Hans Journal of Data Mining 数据挖掘, 2019, 9(3), 88-95
Published Online July 2019 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2019.93011

Hans汉斯

Factors Influencing Passenger Satisfaction Based on Text Data Mining

Yali Jiang, Changjun Li

School of Mathematical Sciences, Ocean University of China, Qingdao Shandong
Email: 1451621830@qq.com

Received: Jun. 26th, 2019; accepted: Jul. 9th, 2019; published: Jul. 16th, 2019

Abstract

With the rapid development of the Internet industry and the aviation industry, more and more people tend to buy tickets on the website, and many passengers will comment on the flight after taking the flight. Based on textual data, this paper studies the characteristics of airline services that affect passenger satisfaction, so as to help airlines improve corresponding services and enhance passenger flight experience. Firstly, this paper uses python crawler technology to crawl the comment data of passengers of China Eastern airlines on CAPSE website. Secondly, high frequency words in comment text are counted. Then the LDA theme model method is applied to obtain the theme keywords, and the service characteristics concerned by passengers are mined from the perspective of users. Then TF-IDF method is used to transform text comments into a word vector matrix based on service characteristics. Finally, through correlation coefficient method and feature importance analysis method based on decision tree, it is found that the key factors affecting passenger satisfaction in airline service are whether the plane is on time, flight attendant service level, cabin environment and so on.

Keywords

Text Mining, Online Reviews, Air Services, Influence Factor

基于文本数据挖掘影响乘客满意度的因素

蒋亚丽，李长军

中国海洋大学数学科学学院，山东 青岛
Email: 1451621830@qq.com

收稿日期：2019年6月26日；录用日期：2019年7月9日；发布日期：2019年7月16日

文章引用：蒋亚丽，李长军．基于文本数据挖掘影响乘客满意度的因素[J]．数据挖掘, 2019, 9(3): 88-95.
DOI: 10.12677/hjdm.2019.93011

摘　要

随着互联网行业和航空行业的高速发展，越来越多的人倾向在网站上购买机票，许多乘客会在乘坐之后对航班进行评论。本文基于文本数据研究影响乘客满意度的航司服务特征，帮助航空公司进行相应服务的改善，提升乘客航程体验。本文利用python爬虫技术爬取CAPSE网站东方航空公司乘客的评论数据，首先对数据进行预处理；其次统计评论文本高频词汇；再应用LDA主题模型方法获取主题关键字，从用户角度挖掘乘客关注的服务特征；然后利用TF-IDF方法将文本评论转化为基于服务特征的词向量矩阵。最后通过相关系数法和基于决策树的特征重要性分析方法，发现航空公司服务中影响乘客满意的关键因素是飞机是否准时、空乘服务水平、客舱环境等问题。

关键词

文本挖掘，在线评论，航空服务，影响因素

1. 引言

随着互联网行业的不断发展和壮大，全球电子商务也在不断的发展。截至 2018 年 12 月，我国网民规模达 8.29 亿，普及率达 59.6%，较 2017 年底提升 3.8 个百分点，全年新增网民 5653 万。截至 2018 年 12 月，我国网络购物用户规模达 6.10 亿，年增长率为 14.4%，网民使用率为 73.6% [1]。在经历多年高速发展后，网络消费市场逐步进入提质升级的发展阶段，供需两端“双升级”正成为行业增长新一轮驱动力。

纵观目前国内航空公司所处的内外竞争局势，可见我国航空公司正面临来自多方的挑战。近年来多个民营航空公司快速壮大，国内航空公司之间低层次竞争加剧、价格体系紊乱。航空公司要想谋求长久的经济利益，必须以不断提升客户满意度为导向，努力提高服务质量和水平。同时，航空服务在商品交易中的地位日益重要，而且有愈演愈烈之势。这意味着航空公司要想与顾客建立一种长期的关系，就必须取悦它的顾客，建立顾客忠诚度[2]。

随着互联网行业的普及化，文本评论数量大幅增长，评论内容无疑会影响着其他乘客对航空公司的选择。在线评论一般具有信息复杂、信息量巨大等特点，如何在海量的用户评论数据中挖掘出有用的信息，了解乘客对于航空公司各项服务的意见和建议，改善乘客对航空公司感官服务不好的想法，是航空公司需要考虑的重大问题之一。本文以中国东方航空公司的乘客在线评论数据为例，进行文本挖掘分析。

2. 数据获取

2.1. 数据来源

数据来源于 CAPSE 民航旅客服务测评中的旅客声音板块。该网站作为独立的第三方民航服务测评网站，以倾听旅客声音，提升服务质量为目标，专注提供民航服务数据咨询及服务解决方案。

论文首先通过 Python 网络爬虫技术获取 CAPSE 网站的中国东方航空的乘客的打分数据和评论数据，共 7562 条。在线评论是影响乘客选择航空公司的重要原因之一，分析乘客在线评论数据可以从中获得乘

客对航空公司的内心想法和感受，从用户体验方面不断完善航空公司的各项服务，提升客户满意度。

2.2. 数据预处理

在对文本数据进行分析之前，需要将评论语料集转化为标准格式，剔除数据噪声。对于在线评论语料中的数字及特殊符号，采用统一转换成易识别的符号或空格，然后依据停用词词典将停用词从语料中清除，预处理具体包括以下步骤。

1) 去噪声数据

由于文本评论的随意性，在文本中可能会出现很多问题，中英文标点混用，在情感表述时添加的各种表情符号，这些噪声数据在整个文本评论挖掘过程中作用不大，要对这些数据进行清洗去除。对文本评论中的表情符号，总结归纳出标点符号的含义并将表情替换为相应的文本。当文本中有英文标点时可能对之后的处理过程有一定的影响，将文本中的英文标点转化为中文标点。

2) 去停用词

停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理文本之前会自动过滤掉某些字或词。基于在线评论数据人工定义停用词字典，去除一些如“的”，“了”等无意义词汇。

3) 分词

本文采用 Python 语言中的结巴分词包进行文本分析，采用精确模式进行中文文本分词与词性标注。精确模式试图对句子进行最精确的切分，适合做文本分析。

4) 文本划分

以乘客的打分数值为依据，1 分和 2 分划分为负面评价，4 分和 5 分划分为负面评价，划分后正面数据共 6316 条，负面数据共 772 条。根据正负情感的评价比例也可以看出，中国东方航空公司给人的整体感觉较好，正面情感评分占比接近 90%。

3. 在线评论文本挖掘

3.1. 词云分布

在对语义进行分析时，提取评论中的高频词是非常有必要的，把握高频语义词对理解文本内容和客户在乘坐飞机过程中的重点关注服务非常重要，但是个性化的词汇不具有代表性[3] [4]。对于不同情感倾向的乘客评论，会有不同的高频词汇，正面情感中的高频词汇体现了乘客对航空公司比较满意的服务特征。反之，负面情感中的高频词汇体现了乘客不满的服务特征，航空公司应致力于维持自己的优势，完善自己的不足。

利用 python 工具，统计预处理后评论词汇的词频，按照词频降序排序，去掉热词中的单字，分别绘制词频排在前 50 热词的词频图，正面和负面倾向评论词云图如图 1 和图 2。

Figure 1. Positive evaluation word cloud
图 1. 正面评价词云

Figure 2. Negative evaluation word cloud
图 2. 负面评价词云

观察词云图发现，在正面评价中，出现较多的服务特征有空乘、服务、准时、值机、安检、机场、餐食、座椅，可以说明这几项服务是乘客在整个航程体验中比较满意的。此外，正面评论中出现较多的评价性词汇有准时抵达、快捷、舒适、好吃，都是正面体验词汇。在负面评价中，出现较多的服务特征有延误、机场、客舱、行李、餐食、卫生、摆渡，说明这几项服务是导致乘客对航空公司留有不好印象的原因，与之相关的体验性词汇是难吃、拥挤、不好、垃圾等负面词汇。

3.2. 基于 LDA 评论分析

LDA 主题模型在机器学习和自然语言处理等领域是用来在一系列文档中发现抽象主题的一种统计模型[5]。运用主题模型，能够发现文本中使用词语的规律，并且把规律相似的文本联系到一起，以寻求非结构化文本集中的有用信息。对于航空公司的评论，代表航空公司的词语如“空乘”、“服务”、“餐食”等会频繁地出现在评论中，运用主题模型，将航空公司的服务特征词汇与评论中情感描述性词汇联系起来，从而深入了解乘客对航空公司评价的聚焦点及乘客对某一特征的情感倾向。

LDA 模型认为每篇文档的每个词都是通过“一定的概率选择了某个主题，并从这个主题中以一定的概率选择了某个词语”。LDA 模型也被称为 3 层贝叶斯概率模型，包含文档(d)、主题(z)和词(w)3 层结构，能够有效地对文本进行建模[6] [7]。通过 LDA 主题模型，能够挖掘数据集中的潜在主题，进而分析数据集的集中关注点和其相关特征词。

虽然 LDA 可以直接对文本做主题分析，但是文本的正面评价和负面评价混淆在一起，并且由于分词粒度的影响(否定词、程度词等)，可能在一个主题下生成一些令人迷惑的词语。因此，将文本分为正面评价和负面评价两个文本，再分别进行 LDA 主题分析。

经过 LDA 主题分析后，评论文本被聚成 5 个主题，每个主题下生成 5 个最有可能出现的词语以及相应的概率，表 1 显示了乘客正面评论文本中的潜在主题，表 2 显示了乘客负面评论文本中的潜在主题。

Table 1. Airlines positive evaluation potential themes
表 1. 航空公司正面评价潜在主题

主题 1	主题 2	主题 3	主题 4	主题 5
好	抵达	餐食	飞机	好
空乘	准时	好吃	延误	服务
服务	安检	服务	到达	餐食
抵达	值机	好吃	提前	空乘
准时	快捷	空乘	起飞	安检

Table 2. Airlines negative evaluation potential themes
表 2. 航空公司负面评价潜在主题

主题 1	主题 2	主题 3	主题 4	主题 5
延误	机场	延误	延误	延误
飞机	东航	差	飞机	难吃
小时	慢	空乘	机场	餐食
航班	安检	态度	小时	差
餐食	小时	小时	航班	小时

根据对航空公司正面评价的 5 个潜在主题的特征词提取，主题 1 主要反映了空乘服务好，主题 2 主要反映了飞机准时抵达和安检值机快捷，主题 3 主要反映了飞机餐食好吃，主题 4 主要反映了飞机提前起飞延误到达，主题 5 主要反映了餐食和服务好。

根据对航空公司负面评价的 5 个潜在主题的特征词提取，主题 1 主要反映了飞机航班延误，主题 2 主要反映了机场安检慢，主题 3 主要反映了空乘态度差，主题 4 主要反映了航班延误，和主题 1 类似，主题 5 主要反映了餐食难吃。在各个主题中，延误多次作为高频词出现，说明在负面评价中，延误无疑是重要的一个原因。

3.3. 影响因素探究

3.3.1. TF-IDF 模型

TF-IDF 是一种用于统计字词对于一个文件的重要度的计算方法。在使用 TF-IDF 方法时需要一个语料库(corpus)，用来模拟语言的使用环境。

词频(TF) = 某个词在文章中的出现次数/文章总词数

逆文档频率(IDF) = log (语料库的文档总数/包含该词的文档总数 + 1)

TF-IDF = 词频(TF)*逆文档频率(IDF)

在 TF-IDF 模型当中，如果某个词出现的频率较高，并且在整个文本中出现的频率较低，那么该词汇就会有较大的 TF-IDF 值。对现在的文本评论内容，采用 TF-IDF 权重构建文档词条矩阵，将文本评论转为词向量矩阵[8]。

由于分词后的词汇数量太多，选取了高频词来做分析。根据词频分析结果，选取“服务、空乘、准时、安检、值机、餐食、座椅、机场、延误、客舱、行李、摆渡”12 个高频词，将 7562 条文本评论转化为 7562*12 维的矩阵向量，通过 Python 编程计算得到词向量矩阵，表 3 展示前 3 条评论的向量矩阵。

Table 3. Section comments on the vector matrix of text
表 3. 部分评论文本的向量矩阵

Contents	服务	空乘	准时	安检	值机	餐食	座椅	机场	延误	客舱	行李	摆渡	star
安检很快，餐食好吃，准时抵达，空乘服务好	0.30	0.31	0.32	0.35	0	0.39	0	0	0	0	0	0	5
空乘人员不通知乘客吃饭时收小桌板，感觉服务不到位的事情还有很多	0.08	0.08	0	0	0	0	0	0	0	0	0	0	5
安检很快，餐食好吃	0	0	0	0.46	0	0.51	0	0	0	0	0	0	5

3.3.2. 相关系数法

相关系数是是研究变量之间线性相关程度的量，由于研究对象的不同，相关系数有多种计算方式，本文计算皮尔逊相关系数，其中皮尔逊相关系数的计算公式如下：

$$r=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}}$$

对转化后的数值向量，计算各个服务特征和满意度评分(star)的相关系数，观察各个特征和对总分的影响性，计算相关系数矩阵如表 4。根据相关系数表可知，延误、客舱、行李、摆渡和评分之间呈现负相关关系，这些关键词的 TF-IDF 值越大，满意度评分(star)越低，其他特征和总分之间呈现正相关关系。其中，对总分影响最大的特征是延误、服务、准时、空乘，说明了航班的准时与否对乘客的总体航程体验满意度影响最大。

Table 4. Correlation coefficient matrix
表 4. 相关系数矩阵

	服务	空乘	准时	安检	值机	餐食	座椅	机场	延误	客舱	行李	摆渡	star
服务	1.00	0.97	0.10	0.12	0.15	0.03	0.14	0.04	−0.21	−0.12	−0.09	−0.07	0.35
空乘	0.97	1.00	0.10	0.12	0.15	0.03	0.14	0.04	−0.20	−0.11	−0.09	−0.08	0.34
准时	0.10	0.10	1.00	0.09	0.14	−0.01	0.07	0.02	−0.21	−0.11	−0.08	−0.08	0.35
安检	0.12	0.12	0.09	1.00	0.22	0.07	0.14	0.14	−0.16	−0.08	−0.07	−0.06	0.26
值机	0.15	0.15	0.14	0.22	1.00	0.06	0.17	0.16	−0.18	−0.10	−0.08	−0.07	0.30
餐食	0.03	0.03	−0.01	0.07	0.06	1.00	0.15	0.17	−0.08	−0.01	−0.06	−0.05	0.11
座椅	0.14	0.14	0.07	0.14	0.17	0.15	1.00	0.18	−0.15	−0.09	−0.07	−0.06	0.26
机场	0.04	0.04	0.02	0.14	0.16	0.17	0.18	1.00	−0.10	−0.05	−0.04	−0.02	0.16
延误	−0.21	−0.20	−0.21	−0.16	−0.18	−0.08	−0.15	−0.10	1.00	0.08	−0.01	0.00	−0.46
客舱	−0.12	−0.11	−0.11	−0.08	−0.10	−0.01	−0.09	−0.05	0.08	1.00	0.01	0.00	−0.22
行李	−0.09	−0.09	−0.08	−0.07	−0.08	−0.06	−0.07	−0.04	−0.01	0.01	1.00	0.06	−0.13
摆渡	−0.07	−0.08	−0.08	−0.06	−0.07	−0.05	−0.06	−0.02	0.00	0.00	0.06	1.00	−0.12
star	0.35	0.34	0.35	0.26	0.30	0.11	0.26	0.16	−0.46	−0.22	−0.13	−0.12	1.00

3.3.3. 决策树法

决策树在特征重要性选择中有很好的效果，使用决策树可以分析出哪个特征在满意度评分中起更大的作用，得出影响乘客满意度的因素，这里利用基尼指数法来进行评价。

将变量重要性评分用 VIM 来表示，将 Gini 指数用 GI 表示，现有 12 个特征 X_1、$X_2\cdots X_{12}$，现在要计算出每个特征X_j的 Gini 指数评分 VIM_j^{Gini}。

Gini 指数的计算公式为

$$GI_m=\sum_{k=1}^{K}\sum_{k'\neq k}P_{mk}P'_{mk}=1-\sum_{k=1}^{K}p_{mk}^2$$

其中，K 表示有 K 个类别，P_{mk} 表示节点 m 中类别 k 所占的比例。直观的说，就是随便从节点 m 中随机抽取两个样本，其类别标志不一致的概率。特征 X_j 在节点 m 的重要性可以表示为加权不纯度的减少

$$VIM_{jm}^{(GIni)} = N_m * GI_m - N_l * GI_l - N_r * GI_r$$

其中，N_m、N_l、N_r分别表示节点 m 左节点 I 和右节点 r 的样本数；GI_l和 GI_r分别表示分枝后两个新节点的 Gini 指数。

这里我们采用 CART 分类树，用基尼指数选择最优特征，同时决定该特征的最优二值切分点。利用 12 个维度的词向量矩阵和分值变量，将数据划分为训练集和测试集，在训练集上训练数据，构建决策树模型，同时利用网格调参获取最优参数，最后在测试集上验证模型的准确率，通过 python 构建决策树，得到的准确率为 78.5%，对一个文本评论，能够准确预测乘客打分分值的概率为 78.5%。

在构建完成的决策树模型之后，对转化后的数值向量利用决策树计算得到特征重要性，绘制重要性分数图形如图 3，查看各个特征对打分的影响性。根据特征重要性分数图形可知，延误对打分的影响性最大，且远远多于其他特征，前四名依次是服务、准时、延误、值机，摆渡、安检和机场对打分的影响性最小，和相关系数得到的结果基本对应。

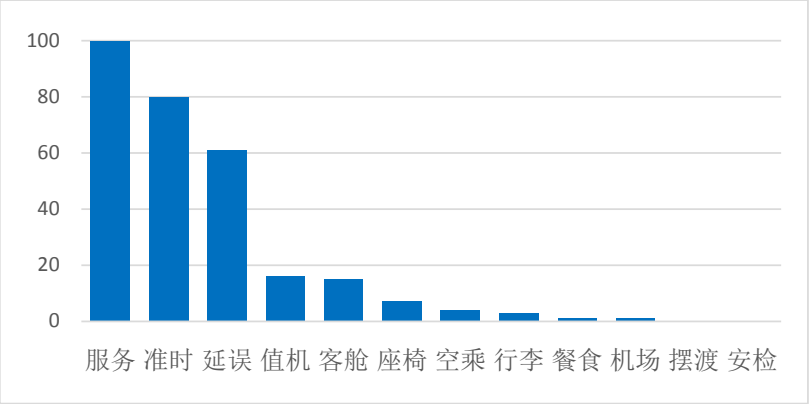

Figure 3. Character importance
图 3. 特征重要性

4. 总结与展望

4.1. 总结

本文在汲取文本挖掘、在线评论与航空公司服务等理论的基础上，以 CAPSE 网站的中国东方航空的在线评论为研究对象，对在线评论文本内容进行特征分类挖掘，建立与检测了基于在线评论文本挖掘的乘客满意度影响因素[9]的分析模型，进一步挖掘了航空公司的服务质量问题。通过本文的一系列研究方法可以得到结论：

1) 在爬取的评论数据中，正面评论文本占接近 90%，说明此航司给人的总体印象非常不错。通过词云图和 LDA 主题模型，对正面评论进行分析可知，让乘客感到满意的内容有空乘服务、准时抵达、值机快捷等，航空公司应该维持这些优势。对负面评论分析可知，让乘客感到不满意的内容有：飞机延误、客舱拥挤、餐食难吃等，航空公司可以适当改进客舱环境和餐食口味。

2) 通过相关系数法和基于决策树的特征分析法可知，航空公司服务中影响乘客满意的关键问题：飞机是否准时、空乘服务水平、客舱环境等问题。

4.2. 展望

本文基于航空领域的进行文本挖掘，取得一定进展，也尚存在很多不足之处。首先，在对文本进行

分词时采用的是现在词汇量比较全面的jieba分词，但是航空领域会有一些专业词汇未纳入分词词汇表中，这会对分词结果产生影响。其次，在将文本转化为词向量时，只是以词频为依据选取关键服务特征，未考虑词汇的其他特性。最后，基于决策时模型计算因素重要性时，对决策树的准确率未进行进一步的提升。这都是以后可以继续努力的方向。

参考文献

[1] 中国互联网络信息中心(CNNIC). 第43次中国互联网络发展状况统计报告[EB/OL]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/P020180305409870339136.pdf, 2018-04-28.

[2] 刘金兰. 顾客满意度与ACSI[M]. 天津: 天津大学出版社, 2006.

[3] 郭立秀. 基于文本挖掘的生鲜电商顾客满意度研究[D]: [硕士学位论文]. 成都: 西南交通大学.

[4] 张振华, 许柏鸣. 基于在线评论文本挖掘的商业竞争情报分析模型构建及应用[J]. 情报科学, 2019, 37(2): 151-155+162.

[5] 王伟, 周咏梅, 阳爱民, 周剑峰, 林江豪. 一种基于LDA主题模型的评论文本情感分类方法[J]. 数据采集与处理, 2017, 32(3): 629-635.

[6] 刘阳. 基于文本挖掘的在线旅游产品销量影响因素分析[D]: [硕士学位论文]. 北京: 首都经济贸易大学.

[7] 张良均, 等. Python数据分析与挖掘实战[M]. 北京: 机械工业出版社, 2015.

[8] 崔永生. 在线评论文本挖掘对电商的影响研究[J]. 中国商论, 2018, 772(33): 23-29.

[9] 吴晖. 航空公司服务质量旅客满意度研究[J]. 现代商业, 2007(24): 175-176.