北京邮电大学

硕士研究生学位论文开题报告

学 号: 2016140003

姓 名: 钱梦莹

学 院: 信息与通信工程学院

专业(领域): 电子与通信工程

研究方向:

导师姓名: 邓钢

攻 读 学 位: 工程硕士

2017年12月19日

|  |  |  |  |
| --- | --- | --- | --- |
| 论文题目 | 用户行为分析平台的设计与实现 | | |
| 选题来源 |  | 论文类型 |  |
| 开题日期 |  | 开题地点 |  |
| **一、立题依据（包括研究目的、意义、国内外研究现状和发展趋势，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）（不少于800字）**  随着移动互联网、物联网、云计算等新兴信息技术在社会、经济各个领域的不断应用，全球数据量正呈现出前所未有的爆发式增长态势。与此同时，数据类型及来源的多样性、数据产生与分析的实时性、数据的价值性等复杂特征日益显著，标志着“大数据”时代的到来。用户画像是大数据技术的重要应用之一，其目标是在多维度上构建针对用户的描述性标签属性。利用这些标签属性，对用户多方面的真实个人特征进行描绘勾勒，可用于描述用户相关的兴趣、特征、行为及偏好。  自2009年新浪微博上线以来，凭借其自身独特的社交性，迅速发展壮大席卷全国，至今为止吸引了超过5亿用户，日均活跃人数超2000万。随着微博的影响力日益增大,提取微博主题的需求也与日俱增。用户作为社交平台的主体，成为推动网络事件发展不可忽视的力量，而用户在使用微博过程中做出的行为、表达的情绪和态度与用户自身以及微博上其他用户紧密有关，因此对微博用户的画像有助于了解用户身份、行为规律、态度、认知情况等，一方面用户的身份特征，从业方向，兴趣特征，消费习惯等信息具有珍贵的商业价值，对于产品推送和基于大数据的精准营销具有重要意义；另一方面，微博平台没有任何经济和框架的限制，在信息选取、关注议题和个人表达方面都保持着一定的独立性，用户通过参与微博事件讨论可以自由表达出自己的态度和立场，对其中蕴含的个人价值观，情绪和社会心态等社会心理类信息进行分析识别，在维护社会秩序、帮助政府部门重点人群进行监控、保护公民利益等方面都能起到重要作用。  目前对微博用户画像技术的研究和利用主要集中在两个方面：  1.心理学家普遍认为，个人的文法和表达习惯与个人情绪和行为方式有关，可以认为是个人心理情况的外化表达。2000年，Amichai-Hamburger和Ben-Artzi[1]首次从实证研究的角度证明了网络行为特征与心理特征(人格)之间存在相关关系。大量研究开始从社交媒体入手对社交媒体用户的性格和人格进行分析画像。 李昂[2]等人根据新浪“微博”的使用功能与数据格式,结合建模目标(心理健康状态)的特点,自行设计了4类45个网络行为特征来测量被试者的抑郁、焦虑水平。同时对用户微博文本进行挖掘，利用汉语改进的LIWC[3]软件来获取被试在上述88个内容类网络行为特征上的取值，以此评估用户的主观幸福感。汪静莹[4]等人设计情绪提取分析模型，宏观地分析微博用户在不同季节和时间的情绪变化，卢婷婷[5]等人分析微博用户的微博文本并建立14维社会态度模型，对用户的社会行为进行评价。中科院相关项目组则将多任务增量回归的分析方法用于文本挖掘，通过收集分析用户好友数，收藏数等10个维度的特征信息，使用多任务增量回归的分析方法，分析预测了用户的大五人格特征[6]。  2.另一方面的研究主要集中在大数据商业价值的利用上，为精准推送精准营销，定位目标人群服务。这一部分的研究重点集中在微博文本中微博主题的提取方面。在利用微博发现用户关注点的研究，高明等人认为每条微博通常涉及一个或多个主题，与 LDA 主题模型匹配，提出了基于 LDA 主题模型推断微博的主题分布和用户关注点[7]。北京大学的王臻皇等人对LDA模型进行了深入研究，提出可通过修改模型参数人为地合并，拆分主题，对主题模型进行干预[8]。 张哲[9]等人分析了用户登陆时长，微博数，关注数等静态信息以此对用户分类，并设计实现了用户画像的单个查询和批量查询系统。  本研究课题的目的是基于微博将繁杂的用户信息，通过数据挖掘从用户关注特征和心理人格特征进行画像，使画像丰满，维度充足，以对用户进行更为准确的定位。画像的主要工作是标签化，即使用高度精炼，语意突出的短文本来为用户打标签。人能很方便的理解这些短标签，这就使得用户画像具有了实际意义，能很好的满足业务需求。  主要参考文献：  [1]Amichai-Hamburger Y A, Ben-Artzi E. The relationship between extraversion and neuroticism and the different uses of the Internet. Comput Human Behav, 2000, 16: 441–449  [2] [基于网络数据分析的心理计算:针对心理健康状态与主观幸福感](http://kns.cnki.net/kcms/detail/detail.aspx?filename=KXTB201511006&dbcode=CJFD&dbname=CJFD2015&v=" \t "http://kns.cnki.net/kcms/detail/frame/kcmstarget)[J]. 李昂,郝碧波,白朔天,朱廷劭.科学通报 . 2015(11)  [3] Gao R, Hao B, Li H, et al. Developing simplified Chinese psychological linguistic analysis dictionary for microblog. In: Imamura K, Usui S, Shirao T, et al., eds. Brain and Health Informatics: International Conference, BHI 2013, Maebashi, Japan. Proceedings. Cham: Springer, 2013. 359–368  [4] 基于微博用户的情绪变化分析,朱廷劭,汪静莹,中国科学院大学学报  [5] 卢婷婷.基于短文本的互联网用户意图识别方法及应用研究[D].济南:济南大学, 2016.  [6][多任务回归在社交媒体挖掘中的应用](http://xueshu.baidu.com/s?wd=paperuri:(a714f2adbcc5a94c274fc093a63ee340)&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http://kns.cnki.net/KCMS/detail/detail.aspx?filename=hebx201409017&dbname=CJFD&dbcode=CJFQ&ie=utf-8&sc_us=13325321085147470988" \t "http://xueshu.baidu.com/_blank)，白朔天，[《哈尔滨工业大学学报》](http://xueshu.baidu.com/usercenter/data/journal?cmd=jump&wd=journaluri:(f82a545a5fa63175)%20%E3%80%8A%E5%93%88%E5%B0%94%E6%BB%A8%E5%B7%A5%E4%B8%9A%E5%A4%A7%E5%AD%A6%E5%AD%A6%E6%8A%A5%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=publish&sort=sc_cited" \t "http://xueshu.baidu.com/_blank" \o "《哈尔滨工业大学学报》),2014,46(9):100-104  [7] 高明,金澈清,钱卫宁等.面向微博系统的实时个性化推荐[J].计算机学报,2014,37(4): 963-975.  [8] Wang Z, Chen S, Yuan X. Visual Analysis for Microblog Topic Modeling. Ruan Jian Xue Bao/Journal of Software (in Chinese)，2013  [9] 张哲，基于微博数据的用户画像系统的设计与实现[D]，湖北：华中科技大学，2015 | | | |

|  |
| --- |
| **二、研究内容和目标（说明课题的具体研究内容，研究目标和效果，以及拟解决的关键科学问题。此部分为重点阐述内容）（不少于2500字）**  随着大数据相关技术的应用越来越广泛，互联网逐渐步入大数据时代，这使每个互联网产业参与者的行为都发生了改变。其中最大的改变莫过于普通民众的行为在大数据分析系统面前将是“可视化”。其中有效地整合民众在社交媒体上的言论和观点，并精准的对用户进行“画像”是高效利用大数据技术维护社会治安，评估舆论方向，进行精准营销等活动的前提。  1.研究内容：  本课题以广大微博用户在微博上留下的海量文本信息为基础，对用户关注，转发微博，发表微博，好友互动等行为进行分析，设计实现了微博用户的画像系统。本课题的研究内容主要包括以下3个部分：  1.1 微博数据的收集与处理  本课题研究的数据集主要来源于对微博数据的爬取，爬取数据的策略初步设定为：从某一用户出发，分别以某一数值为阈值，爬取该用户的粉丝和关注用户列表，并以每一个粉丝（或者关注用户）为新的结点继续查找新的粉丝和关注用户列表进行爬取。  在数据处理部分，主要进行三个步骤的处理，首先将数据进行分开建档，其中用户个人的博文文本储存为一个长文本词袋，为后续的主题模型处理提供原始数据。  其次，对抓取后的数据进行预处理，预处理包括中文分词和停用词过滤两个部分。不同于英文数据中由空格符进行分词，中文的词与词之间没有明显的分隔符号，中文分词的目的再于将中文语句中的词与词分隔开来以供机器识别，中文分词也将直接影响到信息抽取、文本内容分析等的效果。图1为中文分词结果图例。    图1  同时因为用户的博文中为了使语义连贯，通常会使用一些标点符号，所以在进行分词处理后，需要将数据进行清洗。停用词过滤即为去除微博文本中所包含的对用户特征挖掘并无太大意义的词语或符号，这一步对于降低信息冗余度、提高兴趣分析准确度方面是非常必要的。  至此完成所有源数据的收集储藏及准备活动  1.2 利用算法实现用户特征信息模型的构建  由于微博文本信息具有文本短，表情多，不规范用语多，网络词汇多等特点，传统的文本挖掘技术如 VSM、，LSA，PLSA等已经不能有效地处理 Web 2.0 时代下社交网络的文本内容噪声多、用词不规范、词频较少等问题。主题模型可以更好地利用这些文本信息，于在线社交网站文本挖掘方面有着广泛的应用。    图2 LDA原理示意图  如图2所示，LDA(Latent Dirichlet Allocation)是一种文档生成模型。LDA主题模型将文本分为文档层、主题层和单词层，文档和主题相关联主题和单词相关，联假设一篇文档是由主题集合中的各个主题按照一定的比例构成的，而每一个主题，又是由单词表中的单词按照一定的比例混合而成的.通过机器学习的方法可以得到文档的主题。算法输入输出说明如图3所示：    图3 LDA算法说明  LDA 模型通过学习文档集中的单词挖掘出所有潜在的主题信息,并计算出每篇文档相对于这些主题的概率分布，从而判断用户主要对哪些话题抱有兴趣，通过这些信息来挖掘该文档集以外的其他文档的主题分布，可直观看出网民对各种微博主题感兴趣的情况。  1.3 利用回归算法对用户人格建模  使用textmind工具可以很方便地获取特定用户的情绪向量，包括正向情感词、负向情感词、焦虑词、生气词、健康词、关爱词、脏话、惊叹词等情绪传达词的所占比率，根据这些情绪向量，连同微博的静态特征一起，通过回归算法进行有监督的机器学习，可以对用户的大五人格特征进行预测。  回归算法：在机器学习算法中，回归算法是一个重要的应用，主要研究目标问题和假定条件之间的关系。这种技术通常用于预测分析，以及发现变量之间的因果关系。主要包括线性回归算法，逻辑回归算法，高斯回归算法，增量回归算法和岭回归算法等，这些算法有各自的侧重点和使用条件，选择哪一种算法应按具体需求和训练效果决定。  建模的思路为，对输入的训练对象根据不同的回归算法，使用某一种假设模型进行计算分析，这些模型中通常包含有未知的参数，将分析的结果同监督信号进行对比，从而修正模型参数，重复地进行这个学习过程，可以估计出模型所需要的参数，从而达到当新的输入对象到来时推断出正确的输出结果的预测效果。    图4有监督机器学习原理图  20150819114740045  图5某非线性回归算法拟合效果图  研究者通过词汇学的方法，发现大约有五种特质可以涵盖人格描述的所有方面，在大五人格模型中他们作为五个衡量维度依次为：开放性（O）责任心（C）倾向性（E）宜人性（A）神经质性（N）。大五人格可以通过NEO-PI-R评定，该评价模型模型在许多方面都具有广泛的应用价值。目前关于该人格模型的心理学研究已经比较成熟，有很多针对该模型的心理学图谱分析方法，有研究成果认为，同时具有低责任心和低宜人性的人格特征有较多的违法行为；同时具有高倾向性、低宜人性、低责任心人格特征的人，常发生与外界冲突的行为问题，而同时具有高神经质、低责任心人格特征的人则经常表现出由内心冲突引起的问题。大五人格是20世纪90年代以来最活跃的人格研究课题，也是目前对人的基本特质最理想的描述之一，因此本次研究以此作为人格模型对用户人格进行分析预测。  该部分的研究思路为：研究前期通过线上问卷调查的方式，收集一部分微博用户的微博数据及个人的大五人格指标，作为训练集，采用回归算法找到、建立从这些用户的微博文本反映出的情绪向量集X到其大五人格指标的传递矩阵W，并用以预测给定某用户的情绪向量所反映出的大五人格指标。  1.4 设计平台对画像结果进行展示  经上述处理得到的用户画像，将推送到展示层进行可视化展示，展示层的搭建采用前端JavaScript与html绘制页面与SpringMVC框架搭建后台相结合的方式，形象化地展示画像结果。在页面设计部分，基于JavaScript从传统的Jquery到时下新兴的vue有众多插件和脚手架可供选择，而html结合css本身也具有非常强大的功能和延展性，可实现包括3d动画效果在内的大多数视觉特效。后台则主要负责数据库的存取与数据的推送功能，数据的分析和计算都在上述建模过程中进行。初步设定的需求为，输入已知用户微博昵称或数字id，系统即对用户开始进行分析画像，并用户的画像结果包括兴趣爱好，关注内容等进行可视化展示，同时展示用户在五大人格特质上的性格特征，及该用户的主要社交网络情况，从多个角度对画像结果进行展示。  2.技术难点  对于本系统，主要面对3个技术难点，一个是，如何筛选辨别有分析价值的活跃用户；第二个是建立LDA模型对主题进行提取需要正确描述主题；另外一个是如何建立从微博文本特征向量到用户人格向量的模型。   1. 考察现实中微博用户的社交情况，有些用户微博数目过少，这样由于语料过少分析结果会产生很大偏差，没有分析价值，另一些除了用户自身正常吸引的粉丝外，还有很多由系统自动产生的恶意注册的用户，就是所谓的“僵尸粉”，这些微博用户由系统自动生成，无考察分析意义，且“僵尸粉”数量众多。如果爬取数据时收入则会对正常的分析带来很大干扰，因此甄别和屏蔽这些用户十分重要。现下处理筛除“僵尸粉”的技术已经非常成熟，这些识别技术准确率非常高，并考虑多种情况，但是过程一般比较复杂，还要综合考虑分析用户的静态信息，并对文本依照特定的算法进行分析拆解比对，考虑到本项目主要的目标和工作重心不在僵尸用户的识别，且需要爬取的用户数量庞大，应尽量避免在爬虫过程中进行大量计算，因此考虑使用尽量简单的过滤算法效剔除典型僵尸粉和营销号，对语义化非常强烈的僵尸粉则进行正常分析和主题提取。 2. 在创建使用LDA模型提取主题的过程中，最终可以得到的结果是与同一主题有关的语的聚类，以及每个用户的微博文本相对于这些主题的概率分布。如图7所示，通过观察三个主题中的高权重词语，人为地可以理解每个主题的主题词是什么：运动，空间探索和电脑，但是电脑并无法自行总结出这些主题词。LDA的成功很大程度上源自它产生可解释主题的能力，但这些主题被分类产生后如何对其总结命名也是一个非常重要的工作。     图6 LDA主题提取结果示意图   1. 本次与用户人格的研究和预测是基于用户文本内容挖掘的，然而从文本中提取的信息数量庞大，仅使用textmind软件即可提取超过100个文本特征向量，如果将其全部单独作为一个维度对用户人格进行预测则算法计算开销非常庞大，而且实际上考察这些文本向量对用户人格的反映程度，有些可以在很大程度上反映用户某一人格特征，有些则不具有分析价值，因此，如何筛选和过滤有分析价值的向量进行分析也是一个难点。 |

|  |
| --- |
| **三、研究方案设计及可行性分析（包括：研究方法，技术路线，理论分析、计算、实验方法和步骤及其可行性等）（不少于800字）**  本论文拟设计并实现基于微博的用户行为分析系统，并搭建展示平台对结果进行可视化展示。具体的，本论文将沿袭以下的技术路线完成本系统：  1. 对相关领域的技术进行充分调研  充分而有效的调研可以保障在工程设计上不会出现技术路线的偏差，成熟的技术架构在满足设计需求的同时也可以为开发提供可靠的保障。  首先，在确定系统的实现的前景时，对目标领域的相关研究进行了调研，汲取了近几年相关研究的研究成果，明确了研究的主要难点和方向。结合实际中互联网时代用户充分利用大数据技术优势进行精准营销和人格判别的需求，设定了具有现实意义的研究内容。设计平台的主要功能的时候，收集并归纳了用户的要求，对当前互联网常用架构进行了分析与总结，并与导师和领域的专家进行了充分的交流和讨论。  其次，在工程实现的具体过程中，扎实的理论基础具有很重要的作用，对于涉及到的相关技术，提前购买了相关的书籍，进行充分的学习。工程性的项目在实现的过程中会在所难免的出现这样那样的错误，充分利用搜索引擎结合已经了解的理论基础，可以迅速定位到程序的出错点。  经过上述环节的调研和学习，对所要学习的内容、设计的流程和可能遇到的困难都有了一定的了解。   1. 对相关理论的学习和平台搭建的方法   本论文工程项目的数据处理部分需要学习多种建模方法包括LDA和回归算法，前期大量的调研工作中，已对相关算法进行了初步学习，但是这些算法在实际应用中针对所训练的模型的一些特殊要求，可能还需要进行进一步的改编或完善。而在界面搭建部分，对于前后端涉及到的技术已经有了一定的了解，并拥有相关经验，对搭建平台所使用的编程语言包括JavaScript和Java也都进行过学习。  3．分阶段分模块对项目进行测试评估  在工程项目的实现过程中，进行每个模块单独设计搭建的架构思路，每一个功能模块搭建完成之后即对该模块进行测试，测试成功再进行下一个模块的搭建，这样可以保证不同功能间的交叉干扰。在设计每个功能模块时，都先用少量的数据按照算法训练模型，观察训练效果，确保方案可行再对大量的数据进行处理，这样可以避免在不正确的方向上浪费过多时间。 |

|  |
| --- |
| **四、本研究课题可能的创新之处（不少于500字）**  关于创新性主要体现在以下几个方面：  1.独立设计并实现基于用户文本数据的用户画像系统  本系统从用户特征属性信息，人格属性等多维度的对特定用户进行全方面的画像，不仅能够分析出用户的主要兴趣和关注点，用于精准营销和信息推送等商业行为，还可以预测推断用户的人格特性，若将二者结合进行分析，可在一定程度上推定用户身份，相比于目前的研究大多数针对某一特定角度进行分析画像，本研究极大地丰满了画像的内容，为后续对用户进行分类或者甄别提供了支持。  2.向量设定  一方面，创新性地使用由textmind提取的情绪词作为量化大五人格特征的基础向量，利用机器学习技术挖掘不同用户情绪词比例关系反映出的人格特征。情绪词例如愤怒性词汇，包容性词汇，积极性词汇，消极性词汇及其出现的比率一定程度上对微博用户个人的人格属性有着相当密切的关系。人类自身的性格和态度也是依赖这些词汇进行外化表达的，相对于同类画像研究使用的微博登陆时常，登录频率的研究或者简单进行积极情绪和消极情绪区分的研究都更具有参考性和实际价值。  另一方面，这些从文本分析中得到的向量数量众多，考虑到并不是每个向量都是对人格的有效反映，至少反应的程度有所不同，相对于目前的研究不加分辨地使用所有向量进行分析，筛选有效的向量，或者对不同的向量赋予一定的权重可以极大地减少计算开销，并有可能改善模型的训练效果。 |
| **五、研究基础与工作条件（1.与本项目相关的研究工作积累基础 2.包括已具备的实验条件，尚缺少的实验条件和拟解决途径）**  1. 技术支撑  自从 Twitter 于 2006 年正式上线面向公众开放以来,微博(Microblog)就受到了广泛关注,并微博文本内容包含了大量的信息,从中挖掘的信息可以用来进行突发事件监测、事件发展态势预测、精准营销等.目前基于大数据的文本分析和主题挖掘都有大量的研究，算法和语料库充足完备。  而经历了互联网将近20年的高速发展，现有的用于实现Java Web的技术也早已趋于成熟，企业级轻量化网站架构的设计与实现很多已经进行了开发流程的标准化。很多新兴的框架和搭建手脚架也为平台的搭建提供了方便。  2. 合作交流以及开源技术  机器学习和文本挖掘都是目前大热的技术领域，具有大量开源算法和模型，也有成熟的学习路线，同时广大技术论坛和社区也提供开源代码，充分利用现有的开源代码进行相关的整合与修改会大大加快项目进展。  平台搭建部分，一方面时下许多新兴的网站架构框架对许多常用功能都进行了集成和封装，可以直接进行调用，一定程度上减少网站搭建的代码量和时间，另一方面，关于传统的Javascript和html应用研究已经进行得非常深入，可满足所有的可视化需求。  3.经验储备  本人具有一定的网站搭建经验，并且经过前一阶段的调研与学习，对机器学习的分类算法和回归算法都有一定的了解，已经完成了主题模型部分的初步设计，使用小部分数据进行实验性处理，取得了一定的进展。 |

**学位论文工作计划**

|  |  |  |
| --- | --- | --- |
| 时间 | 研究内容 | 预期效果 |
| 2017年12月至2018年3月 | 学习调研阶段。学习调研网站平台搭建技术相关技术，学习机器学习及主题模型相关算法，所用软件，同时查看技术博客以及相关论文。 | 对项目中涉及到的知识、框架以及前端知识算法有深入的了解，掌握调试方法，了解使用机制。 |
| 2018年3月至2018年6月 | 方案设计阶段。根据第一阶段的调研结果，确定项目不同模块的选型，确定所使用的算法，语言，框架，并着手进行搭建。 | 根据方案，验证框架切实可行，取得初步进展。 |
| 2018年6月至2018年12月 | 程序开发阶段。按照设计需求，逐步实现程序功能，解决开发过程中出现的问题。修正框架中不合理的地方。 | 程序开发顺利，全部功能得以实现，系统运行流畅。 |
| 2018年12月至2019年3月 | 验证完善阶段。进行程序的验证测试，完善平台的中细节。开始论文的撰写工作。 | 对系统进行进一步的优化，并将成果转化为学术论文。 |
| 2019年3月至2019年6月 | 论文撰写阶段。整理课题进行期间的研究资料，撰写硕士学位论文。 | 完成硕士学位论文的撰写工作。 |
| 2019年3月至2019年6月 | 论文修改整合阶段。向导师及相关人员争取修改意见，进行最后的修改。 | 完成硕士学位论文。 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 评  定  小  组  成  员 | 姓 名 | 职 称 | 单位名称 | 职务 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| 导师意见： | | | | |
| 同意开题 | | | | |
| 导师（签名）：  日期： 年 月 日 | | | | |
| 开题报告小组意见： | | | | |
| 组长（签名）：  日期： 年 月 日 | | | | |
| 学院意见（签章）： | | | | |
| 负责人：  日期： 年 月 日 | | | | |