哈爾濱工業大學

毕业设计（论文）中期报告

题 目：汽车之家虚假评论信息的甄别

专　业 计算机科学与技术

学　生 魏鸿焱

学　号 **1120310506**

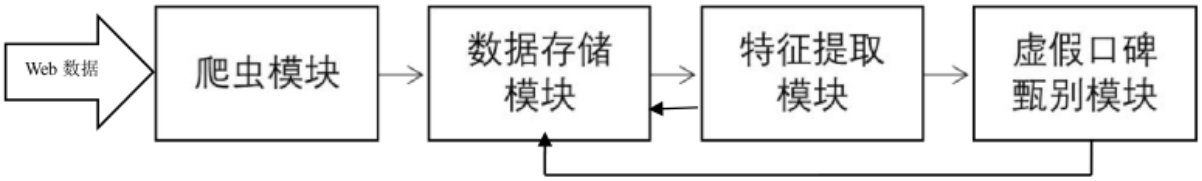
指导教师 刘旭东

日　期 **2016** 年 **4** 月 **22** 日

哈尔滨工业大学教务处制

1．论文工作是否按开题报告预定的内容及进度安排进行

虚假评论甄别大致需要如下几个模块来完成不同的工作：



开题报告项目进度安排表

3-5 周	学习爬虫、数据库、机器学习相关知识；分析汽车之家网站帖子格式内容给出数据库模型。
6-7 周	开发并实现爬虫模块与数据存储模块，尽可能抓取可观数量的帖子评论保存于数据库中。要求爬虫运行速度快并且抓取结果准确,数据库高可用。
8-10 周	开发特征提取模块，将数据存储模块中的口碑数据进行分词，提取文本特征。
11-12 周	进一步学习机器学习相关知识，选择适当的分类算法对口碑进行可信性分类，结合特征提取模块进行调试，训练分类器以达到较高的分类准确性。
12-14 周	将数据存储模块中的帖子数据分类完成，得到垃圾评论集合，与评论分类工具。总结实验。

目前已经完成了爬虫模块与数据存储模块的设计与实现。

2．已完成的研究工作及成果

爬虫模块设计思路：

在 linux+pycharm 环境下使用 python 语言进行开发，使用 urllib 工具从要抓取的页面 url 获得 html 文档，使用 Beautiful Soup 解析 html 文档并且从文档中提取出需要的数据，最终将得到的数据插入 mysql 数据库完成持久化。

代码托管地址：https://github.com/weiazm/crawler.git

爬虫代码结构：

Constant.py：存放字典常量如数据库链接配置、请求头信息与 BBSContent 实体类等内容。

HtmlUtil.py：获得 urllib 请求返回的 gzip 格式压缩网页，并且解压缩 gzip，按文档编码方式解码，获得未经处理的 html 文档。

SoupUtil.py：处理模块。使用 Beautiful Soup 解析 html 文档，计算并提取想要的数据，将数据存储到 BBSContent 实体类中。

StringUtil.py：处理字符串模块。去除制表符、回车符号、前后空格；完成汉子与数字之间的转换；构造分页的链接字符串。

SqlUtil.py：持久化工具模块。将实体类插入数据库、更新数据库表项，以及查询数据库。

RunnableGetContent.py：运行主模块。从数据库中获得要处理的内容以及调用其他模块，处理异常、记录日志，并且保证整个爬虫模块能够持续处理。

数据存储模块设计思路：

目前主要的表有四个，表结构依次如下（以下仅展示重要字段）：

car_id_brand

car_id	int(11)	汽车 id
brand	varchar(45)	汽车品牌
category	varchar(45)	汽车类型

汽车品牌表：存放了 22 个车型的汽车 id 、品牌。根据汽车id 来找到其论坛地址，进而获得帖子链接。

22 条数据

forum_links

bbs_id	int(11)	帖子 id
author_uid	int(11)	发帖人 uid
release_time	datetime	发表时间
reply_num	int(11)	帖子回复数量
click_num	int(11)	帖子点击数量
last_reply_time	datetime	最后回复时间
last_reply_uid	int(11)	最后回复用户 uid
title	varchar(45)	标题内容

主帖表：由论坛地址获得的每个帖子链接保存在这个表中。保存着主贴所对应的数据，这个表只有主贴的信息，没有帖子内容。我们可以根据 bbs_id 与页码拼出帖子的链接，为下一步抓取楼层内容做好准备。

1735822 条

bbs_content

bbs_id	int(11)	帖子 id
uid	int(11)	层主用户 id
from_floor	int(11)	层主所在楼层
to_floor	int(11)	层主回复楼层
reply_time	datetime	回复时间
content	varchar(5000)	回复内容
device	varchar(45)	来自设备

帖子内容表：保存了论坛中所有帖子的内容信息。from_floor 为 0 表示帖子第一条的内容。content 中包括接下来要做情感分析的回复内容。来自设备包括网页、安卓、苹果设备、平板等，便于后期做统计数据。

32716800 条

user

uid	int(11)	用户 id
auth	tinyint(4)	是否是认证车主
num_of_follows	int(11)	关注人数
num_of_fans	int(11)	粉丝人数
create_time	datetime	注册时间
num_of_bbs	int(11)	发帖数量
num_of_reply	int(11)	回复数量
first_post_time	datetime	首次发帖时间
last_post_time	datetime	最后发帖时间
avg_num_of_bbs	int(11)	每日帖子数量

用户表：保存了存在于 bbs_content 表中的用户信息。这个表中包括了一些需要统计的数据（无法由爬虫直接获得），需要对数据进行处理后计算出来。其他数据如是否认证车主、关注，粉丝人数以及几个时间信息等数字内容可以用来做分类鉴定是否为 spammer 的参数。

正在运行未统计

数据库可以根据需求来补充新表或字段，一些统计信息随着后期的计算会添加到数据库中。

遇到的困难以及解决办法：

1. 程序运行时间：

考虑到网络因素，若处理完一个仅有两页的帖子需要 1 秒，则 170 万左右的帖子处理完需要连续不间断处理 472 小时，约 20 天。而且公寓夜晚熄灯断电，预估的时间超出了所能承受的最长时间需求。

解决办法：将主贴数据分块，十万条分为一块，每个数据块分一个爬虫程序来处理，18 个爬虫程序在多台电脑上同时运行，最终总耗时约一周。

2. 代码运行意外崩溃：

代码执行过程中可能遇到网络中断、linux 段错误等问题。影响程序继续运行，以及如何将程序从中断的地方继续运行。

解决办法：建立一个 count 表共 18 条数据对应 18 个爬虫。爬虫处理前查询 count 表找到要处理的 id，爬虫处理完对应的 id 的帖子的时候，更新 count 表。写一个 monitor 程序，定时监督数据库，若一定时间内（500 秒）数据库 count 表某一项未被更新则可以断定为该项对应爬虫程序中断，自动执行脚本重启该程序。优化代码结构，提高程序运行的稳定性。

3. mysql 错误恢复：

当爬虫程序在执行过程中关机可能会导致 mysql 数据库 innodb 表损坏，mysql-service 无法启动。

解决办法：搜索解决办法，发现错误日志里面提示出现了坏页，导致数据库崩溃。配置 mysql 忽略检查到的坏页并重启数据库修复索引。定期做好数据库备份。

4. 帖子页面错误：

网站内容格式会更新，14 年的帖子与 15 年的帖子格式会有部分不同。而且部分帖子会被网站管理员删除。

解决办法：修改程序，增强兼容性。对于已经删除的帖子，完善异常机制，记录日志并且忽略。

3．后期拟完成的研究工作及进度安排

9-10 周开发特征提取模块，对文本进行分词以及对数据进行处理，提取文本的情感倾向、语言模糊性等特征保存在数据库中。以及使用数据挖掘的分类技术对帖子性质进行分类，如游记、问题、调查、提车、评论等。

11-12 周进一步学学习机器学习相关内容，设计并实现分类器模块，提供样本数据进行训练以及进行分类处理。

12-14 周完成虚假垃圾评论甄别工具，总结项目。

4．存在的问题与困难

目前存在的问题：

user 表内容还不完善，抓取内容还需要一定的时间来运行程序。

困难：

考虑到数据量比较大，处理时间可能过长，未必需要将全部数据处理完成可以选择特定车型如该车论坛比较活跃、帖子数量可观，或满足一定条件的数据如回帖人数大于一定数量的帖子针对性地进行处理，从而得到就具有代表性的结果。

5．论文按时完成的可能性

目前数据可以说已经准备完成，后期便只剩下处理数据的工作。按照进度安排可以按时完成。