

中图分类号:

UDC:

学校代码: 10055

密级: 公开

南开大学  
硕士学位论文

基于互联网日志的用户行为分析算法和用户模型的研究

Behavior analysis and user modeling based on Internet access  
logs

论文作者 孙卓豪 指导教师 章辉

申请学位 学士 培养单位

学科专业 电子信息科学与技术 研究方向

答辩委员会主席  评阅人

南开大学研究生院

二〇一六年五月

## 南开大学学位论文使用授权书

根据《南开大学关于研究生学位论文收藏和利用管理办法》，我校的博士、硕士学位获得者均须向南开大学提交本人的学位论文纸质本及相应电子版。

本人完全了解南开大学有关研究生学位论文收藏和利用的管理规定。南开大学拥有在《著作权法》规定范围内的学位论文使用权，即：(1) 学位获得者必须按规定提交学位论文(包括纸质印刷本及电子版)，学校可以采用影印、缩印或其他复制手段保存研究生学位论文，并编入《南开大学博硕士学位论文全文数据库》；(2) 为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆等场所提供校内师生阅读，在校园网上提供论文目录检索、文摘以及论文全文浏览、下载等免费信息服务；(3) 根据教育部有关规定，南开大学向教育部指定单位提交公开的学位论文；(4) 学位论文作者授权学校向中国科技信息研究所和中国学术期刊(光盘) 电子出版社提交规定范围的学位论文及其电子版并收入相应学位论文数据库，通过其相关网站对外进行信息服务。同时本人保留在其他媒体发表论文的权利。

非公开学位论文，保密期限内不向外提交和提供服务，解密后提交和服务同公开论文。

论文电子版提交至校图书馆网站：<http://202.113.20.161:8001/index.htm>。

本人承诺：本人的学位论文是在南开大学学习期间创作完成的作品，并已通过论文答辩；提交的学位论文电子版与纸质本论文的内容一致，如因不同造成不良后果由本人自负。

本人同意遵守上述规定。本授权书签署一式两份，由研究生院和图书馆留存。

作者暨授权人签字：\_\_\_\_\_

20     年     月     日

### 南开大学研究生学位论文作者信息

论 文 题 目	基于互联网日志的用户行为分析算法和用户模型的研究				
姓 名	孙卓豪	学号	1210403	答辩日期	
论 文 类 别	博士 <input type="checkbox"/> 学历硕士 <input type="checkbox"/> 硕士专业学位 <input type="checkbox"/> 高校教师 <input type="checkbox"/> 同等学力硕士 <input type="checkbox"/>				
院 / 系 / 所			专 业		
联 系 电 话			Email		
通讯地址 (邮编):					
备注:					

注:本授权书适用我校授予的所有博士、硕士的学位论文。由作者填写(一式两份)签字后交校图书馆，非公开学位论文须附《南开大学研究生申请非公开学位论文审批表》。

# 南开大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下进行研究工作所取得的研究成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律 responsibility 由本人承担。

学位论文作者签名：\_\_\_\_\_ 年 月 日

## 非公开学位论文标注说明

根据南开大学有关规定，非公开学位论文须经指导教师同意、作者本人申请和相关部门批准方能标注。未经批准的均为公开学位论文，公开学位论文本说明为空白。

论文题目			
申请密级	<input type="checkbox"/> 限制 (≤2 年) <input type="checkbox"/> 秘密 (≤10 年) <input type="checkbox"/> 机密 (≤20 年)		
保密期限	20 年 月 日至 20 年 月 日		
审批表编号		批准日期	20 年 月 日

南开大学学位办公室盖章 (有效)

限制 ☐ 2 年（最长 2 年，可少于 2 年）

秘密 ☐ 10 年（最长 5 年，可少于 5 年）

机密 ☐ 20 年（最长 10 年，可少于 10 年）

## 中文摘要

随着互联网的高速发展，用户留下越来越多的网络行为信息。基于这些用户信息，分析用户行为和研究用户模型，研发适合海量数据的自然语义分析、文本挖掘、位置分析等算法对预测用户行为及商品推荐至关重要。本文针对大数据精准营销的难点，开展数据挖掘技术、模型与算法、场景化实践的实证研究，构建面向新媒体业务的场景化智能营销平台，并针对新媒体公司自有海量数据的特征与需求进行系统优化，提高资源利用率及推荐效果，从而全面提升公司的用户运营和内容运营能力。

**关键词：** 海量数据处理；自然语义分析；数据挖掘

## **Abstract**

Today Internet users, typically 4G mobile users, leave billions tons of data on Servers. Base on such big data, it's urgent to develop algorithm about natural semantic analysis, text mining to predict users' behavior and recommand related content. This paper focus on big data analysis, develop some data mining algorithm and build a bunch of system to pop up Internet company's profit.

**Key Words:** Big data; natural semantic analysis; data mining

## 目录

## 第一章 背景及研究现状

### 第一节 背景

随着互联网的普及，人们每天花费越来越多时间通过网络工作，娱乐。目前，由于芯片技术极速发展，全球大概 30 亿人拥有计算能力相当于 1980 年房间般大小的超级计算机的智能手机。这些移动手机用户使用各地运营商大力推广的 4G 上网服务，实现随时随地上网的梦想。除了电脑和手机，人们还将可穿戴设备及各类物联网接入互联网中。由此，互联网各网络节点每天都能采集到大量用户的上网信息。基于这些用户信息，分析用户行为和研究用户模型，我们能挖掘出用户的兴趣爱好，由此推荐相关的内容，从而提高互联网的交互能力以及网络运营商内容运营能力。

### 第二节 本文结构

本文第一章主要介绍背景。第二章主要介绍相关研究及系统。第三章主要介绍网络结构，爬虫以及网站分类，记录数据格式化，清洗，保存。第四章主要介绍基于 Hadoop 的推荐算法。第五章主要介绍系统工作流程，以及实证分析。第六章是总结以及文中提出的问题。

## 第二章 相关研究及系统

### 第一节 研究现状

基于日志挖掘的网站分类目录用户心智模型研究基于爬虫和网站分类的主题信息源发现方法找一些关于关于推荐算法的，分类的，语义分析的论文以及发展历程

### 第二节 相关系统介绍



## 第三章 用户上网数据采集与处理

本章主要介绍用户上网如何留下信息，如何将上网信息转变为推荐系统所需要的训练集。其中包括了以下内容：网络流量信息采集、打开 URL 后发生了什么、如何给原数据库中没有的 url 分类、如何格式化、存储

### 第一节 网络信息传递流程

互联网用户通过各类软件发送、接受信息。对于不同的软件，有不同的信息发送、接受格式, 称作协议。以大多数网络浏览器使用的基本 HTTP 协议来说, 其请求的头部有以下常用的信息

表 3.1 NKThesis 预调用的宏包

头部名称	描述
Accept	Content-Types that are acceptable for the response
Content-Type	The MIME type of the body of the request (used with POST and PUT requests)
Date	The date and time that the message was sent
Host	The domain name of the server (for virtual hosting), and the TCP port number on which the server is listening. The port number may be omitted if the port is the standard port for the service requested
Referer	This is the address of the previous web page from which a link to the currently requested page was followed
User-Agent	The user agent string of the user agent

根据以上 HTTP 请求头部我们可以获得用户请求网站 URL，请求时间，使用浏览器类型信息。当然对于纪录了所有网络流量包的情况，我们甚至可以解压信息包，获得网页信息。对于使用其他协议的软件来说，只要按照该协议的约定，也能正确解压信息包，获取其中内容。

## 第二节 网络流量信息采集

从“基于 Hadoop 的网络流量分析系统的研究与应用.pdf”的 [17]-[22] 介绍了一堆在“基于流量监测的网络用户行为分析.pdf”的 P22 页有插图基于网络流量监测的移动互联网特征研究.pdfP28 页有介绍

## 第四章 基于大数据处理的推荐算法

本章首先简单描述了经典推荐算法，然后将其推广至包含超出用户兴趣矩阵之外信息的推荐算法，最后将其应用到 Hadoop 上推荐算法。

## 第五章 系统工作流程与实证分析

### 第一节 系统整合

基于前几章的描述，我们可以综合构建一个数据挖掘系统，发现用户兴趣，推荐相关内容。



### 第二节 真实数据测试

抱歉暂时没申请到机器以及数据

## 第六章 总结及未来展望

本文通过描述几个算法，最后构建了一个系统。本文所介绍的系统中存在以下问题未解决：

## 第七章 The Tikz Package

The pdf package, where “pdf” is supposed to mean “portable graphics format” (or “pretty, good, functional” if you prefer...), is a package for creating graphics in an “inline” manner. It defines a number of T<sub>E</sub>X commands that draw graphics. For example, the code `\tikz \draw (0pt,0pt) -- (20pt,6pt);` yields the line  and the code `\tikz \fill[orange] (1ex,1ex) circle (1ex);` yields .

In a sense, when you use pdf you “program” your graphics, just as you “program” your document when you use T<sub>E</sub>X. You get all the advantages of the “T<sub>E</sub>X-approach to typesetting” for your graphics: quick creation of simple graphics, precise positioning, the use of macros, often superior typography. You also inherit all the disadvantages: steep learning curve, no wysiwyg, small changes require a long recompilation time, and the code does not really “show” how things will look like.

## 致谢

感谢您使用本模板。

## 个人简历