## 本科毕业设计(论文)任务书

|  |  |  |  |
| --- | --- | --- | --- |
| 学 号 | 1802250006 | 姓名 | 胡婷霞 |
| 届 别 | 2018 | 专 业 | 数据科学与大数据技术 |
| 毕业设计  (论文)题目 | 基于社交数据的用户画像系统设计与实现 | | |
| 主  要  研  究  内  容  及  任  务  要  求 | 一、主要研究内容  基于在线社交的用户画像系统设计主要由数据采集和数据挖掘两大部分，通过其中不同的模块进行功能组合，系统最终实现对用户的隐形爱好挖掘与既定属性预测两个方面的画像。  二、任务要求  （一）数据采集模块设计  1、用户数据模型  在确定了系统画像对象以及画像方向及属性后，通过引入用户数据模型，确定社交网络用户画像方向与数据信息，即对具有画像标签相关特征的用户群体进行资料获取，分为以下几步:  （1）获取目标画像用户群体；  （2）获取用户页面的基本数据；  （3）获取用户页面的详细数据；  （4）获取用户质量数据。  2、数据采集与处理模型  数据采集与处理模型基于用户数据模型定义的所需画像系统的数据样本进行数据采集。结合不同的社交网站结构进行分析发现，编写网络爬虫的方式进行数据采集多需要模拟登陆才能获取到数据,对每个用户数据获取时需要发起大量的HTTP请求，并且时常面临反爬虫策略。  通过数据采集模块获取到样本数据后,考虑到脏数据对画像结果的影响需要进行数据预处理,来保障数据挖掘算法的稳定性,系统通过利用用户数据模型中的用户质量数据，对僵尸、潜水账号进行过滤清洗。  （二）数据挖掘模块设计  1、隐性偏好挖掘模型  用户的隐性偏好挖掘为不设画像方向，意为对系统画像目标的测试集进行直接的文本挖掘、分析，从而得到测试集的偏好特征，并以此得到用户的画像。隐形偏好挖掘模型在本画像系统设计中是基于分词模块实现。  当数据采集与处理模型获取画像目标的文本数据集汇总后,通过引入第三方分词类库对文本语句进行分词，无属性的词语势必不能作为隐性偏好，所以便需要对这类词语进行过滤。隐性偏好挖掘模型通过引入总结出的中文停词表，实现对分词列表中无意义词汇过滤。最终通过展现过滤后按词频排序的词语列表,实现目标用户群体的隐性偏好挖掘。 | | |
| 主  要  参  考  资  料 | 1、王宪朋.基于视仁大数据的用户画像构建[J].电视技术,2017,41(6):20-23  2、余孟杰.产品研发中用户画像的数据建模[J].设计艺术研究,2014,4(6):62-64  3、百 度 百 科 . 大 数 据  4、 X. Qin, B. Qin, C. Li, et al. Landscape of Unified Big Data Platforms[J]. 2014  5、杰. Hadoop 大数据挖掘从入䰘到进䱦实战[M]. 机械工业出版社, 2018, 24-25  6、 田承平. 基于 B/S 架构的统一䝽置管理系统的设计与实现[D]. 南京:东南大学, 2014,15-25  7、 张智, 嗊宇. 分布式存储系统 HBase 关䭞技术研究[J]. 现代计算机(专业版)(32):35-39  8、 张䟾. Hadoop Hive 实现日志数据统计[J]. 电脑编程技巧与维护(4):115-117  9、 S. Hoffman. Apache Flume: Distributed Log Collection for Hadoop[J]. 2015 | | |
| 进  程  计  划 | 11月，对项目进行初步分析，构思，查找资料  12月，将完成项目二分之一  1月，完成项目，对项目进行总结。 | | |
| 教研室  主任  意见 | 教研室主任签名： 年 月 日 | | |
| 签  名 | 学生签名： 年 月 日  指导教师签名： 年 月 日 | | |

说明：本表须双面打印。