[EB/OL]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/P020180305409870339136.pdf, 2018-03-05.
[2] 袁兴福, 张鹏翼, 刘洪莲, 等. 基于点击流的电商用户会话建模[J]. 图书情报工作, 2015, 59(1): 119-126.
[3] 张波, 巫莉莉, 周敏. 基于Web使用挖掘的用户行为分析[J]. 计算机科学, 2006, 33(8): 213-214.
[4] 马晓艳, 唐雁. 一种基于用户浏览路径的Web用户聚类方法[J]. 西南师范大学学报(自然科学版), 2009, 34(3): 93-97.
[5] 朱志国. 基于URL语义分析的Web用户会话识别方法[J]. 大连理工大学学报, 2011, 51(3): 440-446.
[6] 管恩政, 常晓宇, 王喆, 等. 快速频繁序列模式挖掘算法[J]. 吉林大学学报: 理学版, 2005, 43(6):

768-772.
[7] 朱志国. 基于隐马尔可夫链模型的电子商务用户兴趣导航模式发现[J]. 中国管理科学, 2014, 22(4): 67-73.
[8] Montgomery, A.L., Li, S., Srinivasan, K., et al. (2004) Modeling Online Browsing and Path Analysis Using Clickstream Data. Marketing Science, 23, 79-595. https://doi.org/10.1287/mksc.1040.0073
[9] Guo, X.Z. and Sun, Y.G. (2009) Research of Intrusion Detection Based on Neural Network Optimized and Genetic Algorithm. Computer Knowledge & Technolo-gy.
[10]张文君, 王军, 徐山川. 电商用户需求状态的聚类分析——以淘宝网女装为例[J]. 现代图书情报技术, 2015, 31(3): 67-74.
[11]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(4): 605-608.