



开放科学
(资源服务)
标识码
(OSID)

融合大数据分析的用户画像构建

蔡皎洁^{1,2}

1. 湖北工程学院 孝感 432000;
2. 湖北小微企业发展研究中心 孝感 432000

摘要: [目的/意义] 本文从融合大数据分析的角度,深入用户偏好形成的内部机理,构建包含因果关系标签的用户画像,提高其在应用中的预测能力。[方法/过程] 通过大数据分析获取用户偏好的关联关系,为量表式小数据分析提供理论假设素材并获取用户偏好的因果关系,通过语义集成形成完整的用户偏好标签体系。[结果/结论] 以互联网股票投资领域进行数据实验,本文所提出的理论框架和方法能够深入用户偏好形成的心理,提高用户画像分类预测的能力。

关键词: 大数据分析; 小数据分析; 相关关系; 因果关系; 用户画像

中图分类号: G35; TP391

User Profile Construction Based on Big and Small Data Analysis

CAI Jiaojie^{1,2}

1. Hubei Engineering University, Xiaogan 432000, China;
2. Hubei Small and Micro Enterprise Development Research Center, Xiaogan 432000, China

Abstract: [Purpose/Significance] From the perspective of fusion of big and small data analysis, this paper goes deep into the internal mechanism of user preference formation, constructs user portraits containing causal relationship labels, and improves its predictive ability in applications. [Method/process] The correlation relationship of user preferences is obtained through big data analysis, which provides theoretical hypothesis materials for scale-based small data analysis and obtains the causal relationship of user preferences, and forms a complete user preference label system through semantic integration. [Result/conclusion] Through the data experiment in the field of Internet stock investment, the theoretical framework and method proposed in this paper can go deep into the psychology of the formation of user preferences and improve the ability of user portrait classification and prediction.

Keywords: Big Data Analysis; small data analysis; correlation; causality; user profile

基金项目 湖北省科技厅软科学项目“数字普惠金融助力精准扶贫对策研究”(2019ADC148); 教育厅规划办: 省社科基金前期资助项目“数字普惠金融助力精准扶贫模型研究”(19ZD057)。

作者简介 蔡皎洁(1982-), 博士、副教授, 研究方向为网络金融, 数据挖掘, E-mail: caijiaojie0011@aliyun.com.cn。

引用格式 蔡皎洁. 融合大数据分析的用户画像构建[J]. 情报工程, 2022, 8(1): 100-110.

引言

早期用户画像模型一般是基于设计师的主观想象进行构建,而设计师主观想象的直接表现就是通过问卷设计来了解用户喜好。随着移动互联网及社交网络的迅速发展,用户喜好表现于散布网络平台上的行为数据中,不同于以往问卷、访谈等形式所“制造”的数据,网络数据呈现出相对客观中立、海量多结构化、实时变化等特征,这为传统定量及定性分析带来了技术、模型及研究范式等多方面的挑战。随着机器学习、文本挖掘、自然语言处理等多种新型数据分析技术的兴起,开启了基于大数据分析的用户画像研究模式。大量研究表明,大数据分析的优势在于挖掘事物间相关关系,以往基于问卷等小数据分析优势在于论证事物间的因果关联,但两者并非取代关系,而是互补和增进关系。在用户画像构建中,如何将两种不同的数据分析思路、方法与研究范式融合在一起,提高对用户偏好分析的深度与宽度,是本文研究的目标。从大小数据分析的特征及与用户画像构建的语义关联出发,构建基于大小数据分析的用户画像研究框架,并以互联网股票投资领域为例,进行研究框架与方法实践,实现语义完整的用户画像构建。本文为用户画像的构建提供了一种全新的研究思路。

1 相关研究概述

1.1 大小数据分析研究概述

本文的小数据仅指问卷样本单结构化数据,

而非在体量上“小”的针对单个事物所抽取的多结构化大数据。在社会科学研究领域,大小数据分析优劣势的争论主要有:

(1) 大数据分析是否能够真正代表社会本源^[1]。反对者认为“全样本数据仅是一个美丽的谎言,数据并不能代表社会生活的全部”^[2-3]。

“数据为王,技术主导”^[4]的观点也得以反驳,如 Silver^[5]认为数据驱动的分析低估了人的重要性。

(2) 大数据分析是否能够抛弃小数据分析中的理论假设。齐磊磊^[6]、彭理强^[7]、方环非^[8]等学者在充分剖析实证分析理论的历史来源及国内外代表性专家及著作理论观点的基础上,指出大数据分析是为了让理论在内涵及外延上得以更好的拓展,理论仍保持核心地位。

(3) 大数据分析强调“相关关系”是否可以忽略小数据分析的“因果关系”。史蒂夫·洛尔^[9]指出仅凭相关性是不够的,还要对因果关系产生启发性认识,两者需密切配合。Josh Cowls 与 Ralph Schroeder^[10]、彭知辉^[11]通过广泛的文献及访谈研究,指出相关关系是一种比因果关系更为广泛的概念,它是因果关系的派生关系,因果起着根本性的解释作用。

1.2 大小数据分析与用户画像构建研究概述

小数据分析有定性和定量分析两种形式,其中定性分析方法被广泛应用到用户画像构建中,张艳丰^[12]等人基于 SSO 理论框架,利用扎根理论从构想问题的文本式回答资料中抽取用户画像标签,通过问卷采集用户数据

实现不同群体用户画像分类。而采用“提出假设—构建模型—收集数据—验证假设”程式化定量分析过程构建用户画像的成果偏少,袁润^[13]采集科学网博客可量化的用户属性和行为数据,通过构建计量模型,进行统计学分析和验证,来构建学术博客用户画像。无论是定性还是定量分析,小数据分析的优势是:(1)具有长期反复论证的理论模型做基础;(2)数据可信度较高;(3)体现变量间因果关系。缺点是:(1)理论在反复验证中不断精细化,而难以创新;(2)数据范围有限,且主观性较高。因此,将大小数据结合来构建用户画像可以做到优势互补,大数据提供了让理论模型创新的数据条件,传统的理论模型也可以支撑大数据分析,通过理论解释现象^[14]。融合大小数据分析思维的用户画像构建,是要解决有关用户偏好的“相关关系”挖掘与“因果关系”分析的标签融合问题。关于这一问题有两个研究趋势:

(1) 基于传统理论引导大数据分类,获取不同分类标签,这一过程类似“有监督”式学习。林燕霞^[15]利用社会认同理论作为用户群体分类的理论基础,利用 LDA 主题模型、多元 logistic 回归等技术抽取群体用户分类偏好标签。李伟卿^[16]基于感知价值对网络消费者偏好进行预测,通过对消费者感知价值维度的分解,对海量用户行为数据进行目标价值分类,利用深度学习从不同感知价值数据中获取对应标签。

(2) 基于大数据构建用户画像过程中融

入情境因素,情境绝不仅指客观物理环境,而是包含了认知、意识、动机等主观因素^[17]。通过情境建模^[18]将用户画像由单一客观标签转为“客观—主观”多维标签。

2 融合大小数据分析的用户画像构建框架

围绕某特定领域构建“全貌”的用户模型,即要反映出用户感兴趣的领域主题是什么,又能反映出感兴趣的原因是什么。在用户画像构建中大小数据融合机制如下:

(1) 利用大数据分析从用户多结构化数据中获取偏好标签。由于实验环境所限,本文选择用户评论和行为数据为大数据来源,主要体现出与问卷样本小数据在研究范式上的差别,在数据量和数据结构上仍与大数据有差距。利用修正的 LDA 模型无监督学习偏好主题间的相关关系。

(2) 根据 LDA 主题模型中词频的分布,参考领域本体进行主题粒度设计,将主题引入传统的理论模型实现优化更新。

(3) 基于更新后的理论模型,提出研究假设,设计问卷收集数据,实现主题间因果关系论证。

(4) 对第一步生成的偏好主题间的相关关系和第三步生成的偏好主题间的因果关系,参照领域本体实现语义集成,构建语义完整的用户画像。其逻辑框架如图 1 所示,并在下文以互联网股票投资博客平台为例,按照该逻辑结构实现数据分析与实践。

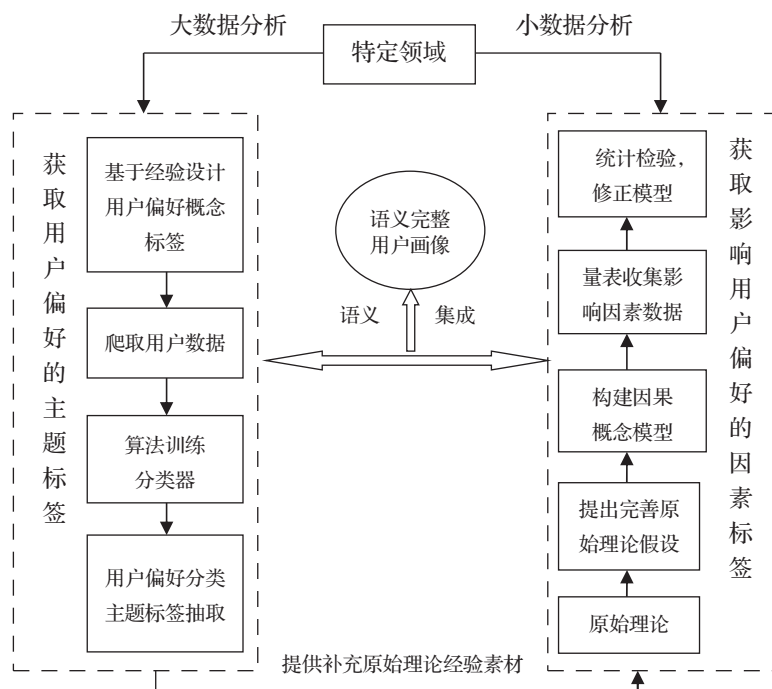


图1 融合大小数据分析的用户画像构建框架

3 基于修正的 LDA 模型获取用户偏好相关关系

3.1 问题的提出与概念标签设计

用户画像具有较强的领域性，不同领域的核心诉求不同，对概念标签设计的要求不同。以互联网股票投资领域为例，研究投资者情感分类的主题标签对刻画精准用户画像有着重要价值。因此，偏好及行为属性标签要比基础信息标签对用户画像构建更重要，基于经验及可获取的数据，将概念标签维度设计如下：

（1）内容偏好维度。主要是用户评论形成的短文本数据，对购买的股票或有关股票的新闻热点、网络大V的专评文章等形成评论或解读内容。这些内容中隐含着持有不同情感的群体用户对股票特征的偏好趋向。

（2）行为偏好维度。用户通过“关注”“点赞”“奖赏”及“转发”等动态行为，构建自己的社会化网络。拥有相似偏好或动机的用户会通过上述行为关联到一起形成群体用户簇，表达相似的兴趣趋向。

3.2 数据爬取与偏好情感分类

由于实验条件所限，仅反映出非结构化文本数据和行为数据与传统问卷数据在分析流程和研究范式上的不同，利用八爪鱼采集器从东方财富股吧官网爬取10000条用户评论，选取“众数”作为集中趋势分析指标，删除不包含无效关键字的评论，保留有效评论8350条。由于针对一个平台的用户数据，评论主题的集中趋势度高，而离散程度较低，主题分布受数据量影响较小。利用中科院NLPIR系统对评论数据进行偏好情感分析，部分结果截图如图3所示。

文件(F)	编辑(E)	格式(O)	查看(V)	帮助(H)						
通用情感分析										
文档总数	257	负面总数	233	负面占比	90.66%	正面总数	24	正面占比	9.34%	
标题	出处	发表时间	情感得分	正面得分	负面得分	原始链接	本地文件名			
工作簿29	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿29.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿29.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿29.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿29.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿29.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿29.txt	20210206 %H:02:55	-6	0	-6
C:\Users\lenovo\Desktop\预书库\站点采集\工作簿29.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿29.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿29.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿29.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿29.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿29.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿29.txt	20210206 %H:52:29	-6	0	-6
C:\Users\lenovo\Desktop\预书库\站点采集\工作簿26.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿26.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿26.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿26.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿26.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿26.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿26.txt	20210206 %H:15:19	-5		
股票代码-评论采集2	C:\Users\lenovo\Desktop\预书库\站点采集\股票代码-评论采集2.txt	C:\Users\lenovo\Desktop\预书库\站点采集\股票代码-评论采集2.txt	C:\Users\lenovo\Desktop\预书库\站点采集\股票代码-评论采集2.txt	C:\Users\lenovo\Desktop\预书库\站点采集\股票代码-评论采集2.txt	C:\Users\lenovo\Desktop\预书库\站点采集\股票代码-评论采集2.txt	C:\Users\lenovo\Desktop\预书库\站点采集\股票代码-评论采集2.txt	20210206 %H:06:06	-5	1	-6
股票代码-评论采集2.txt	C:\Users\lenovo\Desktop\预书库\站点采集\股票代码-评论采集2.txt	C:\Users\lenovo\Desktop\预书库\站点采集\股票代码-评论采集2.txt	C:\Users\lenovo\Desktop\预书库\站点采集\股票代码-评论采集2.txt	C:\Users\lenovo\Desktop\预书库\站点采集\股票代码-评论采集2.txt	C:\Users\lenovo\Desktop\预书库\站点采集\股票代码-评论采集2.txt	C:\Users\lenovo\Desktop\预书库\站点采集\股票代码-评论采集2.txt	20210206 %H:48:47	-1	0	-1
工作簿32	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿32.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿32.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿32.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿32.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿32.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿32.txt	20210206 %H:36:34	0	0	0
C:\Users\lenovo\Desktop\预书库\站点采集\工作簿32.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿32.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿32.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿32.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿32.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿32.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿32.txt	20210206 %H:51:51	-1	1	-1
工作簿8	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿8.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿8.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿8.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿8.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿8.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿8.txt	20210206 %H:53:58	3	5	-1
C:\Users\lenovo\Desktop\预书库\站点采集\工作簿8.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿8.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿8.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿8.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿8.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿8.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿8.txt	20210206 %H:55:21	1	3	-2
工作簿15	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿15.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿15.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿15.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿15.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿15.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿15.txt	20210206 %H:56:32	0	1	-1
C:\Users\lenovo\Desktop\预书库\站点采集\工作簿15.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿15.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿15.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿15.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿15.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿15.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿15.txt	20210206 %H:57:34	2	3	-1
工作簿10	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿10.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿10.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿10.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿10.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿10.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿10.txt	20210206 %H:44:36	-5	0	-6
C:\Users\lenovo\Desktop\预书库\站点采集\工作簿10.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿10.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿10.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿10.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿10.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿10.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿10.txt	20210206 %H:37:40	5	7	-2
工作簿13	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿13.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿13.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿13.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿13.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿13.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿13.txt	20210206 %H:40:21	-1	0	-1
C:\Users\lenovo\Desktop\预书库\站点采集\工作簿13.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿13.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿13.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿13.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿13.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿13.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿13.txt	20210206 %H:41:29	-6	0	-6
工作簿14	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿14.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿14.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿14.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿14.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿14.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿14.txt				
C:\Users\lenovo\Desktop\预书库\站点采集\工作簿14.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿14.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿14.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿14.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿14.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿14.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿14.txt				
工作簿20	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿20.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿20.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿20.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿20.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿20.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿20.txt				
C:\Users\lenovo\Desktop\预书库\站点采集\工作簿20.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿20.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿20.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿20.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿20.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿20.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿20.txt				
工作簿16	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿16.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿16.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿16.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿16.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿16.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿16.txt				
C:\Users\lenovo\Desktop\预书库\站点采集\工作簿16.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿16.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿16.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿16.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿16.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿16.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿16.txt				
工作簿17	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿17.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿17.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿17.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿17.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿17.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿17.txt				
C:\Users\lenovo\Desktop\预书库\站点采集\工作簿17.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿17.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿17.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿17.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿17.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿17.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿17.txt				
工作簿18	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿18.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿18.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿18.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿18.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿18.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿18.txt				
C:\Users\lenovo\Desktop\预书库\站点采集\工作簿18.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿18.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿18.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿18.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿18.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿18.txt	C:\Users\lenovo\Desktop\预书库\站点采集\工作簿18.txt				

图2 基于NLPIR系统对评论数据进行情感分析

3.3 融合内容与行为的LDA修正模型设计

LDA主题模型可有处理短文本数据中主题挖掘, 本文将用户评论内容与用户行为相结合, 参考郭光明在博士论文中所构建的LDA修正模型LUBD-CM^[19], 但考虑到: (1) 用户行为与所关注到的第K个主题相关, 可将行为参数元组 $\langle b_1, b_2, \dots, b_L \rangle$ 进行归一化处理为参数B, B代表不同用户行为发生的个数。(2) 考虑背景词对用户偏好分类的价值, 将背景词也归入主题词中。通过对LUBD-CM模型的修改与设计, 本文应用的LDA修正模型概率有向图如图3所示。

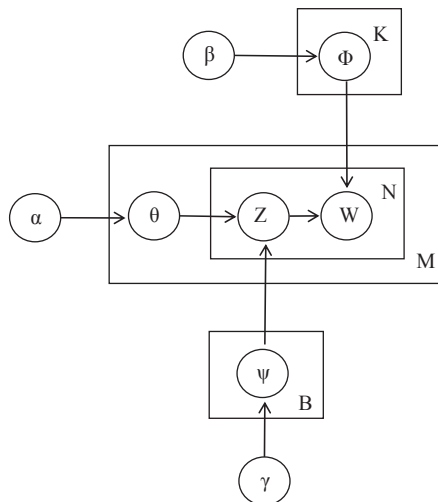


图3 融合内容与行为的LDA模型概率图

对于第m篇评论, z代表一次实验产生的

主题随机变量, $z \sim Multi(z | \bar{\theta}_m)$, $\bar{\theta}_m$ 代表第m篇评论的多项式分布参数, 长度为K, θ 是一个M*K的矩阵, 每一行代表一篇评论的多项式分布参数; 对于第k个主题, w代表一次实验产生的词随机变量, $w \sim Multi(w | \bar{\Phi}_k)$, $\bar{\Phi}_k$ 代表第k个主题的多项式分布, 长度为N, Φ 是一个K*N的矩阵, 每一行代表一个主题的多项式分布参数。加入行为参数B后, 将修正z与w值, 用户通过不同行为表达对一篇文章或一个主题的认同度, $z \sim Multi(z | \bar{\theta}_m, \bar{\Psi}_b)$, $w \sim Multi(w | \bar{\Phi}_k, \bar{\Psi}_b)$, $\bar{\Psi}_b$ 代表第b个行为的多项式分布, 长度取K与N的并集, 即 $K \cup N$, Ψ 是一个B*K*M的三维矩阵, 每一行代表一个主题或一篇评论的多项式分布参数。LDA修正模型的目标, 给定M个用户评论集, 求出隐含于其中的主题z, 概论公式如下^[20]:

$$p(z|w) = \frac{p(w, z)}{\sum_z p(w, z)} \quad (1)$$

采用Gibbs抽样学习公式(1)中的参数, 根据贝叶斯理论, 随机变量 $\bar{\theta}_m$ 、 $\bar{\Phi}_k$ 、 $\bar{\Psi}_b$ 的先验概率分布均服从Dirichlet分布,

$\overline{\Phi}_k \sim \text{Dirichlet}(\beta)$, $\overline{\theta}_m \sim \text{Dirichlet}(\alpha)$,
 $\overline{\Psi}_b \sim \text{Dirichlet}(\gamma)$, α 、 β 、 γ 分别为评论集中主题的先验超参数、主题集中词汇的先验超参数、行为集中主题与词汇并集的先验超参数。上述修正 LDA 模型的 Gibbs 抽样公式如下:

$$P(z_j = k | \overline{z}_{-i}, \overline{w}, \alpha \beta \gamma) \propto \Phi_{kt} \propto \theta_{mk} \propto \Psi_{b(k \cup N)} \quad (2)$$

其中, z_{-i} 代表第 m 条评论里去除了主题为 z_i

的其他所有主题, t 代表 N 词表中的词汇 t 。

3.4 与不同情感偏好相关的主题标签抽取

利用公式(2), 分别对偏好分类处理后所生成的正面数据集和负面数据集实现主题抽取。使用 Python 的 Gensim 库生成修正 LDA 主题模型抽取的主题及相应概率结果如表 1 所示。

表 1 正负面情感评价对应的主题词表

正面情感偏好主题						负面情感偏好主题					
主题1		主题2		主题3		主题1		主题2		主题3	
涨	0.057	上市	0.053	股价	0.068	跌停	0.073	股价	0.059	业绩	0.047
业绩	0.043	股价	0.048	涨	0.053	业绩	0.057	下跌	0.055	上市	0.039
效益	0.038	涨	0.036	业绩	0.048	股价	0.042	减持	0.037	收益	0.035
股价	0.027	业绩	0.028	收益	0.039	涨停	0.035	业绩	0.032	太低	0.029
涨停	0.013	稳定	0.013	利好	0.033	亏损	0.031	佣金	0.028	利空	0.017
绩优股	0.011	增持	0.012	走势	0.021	垃圾	0.017	疫情	0.017	股价	0.012
信心	0.008	红柱	0.010	长期	0.013	退市	0.012	利空	0.012	套现	0.010
反弹	0.005	高	0.008	激励	0.007	差	0.008	收益	0.009	熊市	0.007
牛市	0.003	龙头	0.006	高	0.005	骗人	0.003	亏损	0.007	暴雷	0.005
能赚	0.001	反弹	0.003	看好	0.001	洗盘	0.002	垃圾	0.005	高估	0.002

可见, 对股票投资持有积极情感偏好的用户对股价上涨、上市公司业绩优良、利好信息、收益高、大盘走势稳定等投资属性标签有较高关注度; 对股票投资持有消极情感偏好的用户对股价下跌、上市公司业绩亏损、利空信息、收益太低、佣金过高、疫情等投资属性标签有较高关注度。总之, 与用户投资情感相关的投资属性标签有: 股价、上市公司业绩、收益、市场利好利空、大盘走势、疫情等。

4 融合大数据分析结论与感知价值理论的用户偏好因果分析

4.1 基于LDA分析结论优化用户感知价值理论

顾客感知价值理论是由顾客内心的感知所

决定的, 是感知利得与感知利失之间的权衡^[21]。本文将股票投资顾客感知价值定义为感知利益与感知风险权衡的结果。上述 LDA 分析结果可为感知利益和感知风险的构建提供完善素材, 将感知利益划分为投资收益、股价涨幅、上市公司盈利能力、利好政策 4 个维度; 将感知风险划分为上市公司经营风险、本金回收风险、利空政策、疫情影响 4 个维度。

4.2 假设提出及模型构建

Mitchell 等^[22] 通过研究认为人们对金钱的理解会受到个体差异的影响, 不同理财性格将会选择不同方式的理财手段。崔冬亮^[23] 研究发现不同性格的投资者在选择基金时会有明显的

特征区别。因此,将投资性格作为外生变量,根据心理学的解释,从性格的态度特征、意志特征、情感特征和理智特征四个部分中,选取

内向型人格、独立性、情感控制力、学习能力为观察变量。本文所构建的股票投资顾客感知价值模型如图4所示。

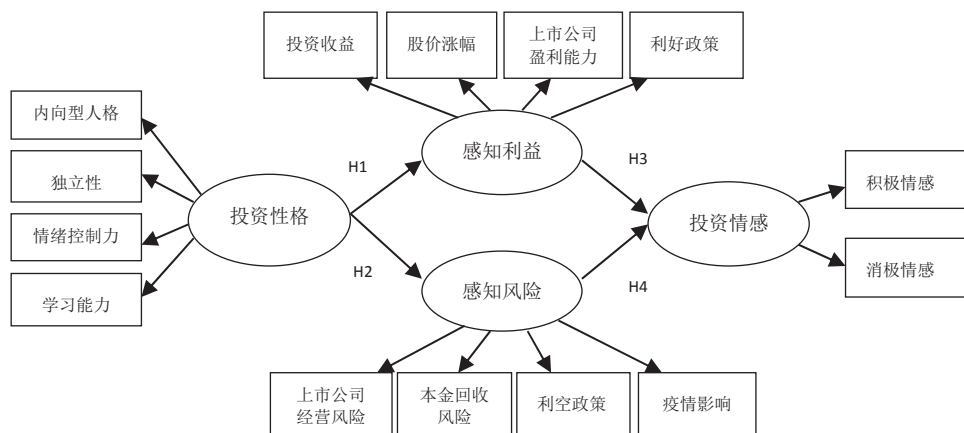


图4 股票投资顾客感知价值模型

该模型需要验证的是:(1)由“数据驱动”获取的主题标签是否对投资情感产生显著影响。(2)引入的外生变量“投资性格”是否对“数据驱动”获取的主题标签产生显著影响。因此,设定的基本假设如下:

H1: 投资性格对股票投资顾客感知利益产生正向影响

H2: 投资性格对股票投资顾客感知风险产生负向影响

H3: 感知利益对投资情感产生正向影响

H4: 感知风险对投资情感产生负向影响

4.3 数据收集与统计分析

4.3.1 问卷设计及信效度分析

参照已有的相关文献量表,利用问卷星平台制作问卷并网络发布,不设定具体人群,收集到有效问卷103份,使用SpssAU平台进行数据分析。首先对量表进行信效度分析,总体

Cronbach's α 值为0.863。另外,各题项的因子载荷值大于0.5,证明量表设计有效且适合因果分析。量表结构及测评值如表2所示。

4.3.2 假设检验及因素标签获取

使用SpssAU系统的结构方程模型SEM功能模块,进行路径分析。采用标准化路径系数指标检测模型中潜变量之间的关系是否成立,以验证假设是否成立。路径系数分析结果如表3所示。

因子A对因子B影响时,标准化路径系数为0.285>0,并且此路径呈现出0.05水平的显著性($z=1.591, P=0.031<0.05$),说明因子A对因子B产生显著的正向影响关系,假设H1成立。同理,因子A对因子C的路径系数为-0.417<0,且在0.01水平下显著,假设H2成立。因子B对因子D的路径系数0.619>0,在0.01水平下显著,假设H3成立。因子C对因子D的路径系数-0.429<0,且在0.05水平下显著,假设H4成立。

表 2 量表结构及效度测评值

潜变量	显变量	题项	标准化载荷系数
投资性格 A	内向型人格 A1	A11: 我性格比较内向	0.728
		A12: 我不喜欢独自一个人工作学习	0.623
	独立性 A2	A21: 我喜欢获得他人对我工作指导示范	0.631
		A22: 根据实际情况制定计划不受他人影响	0.689
	情绪控制力 A3	A31: 别人的赞扬批评对我有很大影响	0.774
		A32: 遇到问题我比较冲动	0.823
	学习能力 A4	A41: 我善于分析	0.645
		A42: 我做事果断, 看个股诊断只抓要点	0.607
感知利益 B	投资收益 B1	B11: 我非常在意投资收益的多少	0.889
		B12: 投资收益高的股票代表未来也是绩优股	0.537
	股价涨幅 B2	B21: 股价涨跌对我情绪有很大影响	0.695
		B22: 股价涨让我有不断增持的冲动	0.523
	上市公司盈利能力 B3	B31: 上市公司业绩舆论对我影响不大	0.514
		B32: 我喜欢龙头企业股票	0.639
	利好政策 B4	B41: 利好政策增强我持股的信心	0.592
感知风险 C	上市公司经营风险 C1	C11: 股票涨跌让我意识到经营风险	0.501
		C12: 上市公司套现让我感觉受骗	0.533
	本金回收风险 C2	C21: 我最担心血本无归的事发生	0.798
		C22: 更看重本金长期投资价值	0.674
	利空政策 C3	C31: 利空政策让我对股市感到焦虑	0.533
	疫情影响 C4	C41: 疫情对我投资股票有一定影响	0.629
投资情绪 D	积极情绪 D1	D12: 我比较看好股票投资	0.683
	消极情绪 D2	D21: 我比较厌恶股票投资	0.536

表 3 模型回归系数汇总表

路径影响关系	C.R.	S.E.	P	标准化路径系数
A -> B	1.591	0.109	0.031	0.285
A -> C	-3.011	0.126	0.005	-0.417
B -> D	4.438	0.083	0.001	0.619
C -> D	-3.172	0.151	0.018	-0.429

由此可见,用户的投资性格直接影响股票投资
感知利益与感知风险,且与“数据驱动”获取的主
题标签,共同影响投资情感。因此,影响用户股票
投资情感分类的性格等因素标签如表 4 所示。

表 4 影响用户股票投资情绪分类的性格标签

性格维度	影响投资情绪分类的性格标签	
	正面情绪投资性格标签	负面情绪投资性格标签
内向型人格	安静、善于思考、做事有计划等	好动、焦虑、做事不加思考等
独立性	独立自主、不受别人左右、果断等	盲目性强、易受暗示、优柔寡断等
情绪控制力	情绪稳定、遇事冷静、情绪独立等	情绪易冲动、有进犯性、情绪依赖他人
学习能力	深思熟虑、看问题全面、理性等	缺乏主见、人云亦云、感性等

5 基于语义集成的用户画像构建

为了使用户画像有更好的扩充性和兼容性,参照领域本体将上述获取的用户偏好相关主题词表与用户偏好因果主题词表进行语义集成,

实现概念的边界扩充,提高用户画像预测的范围和质量。本文对用户投资情感分类的主题词表进行词云化显示,并刻画出与其对应的投资性格与态度因素,互联网股票投资领域用户画像抽象表示如图 5 所示。

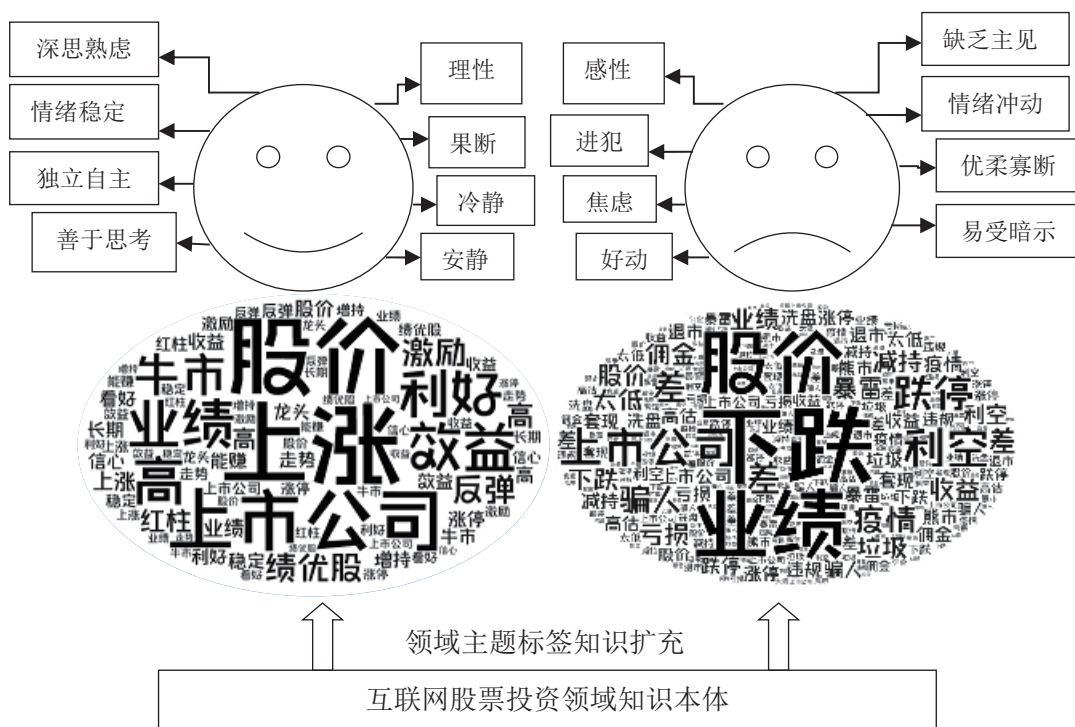


图 5 互联网股票投资领域用户画像抽象表示

利用八爪鱼从东方财富股吧官网爬取 5000 条评论数据作为测试集,分别检测包含相关关系和因果关系标签的用户画像系统(记

为: BSUP) 和去掉性格因素标签只包含相关关系标签的用户画像系统(记为 BUP),对测试集中高价值与低价值投资用户进行分类。

检测结果如表 5 所示,可见 BSUP 用户画像系统性能更优,因此本文所提出基于大数据分析构建融合主题标签和因素标签的用户画像方案可行,由于其描述用户兴趣偏好的完整语义性,更能提升预测质量。

表 5 基于 BSUP 和 BUP 用户画像系统的分类性能对比

用户画像系统	Recall	precision
BSUP	0.75	0.86
BUP	0.52	0.65

6 研究总结

本文研究的创新点是:(1)在大数据分析日趋白热化,甚至可取代小数据分析的理论学说下,找出两者的优势与不足,发现其研究的契合点。(2)提出基于大数据分析的用户画像构建框架,既发挥了“数据驱动”的客观性,又能使大数据分析结果根植于“知识驱动”的因果模型中,相互促进分析的质量。(3)将小数据分析互相嵌入,构建即包含相关关系又包含因果关系的完整语义画像,提升画像预测效果。研究的不足是:(1)以互联网股票投资领域为例,大数据分析中仅收集了用户评论和行为数据,在数据结构和规模上都有所限制。(2)在小数据分析中,所构建的因果关系模型并未涉及到投资性格潜变量的各维度与“数据驱动”获取的主题标签之间的相关关系,导致所构建的用户画像还不够全面精准。

参考文献

- [1] 黄欣荣. 大数据主义者如何看待理论、因果与规律——兼与齐磊磊博士商榷[J]. 理论探索, 2016(6):33-39.
- [2] 杨子飞. “第三洞穴”与“数据主义”——论大数据社会科学的内在矛盾[J]. 自然辩证法研究, 2016, 32(8):63-67.
- [3] 罗小燕, 黄欣荣. 社会科学研究的大数据方法[J]. 系统科学学报, 2017, 25(4):9-12.
- [4] 阎光才. 教育及社会科学研究中的数据——兼议当前的大数据热潮[J]. 北京大学教育评论, 2013, 11(4):78-86.
- [5] Silver N. The signal and the noise: Why so many predictions Fail—But some Don't[M]. New York: Penguin Press HC, 2012.
- [6] 齐磊磊. 大数据主义与大数据经验主义——兼答黄欣荣教授[J]. 山东科技大学学报(社会科学版), 2018, 20(2):16-21.
- [7] 彭理强. 大数据主义与经验主义——兼评齐磊磊博士的“大数据经验主义”[J]. 长沙大学学报, 2019, 33(3):82-85.
- [8] 方环非. 大数据:历史、范式与认识论伦理[J]. 浙江社会科学, 2015(9):113-120.
- [9] 史蒂夫·洛尔. 大数据主义[M]. 胡小锐, 朱胜超译. 北京:中信出版社, 2015.
- [10] Cows J, Schroeder R. Causation, Correlation, and Big Data in Social Science Research[J]. Policy and Internet, 2015, 7(4):447-472.
- [11] 彭知辉. 论大数据思维的内涵及构成[J]. 情报杂志, 2019, 38(6):126-130.
- [12] 张艳丰, 彭丽微, 刘金承, 等. 新媒体环境下移动社交媒体倦怠用户画像实证研究——基于SSO理论的因果关系视角[J]. 情报学报, 2019, 38(10):1092-1100.
- [13] 袁润, 王琦. 学术博客用户画像模型构建与实证——以科学网博客为例[J]. 图书情报工作, 2019, 63(22):13-20.
- [14] 齐佳音. 数据+or模型+:数据驱动与知识驱动的关系探讨[EB/OL]. (2017-3-11)[2021-12-12]. http://blog.sina.com.cn/s/blog_6cf93b650102xbo8.html.
- [15] 林燕霞, 谢湘生. 基于社会认同理论的微博群体用户画像[J]. 情报理论与实践, 2018, 41(3):142-148.

- [16] 李伟卿, 池毛毛, 王伟军. 基于感知价值的网络消费者偏好预测研究 [J]. 管理学报, 2021, 18(6):912-918.
- [17] 苏敬勤, 张琳琳. 情境内涵、分类与情境化研究现状 [J]. 管理学报, 2016, 13(4):491-497.
- [18] 史海燕, 韩秀静. 情境感知推荐系统研究进展 [J]. 情报科学, 2018, 36(7):163-169.
- [19] 郭光明. 基于社交大数据的用户信用画像方法研究 [D]. 合肥: 中国科学技术大学, 2017.
- [21] 李志清. 基于 LDA 主题特征的微博转发预测 [J]. 情报杂志, 2015, 34(9):158-162.
- [22] 孙宁. 网购顾客感知价值驱动因素的探索性研究 [D]. 大连: 东北财经大学, 2010.
- [23] Mitchell, Temce R, Amy E, et al. The Meaning of Money: An Individual Difference Perspective [J]. Academy of Management Review, 1999, 24(3):568-578.
- [24] 崔冬亮. 中国股票型基金市场投资者行为研究——基于感知利益与感知风险 [D]. 北京: 北京大学, 2008.