热门餐厅数据分析与推荐系统研究与实现

开题报告

班级（学号）：计科 1606 班（2016010276）　　姓名：杨芳芳
指导教师：田英爱

一、综述

随着网络时代的不断发展，电子信息技术的应用领域也随之不断扩张，越来越多的商家开始将其店铺信息发布到互联网上，形成 O2O 模式。同时，大量的消费者选择使用手机、电脑等互联网终端，来获取生活服务信息并对店铺信息进行评价。互联网已成为消费者获取商品信息的重要媒介，也成为商家与消费者联系的重要纽带。然而，消费者在获取生活服务信息的同时，也面临着如何从海量信息中如何获取有效信息的问题。

目前，获取生活服务信息的主要来源基本为大众点评，百度等软件。该类别软件的使用者众多，并且使用者对餐饮行业的评价与要求也较高。这使得这些软件在某些方面暴露出了明显的短板：软件可能无法满足所有使用者的个性化需求。因此，本课题是通过使用网络爬虫技术，精确获取大众点评上已有的餐厅信息和评论。随后将已爬取的信息用于研制一套针对用户个性化需求的餐厅推荐系统。该系统将同时具有浏览餐厅基本信息、餐厅词云、个性化推荐等功能，以此满足用户的个性化需求。

二、研究内容

本课题将使用大众点评网站上影响力较大的城市例如北京、上海、广州等，用这些城市餐厅商户的信息作为依据，根据用户偏好实现个性化推荐。本课题的研究内容具体如下：

1、 数据集获取

本项目针对大众点评网站设计爬虫框架，自动获取餐厅信息、用户评价等。此框架保证了大量餐厅的用户评论信息能够快速进行采集。本部分的难点为：更换 IP 代理、自动翻页、验证码识别、用户评论集解析。

2、 城市报告

对热门餐厅的评分系数及各项指标进行数据分析，得出各大城市味觉特征及综合指数等数据，生成城市报告。

3、 基于 TF-IDF 算法的关键词提取

为了从完成采集的用户评论文本获取有效信息，本项目将需要进行主题词提取。由于餐厅评论文本信息复杂，本项目采用 TF-IDF 算法，进行关键词提取，并根据关键词生成词云。

4、 基于 ECharts 的数据可视化

对热门餐厅进行数据对比和分析，研究餐厅受欢迎的因素，并使用 ECharts 进行数据可视化。

5、 构建个性化推荐系统。

根据每个用户的评分偏好和关注主题偏好，对用户推荐符合用户偏好的餐厅。本项目中

将分别运用两种推荐算法：基于内容的推荐算法和协同过滤算法。本项目将系统根据实际运行情况，选择一种比较符合用户偏好的推荐算法进行展示。

三、实现方法及预期目标

[实现方法]

本系统是由多个模块组成的较为基础的餐厅推荐系统，需要在保证整个系统能够正常运行的情况下，从用户角度提升用户的体验舒适度。下面是对主要需要完成的模块进行实现方法的说明。

1、 开发工具

本项目主要采用 PyCharm 开放平台利用 Python 语言来实现的。PyCharm 是一种 Python IDE，带有一整套可以帮助用户在使用 Python 语言开发时提高其效率的工具，比如调试、语法高亮、Project 管理、代码跳转、智能提示、自动完成、单元测试、版本控制。

2、 Web 端实现

因本项目都是在 PyCharm 开放平台上完成，为确保项目能够正常运行，使用 Django 框架进行 Web 开发。Django 是一个开放源代码的 Web 应用框架，由 Python 写成。采用了 MTV 的框架模式，即模型 M，视图 V 和模版 T。

3、 热门餐厅推荐系统的分析与设计

本项目主要通过 Web 前端界面实现用户交互，展示推荐、生成词云等功能。其后端部分主要为数据处理部分，即数据分析、词云生成、推荐模型训练。热门餐厅推荐系统总体数据处理框图如图 1 所示，主要分为三个阶段：数据准备、数据处理、数据可视化及 WEB 展示。

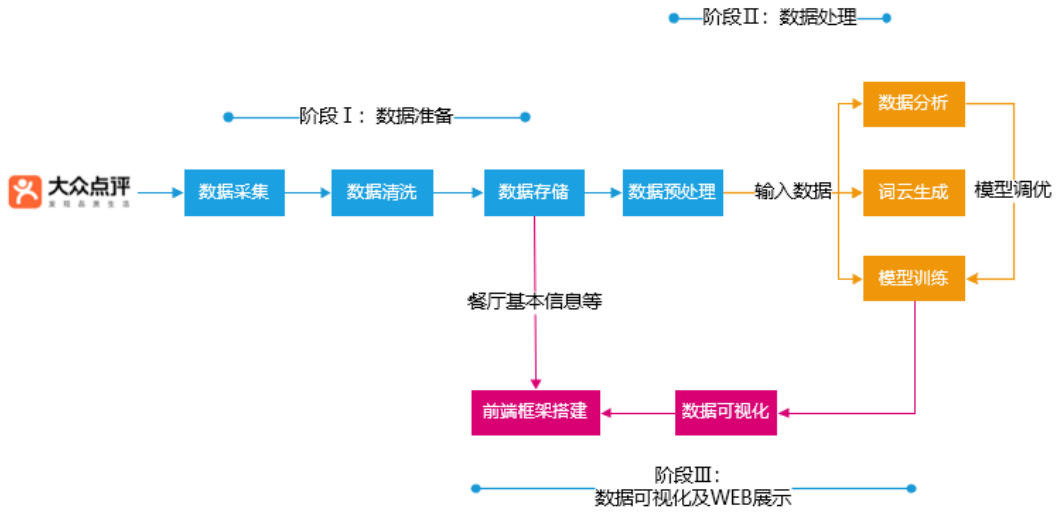

图 1 热门餐厅推荐系统总体数据处理框图

4、 基于 Web Spider 数据获取的分析与设计

本项目针对大众点评美食网站设计爬虫方案，要爬取的数据主要为餐厅基本信息、用户评论等。此爬虫可以依据数字字典格式，爬取到有效信息，其主要流程如图 2 所示。首先在任务开始时，需要设置用户代理池，即 IP 代理，将满足需求的 IP 存入 IP 队列。然后根据想要获取数据的 URL 抓取数据，对数据进行清洗，然后保存到数据库中。

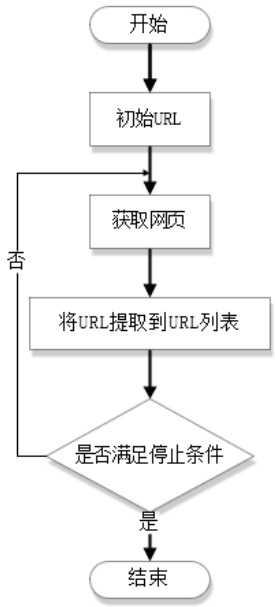

图 2 爬虫框架

5、 基于 TF-IDF 算法的关键词抽取分析与设计

本项目需要根据用户评论，生成词云，而关键词的提取是该模块的重点。

关键词是指语料中与主题高度相关的词语，目前较为经典的关键词提取技术是 TF-IDF 算法。TF-IDF 算法是基于概率统计的方法，其含义为：如果某个词在文本中出现的频率高，而在其他文本中很少出现，则认为该词与文本主题相关，即关键词。

基于上述，采用 jieba 分词器对评论进行分词处理后，使用 TF-IDF 算法统计出关键词，最后根据计算出的关键词生成词云。

6、 个性化推荐系统的设计

本项目采用的推荐算法主要如下：

1) 基于内容的推荐算法

基于内容的推荐首先要确定用户的偏好，获取兴趣偏好往往需要利用数据挖掘等方法通过对用户历史信息的挖掘，得到用户在一定时期比较稳定的喜好，然后将与用户兴趣偏好相匹配的推荐对象推荐给用户。

2) 协同过滤推荐算法

协同过滤算法也是基于用户偏好来进行个性化推荐，同样需要利用数据挖掘等方法获取。然后，计算该用户与其他用户的偏好相似度，找到兴趣偏好相似的用户。最后根据这用户对某类对象的兴趣偏好程度得到该用户的兴趣偏好。

这两种推荐算法都是基于用户偏好进行推荐，因此在该部分中，获取用户的历史偏好是实现推荐的关键。对于首次使用该系统的用户，默认推荐用户所在地的排行榜 TOP 10。用户需要注册/登录才能使用个性化推荐功能。在注册过程中，需要进行问卷填写，首次个性化推荐会根据问卷内容进行用户情感分析，从而进行推荐。用户在后续的浏览过程中，若点击“喜欢”，推荐算法将会根据该动作和历史记录，重新进行推荐。

7、 数据库设计

预期数据库如下表：

餐厅信息：城市、大众点评网址、商铺名称、综合评分、商圈、食品类别、口位评

分、环境评分、服务评分、人均、详细地址。

用户评论：店铺名称、店铺网址、用户名、用户 ID 链接、评定星级、评论描述。

系统用户：邮箱、用户 ID、密码。

8、 基于 Web 引导页的设计

1) 显示热门城市及其城市味觉。

2) 显示推荐餐厅等。

9、 基于 Web 用户界面的设计

1) 用户注册界面。

2) 新用户饮食喜好问卷等。

3) 用户登录界面。

4) 用户可以编辑个人信息，查看收藏夹。

[预期目标]

整个系统可以按系统总框架流程图顺利运行，各模块能够积极响应，Web 客户端运行能够连接服务器并正常进行操作。在框架搭建正确并可实际运行后逐渐完善各模块功能直至达到覆盖基本功能。

四、对进度的具体安排

第 1-3 周：明确毕业设计阶段要完成的任务，查阅资料，完成任务书及开题报告。

第 4-5 周：完成数据收集，进行数据采集和数据清洗。

第 6-7 周：完成数据分析。

第 8-10 周：初步完成推荐餐厅推荐、数据可视化等功能。

第 11-13 周：完善推荐功能，并且调试整个程序。

第 14-15 周：完善程序，修改并提交毕业论文。

第 16 周：准备毕业答辩及相关文档的归档工作，完成毕业答辩。

五、参考文献

[1].王佳安,王可心.李直旭.美食精准搜索与智能推荐平台的设计与实现[J].福建电脑,2019(08).

[2].袁丁,章剑林.吴广建.基于方面级的餐厅用户评论细粒度情感分析[J].软件. 2019(08).

[3].刘伟,陈春林..基于注意模型深度学习的文本情感倾向性研究[A].第 19 届中国系统仿真技术及其应用学术年会论文集（19th CCSSTA 2018）[C].2018 年.

[4].崔垚,融合用户情境及特征信息的餐厅推荐系统设计与实现[D].北京邮电大学.2017 年.

[5].王嘉菲,朱志锋.基于协同过滤算法的视频智能推荐系统[J].湖北大学学报(自然科学版),2019(02).

[6].李吉祺,黄刚.提取关键字改进协同过滤算法的研究与应用[J/OL].计算机技术与发展,2019(06) .

[7].邹红旭,潘冠华,李吟.基于 Spark 框架的改进协同过滤算法[J/OL].计算机技术与发展,2020(05).

[8].黎曦. 基于网络爬虫的论坛数据分析系统的设计与实现[D].华中科技大学,2019.

[9].张晓阳,秦贵和,邹密,孙铭会,高庆洋.基于LDA模型的餐厅推荐方法研究[J].计算机科学,2017,44(07)

[10]. 徐林.基于 Spark MLlib 协同过滤算法的美食推荐系统研究[J].吉林大学学报(信息科学版),2019,37(02).

[11]. Qingyao Ai, Xuanhui Wang, Sebastian Bruch, Nadav Golbandi, Michael Bendersky, and Marc Najork. Learning Groupwise Multivariate Scoring Functions Using Deep Neural Networks. The 2019 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '19). 2019.

[12]. Jalilifard A, Caridá, Vinicius, Mansano A, et al. Semantic Sensitive TF-IDF to Determine Word Relevance in Documents[J]. 2020.

[13]. Shuai Zhang, Lina Yao, Aixin Sun, Yi Tay. Deep learning based recommender system: A survey and new perspectives. ACM Computing. Surveys, 2018.

[14]. Travis Ebesu, Bin Shen, Yi Fang. Collaborative Memory Network for Recommendation Systems.ACM SIGIR,2018.

[15]. Yonghong Tian,Bing Zheng,Yanfang Wang,Yue Zhang,Qi Wu. College Library Personalized Recommendation System Based on Hybrid Recommendation Algorithm[J]. Procedia CIRP,2019,83.

[16]. B. Kupisz and O. Unold,.Collaborative filtering recommendation algorithm based on Hadoop and Spark. 2015 IEEE International Conference on Industrial Technology (ICIT), Seville, 2015.

指导教师:(签署意见并签字) 年 月 日

督导教师:(签署意见并签字) 年 月 日

领导小组审查意见:

审查人签字: 年 月 日