哈爾濱工業大學

毕业设计（论文）开题报告

题 目： 汽车之家虚假口碑信息的甄别

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 3 月 10 日

哈尔滨工业大学教务处制

1．课题来源及研究的目的和意义

1.1 课题来源

评论，即用户针对事物进行主观或客观的印象阐述，能够帮助人们去了解某一事物。如今在互联网上的各种信息分享、传播及获取平台上均有海量的评论信息。互联网用户由简单的信息接收者，发展成为信息的发布、监督者。一方面，平台信息的共享性、实时性、互动性以及传播方式的多样性深刻地影响了人们的生活方式，极大地提高了获取信息的效率。另一方面，评论信息的急剧增长与信息的快速传播也引发了许多不容忽视的问题。其中包含主观与客观成分的评论信息的可信度问题便是其中亟待解决的问题。如何快速地选择出有价值、可信度高的评论，判断并识别虚假评论等，已成为个人乃至企业密切关注的问题。

1.2 研究的目的

随着互联网行业的发展，用户在互联网上的作用也越来越重要，互联网也在影响着消费者的消费观念，越来越多的消费者倾向于网上购物。就电子商务网站的某一产品的评论而言，过度褒奖的评论能够引导用户对其产生较好的印象，进而加强其对该产品的消费欲望，而过度贬低的评论会令用户无法真正了解到产品的特征性质，进而对该产品的营销产生不良影响。由此涉及到的利益问题会令部分商家为了达到自己的商业目的雇佣“水军”对自家或他家产品做出吹捧或者诋毁的评论来误导用户对该产品的看法，混淆视听，从而误导潜在消费者。虚假评论甄别旨在解决这一问题，将垃圾评论从评论文本中找到并排除，保留真实的评论，为用户提供可靠的参考内容，同时也为构建和谐互联网贡献力量。

1.3 研究的意义

口碑的意义在于其蕴含着巨大的商业价值导向，是信息时代消费者购买前进行决策的重要依据。虽然在互联网平台中，大部分的口碑信息是真实可靠的，但是依然有少部分商家进过幕后操作产生的虚假口碑。互联网上的信息量巨大且内容混杂，若用户针对某一个产品的一条评论人工地进行判断则需要通读全部评论，对其把握好产品评论大致基调，再对该条评论的可信性做定夺。这种方法必然费时费力，且手动操作亦有发生错误的可能。而如今处于互联网时代的我们，研究并且设计一个能够快速并准确地对商品的评论进行可信度分类的工具将具备很大的实用价值。研究的意义在于虚假口碑甄别工具排除了虚假的评论内容，节省了用户对评论内容的可信性做出判断的时间，降低了其思考代价。从使用网站的消费者角度讲，甄别出虚假口碑信息能使消费者用户获得更可靠的参考信息；从出售商品的商家角度讲，真实的口碑评论能够帮助商家用户及时地了解市场现状，并且准确地获得消费者的反馈意见，以便其应对市场做出更好的行动；从维护网站的管理员角度讲，排除虚假的评论信息能够帮助其建立一个更公正透明的网络平台，维护良好的网络环境，以便吸引更多用户使用。在另一个层次上，对评论可信度的研究亦会反映出一些对发出评论的用户的可信度的看法，在必要的情况下网站管理员可以对该类用户做出处理措施，如

警告、删除评论、禁用账号等。

2．国内外在该方向的研究现状及分析

近年来，随着电子商务的迅猛发展，评论的可信度计算问题吸引了来自经济、社会和计算机等多领域学者的关注。研究的方向一般是 web 信息的可信度计算，现有研究对象有 deep web 数据、博客 、wikipedia 文章、新闻及电影以及电子商务平台上的评论等。

如何计算评论的可信度一般有以下几个研究思路：

1）基于评分的可信度计算，即依据用户评分来计算 web 内容的可信度；

2）基于口碑的可信度评估，通过识别用户的恶意口碑来得到可信度高的评论；

3）基于评论数据中语义相关度的可信度计算；

4）基于历史数据的可信度估算等。

有国外学者基于评论的语义如评论的信息量（information）、可读性（readability）、和主观性（subjectiveness）这三个方面的特征，使用支持向量机法建立机器学习模型来区分评论的质量。另有 Kim S-M 等人基于评分来评估评论的质量。如 Amazon.com 中“3204 of 3272 people found the following review helpful”表示有 3272 位读者对该评论做出了评价，其中 3204 位用户认为该评论有用。Ghose A 等人基于历史数据，如可信度高的用户做出的评论可信度会相对高一些，建立了以消费者为导向和以商家为导向的排名机制。Myle Ott 等人结合了心理学与计算机语言学，研究并比较了三种检测虚假评论的方法，并最终得到了一个能满足其 90%的虚假评论样本的分类器，除此之外他们还进行了几个理论贡献，包括发现了虚假评论与富于想象力的文学作品之间的关系。

国内的学者孙升芸等将产品评论和互联网上其他常见的垃圾信息进行对比，并把垃圾评论的检测、质量判断和情感分析等相关工作进行了比较分析。然后从产品垃圾评论检测的数据集、检测方法两个角度对相关工作进行了概述和分析。在另一份工作中，唐晓婷等人提出了一种基于用户的特征分析的评论可信度计算算法。该算法首先根据语义特征，对历史评论者进行用户社区挖掘，得到在某种准确度下评论过某对象的用户公共特征，形成用户模板；其次，对任意给定新评论，通过其评论者和用户公共特征模板的匹配程度，并综合该评论者可信度、评论的语义相关性等计算该评论的可信度。

也有学者对微博进行研究。如当可信度低的微博被转发时一般会保留原始作者的观点；当可信度低的微博信息转发时通常附加上转发者的观点。由此可以通过计算转发微博相比原始微博的保留率来评估微博信息的可信性。在这方面国外学者更倾向于研究 Twitter 用户发布微博的行为特征和信息传播中的特点。他们发现谣言大多集中在话题相关的讨论区中，可以基于上下文信息如文本语义、转发、关注、评论的数量来计算搜索内容的可信度。

Bing Liu 在虚假口碑甄别方面的工作具有一定的代表性。他使用了有监督学习、模式挖掘、基于图形的方法以及关系建模来解决这个问题。下面是其研究该问题时的一些主要关注点：

1）评论内容：包括词汇特征、不同评价者评价内容和风格的相似性、语义上的不一致

2）评论者不正常的行为：网站上的公开数据（评论者 id、评论时间、评论频率、产品的初次评论者们等）、网站上的隐藏数据（IP 地址、MAC 地址、发表评论所需时间、评论者的真实位置等)。

3）产品相关特征：产品描述、产品销售数量、销售排名。

4）其他关系：评论者、评论、实体之间的复杂关系。

Bing Liu 同时也指出了这项研究面临的一些挑战。如人工地发现虚假评论是非常困难的，尤其对于那些写得非常好的虚假评论。人工标注虚假评论非常困难，这就导致难以获得 Gold Standard 的数据集。

3．主要研究内容

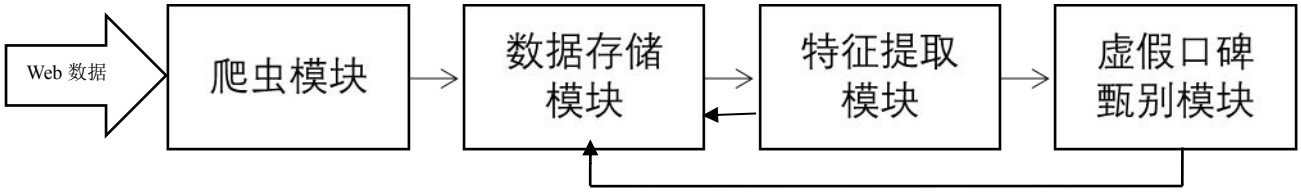
互联网上的虚假事实陈述严重影响人们有效地获取信息。本题目的主要研究内容是甄别出汽车之家网站中的虚假口碑信息。从评论对象的一些历史口碑中使用机器学习方法建立一个分类模型，待训练结果达到满意程度后，以此作为标准来对全部口碑进行分类，最终得到两类口碑集合，即真实的口碑集与虚假的口碑集。

选择汽车之家网站作为计算评论可信度的平台是因为该网站上的口碑评论内容格式整齐，便于处理，可以将 web 信息提取到关系型数据库中规范其形式，便于针对性地对特定商品或特定用户的研究。而且该网站在汽车社区类网站中在信息规模与用户数量上具有代表性，信息更新速度快，汽车类型与入驻商家也非常丰富。在该平台上进行的研究有一定的实用价值。

4. 研究方案

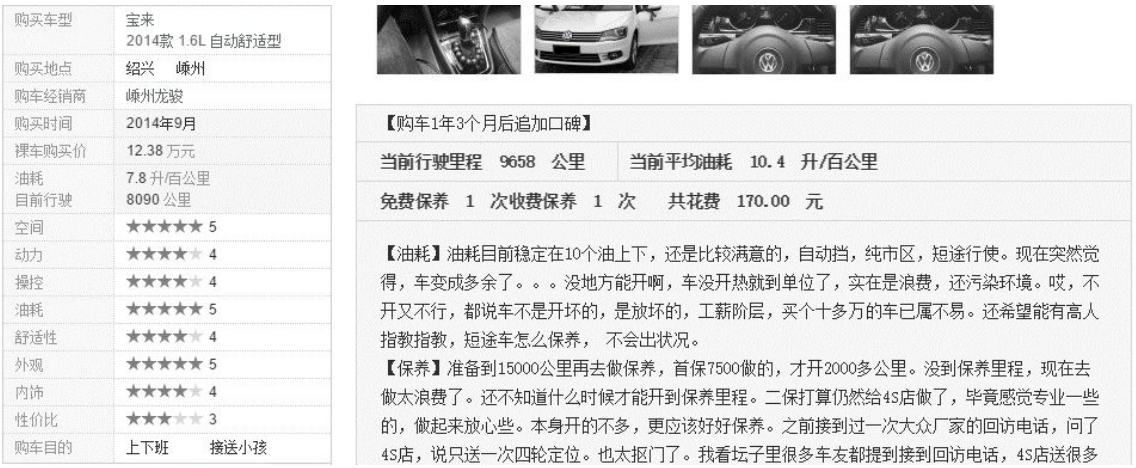
虚假口碑甄别大致需要如下几个模块来完成不同的工作：

爬虫模块、数据存储模块、特征提取模块、虚假口碑甄别模块



每个模块的工作大致如下：

1） 爬虫模块使用 python 来实现，将网页上的口碑信息抓取下来并保存到数据存储模块。汽车之家网站上的一条口碑信息如下：



购买车型	宝来 2014款 1.6L 自动舒适型
购买地点	绍兴　嵊州
购车经销商	嵊州龙骏
购买时间	2014年9月
裸车购买价	12.38 万元
油耗 目前行驶	7.8 升/百公里 8090 公里
空间	★★★★★ 5
动力	★★★★☆ 4
操控	★★★★☆ 4
油耗	★★★★★ 5
舒适性	★★★★☆ 4
外观	★★★★★ 5
内饰	★★★★☆ 4
性价比	★★★☆☆ 3
购车目的	上下班　　接送小孩

【购车1年3个月后追加口碑】

当前行驶里程　9658　公里　　当前平均油耗　10.4　升/百公里

免费保养　1　次收费保养　1　次　　共花费　170.00　元

【油耗】油耗目前稳定在10个油上下，还是比较满意的，自动挡，纯市区，短途行使。现在突然觉得，车变成多余了。。。没地方能开啊，车没开热就到单位了，实在是浪费，还污染环境。哎，不开又不行，都说车不是开坏的，是放坏的，工薪阶层，买个十多万的车已属不易。还希望能有高人指教指教，短途车怎么保养，　不会出状况。

【保养】准备到15000公里再去做保养，首保7500做的，才开2000多公里。没到保养里程，现在去做太浪费了。还不知道什么时候才能开到保养里程。二保打算仍然给4S店做了，毕竟感觉专业一些的，做起来放心些。本身开的不多，更应该好好保养。之前接到过一次大众厂家的回访电话，问了4S店，说只送一次四轮定位。也太抠门了。我看坛子里很多车友都提到接到回访电话，4S店送很多

口碑信息中包括了几个维度：<油耗>、<保养>、<故障>、<空间>、<动力>、<舒适性>、<外观>、<内饰>等，不同维度的口碑包括了评分与用户填写的内容。爬虫模块可以使用 Beautiful Soup 工具来抓取网页数据。

2） 数据存储模块即数据库。我们选择使用 mysql 数据库来实现。首先依据汽车之家网站上的口碑的格式设计出对应的关系型数据库模型，并且依据该模型建立数据库，要求做到将口碑整体内容详细并准确地进行分解，尽可能还原该网站的原本数据模型。

3） 特征提取模块同样使用 python 来实现，首先使用 jieba 分词工具将口碑中的句子精确地切开或将文本中的词语全部扫描出来。将文本进行分词后文本就是一个由每个词组成的长数组：[word1, word2, word3…… wordn]。之后就可以使用 nltk 里面的各种方法来处理这个文本。比如用 FreqDist 统计文本词频，用 bigrams 把文本变成双词组的形式：[(word1, word2), (word2, word3), (word3, word4)……(wordn-1, wordn)]。然后可以用这些来计算文本词语的信息熵、互信息以及评论者的心态、评论内容的模糊度、评论内容的相似性等信息。

4） 最后是虚假口碑甄别模块。该模块使用机器学习工具 scikit-learn 来对文本进行分类，该工具提供了丰富的分类算法。我们可以用特征提取模块得到的结果作为选择机器学习的特征，选择适当的分类算法构建分类器，手动地对其进行训练与调试，最终做出对文本的分类。分类算法可以使用 Naïve Bayes、KNN 或 SVM 等。

爬虫模块将网页上的口碑内容抓取下来并且根据其维度将数据保存到数据存储模块中，特征提取模块从数据存储模块中获得数据再将文本特征保存到数据存储模块中，虚假口碑甄别模块其本质是个分类器，将口碑的文本特征作为输入参数，最终将分类结果保存到数据库中。由于数据存储模块由于与其他模块之间的调用非常频繁，故作为核心模块需要其有良好的性能与高可靠性，对该模块的设计与实现将有较高的要求。

5．进度安排，预期达到的目标

3-5 周	学习爬虫、数据库、机器学习相关知识；分析汽车之家网站口碑格式内容，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口碑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口碑数据分类完成，得到虚假口碑集合。总结实验。

6．课题已具备和所需的条件、经费

目前该课题已经具备所需要的硬件设施，研究中所需的编程部分将使用 python 语言来实现，在数据挖掘知识方面需要一定的学习代价，目前暂无其他所需条件。

7．研究过程中可能遇到的困难和问题，解决的措施

可能遇到的问题：

1）数据挖掘的分类算法有很多，研究过程中在算法的选择上可能会有一定困惑。并且算法的实现可能会很复杂或耗费时间。

2）互联网平台上的口碑数据量巨大，由于在时间与空间方面的限制无法将数据完全地下载到本地进行处理。

3）由于电脑的不稳定，代码运行过程中可能会宕机。

解决的措施：

1）仔细学习该部分的知识，找到适用于该数据模型的算法，在理解算法原理后尽量使用开源工具避免造轮子。

2）选择出具有一定规模的有代表性的口碑数据进行处理。

3）考虑容错机制，即如何从宕机中恢复继续处理。并且提高电脑的稳定性。

8．主要参考文献

[1] 孙升芸, 田萱. 产品垃圾评论检测研究综述[J]. 计算机科学, 2011(B10):198-201.

[2] 唐晓婷,吴爱华,曾卫明. 不一致数据库中基于用户语义模板的评论可信度计算. 燕山大学学报.2014.11.38(6):523-543.

[3] 蒋盛益,陈东沂,庞观松. 微博信息可信度分析研究综述. 图书情报工作.2013.6.57(12):136-142.

[4] 周中华,张惠然,谢江. 基于 Python 的新浪微博数据爬虫. 计算机应用.2014.11.34(11):3131-3134.

[5] 李秀娟,田川,冯欣. 数据挖掘分类技术研究与分析. 现代电子技

术.2010.10.(331):86-88.

[6] 林闯,田立勤,王元卓. 可信网络中用户行为可信的研究. 计算机研究与发展.2008.12.45(12):2033-2043.

[7] 沈昌祥,张焕国,王怀民. 可信计算的研究与发展. 中国科学.2010.01.40(2):139-166.

[8] 杨定中,赵刚,王泰. 网络爬虫在 Web 信息搜索与数据挖掘中应用. 计算机工程与设计.2009.12.30(24):5658-5662.

[9] 李玲俐. 数据挖掘中分类算法综述. 重庆师范大学学报. 2011.07.28(4).44-47.

[10] Ott M, Choi Y, Cardie C, et al. Finding deceptive opinion spam by any stretch of the imagination[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. Association for Computational Linguistics, 2011:309-319.

[11] Y Suzuki. A Credibility Assessment for Message Streams on Microblogs. International Conference on P2p, Parallel, Grid, Cloud, & Internet Computing.2010:527-530.

[12] Cho J, Kwon K, Park Y. Q-rater: A collaborative reputation system based on source credibility theory [J]. Expert Systems with Applications. 2009,36 (2): 3751-3760.

[13] Yang K C C. Factors influencing Internet users' perceived credibility of news-related blogs in Taiwan [J]. Telematics and Informatics. 2007,24 (2): 69-85.