基于 LAD 模型的“智慧政务”中的文本挖掘应用

摘要

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

在本次数据挖掘过程中，由于群众反映问题的形式是多样化的，在处理网络问政平台的群众留言时，首先运用自然语言进行处理按照一定的划分体系对留言进行分类，以便后续将群众留言分派给相应的职能部门进行处理。

接着，我们运用 LAD 模型进行文本挖掘对附件 3 某一时段内反映特定地点或特定人群问题的留言来分类，定义合理的热度评价指标，并给出评价结果，并给出排名前 5 的热点问题。给出相应热点问题对应的留言信息。

最后，根据相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

本文主要利用自然语言处理和文本挖掘的方法解决群众留言问题并给予答复意见。

关键词：留言；热点问题；自然语言；文本挖掘；LAD 主题模型

目　录

一、引言

随着经济的不断发展，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。如何对这些网络问政平台的群众的留言文本数据进行有效的挖掘，使之造福于社会大众是一个急需解决的问题。自然语言处理作为人工智能领域的一个重要分支，其目标是使机器能够像人一样，理解自然语言。目前我们已经在包括文本挖掘、语音识别、信息检索、机器翻译等等领域取得突出成果，功能强大，应用广泛。使用自然语言处理技术对文本进行挖掘是一种有效的方式。

群众留言的数量是海量的、有噪声的、随机的、模糊的非结构化文本数据，而从这些海量文本中提取出潜在的、未知的、可能有价值的信息的挖掘技术就是文本挖掘[1]，文本挖掘技术通过将文本数据由复杂随意转变为简单有序、从无法量化转变为定量分析，从而得到可以利用的信息。

二、研究目标

本文针对目前群众留言问题的文本数据存在的处理难、分析难等问题，使用基于 LDA 模型的主题挖掘方法，通过相应的数据预处理，找出群众在某一时刻的重视的主题进行分析。并对留言的回答给予一套评价方案。

三、分析方法与过程

（一）数据的预处理

获得数据后，我们先对所给数据进行预处理。数据中存在大量重复数据甚至存在没有价值的数据，这样的数据对于后期的分析造成很大的影响，得到的结果也必然存在问题，所以上述所说的数据预处理及对这些无用的数据进行去除。

文本挖掘的一个重要准备工作就是数据预处理，为了得到 LDA 模型可输入的参数形式，必须对原数据进行结构化和标准化处理。完善的预处理过程可以使得 LDA 模型结果更有效性。

本文进行文本数据的预处理有 3 个步骤。1.筛选出文本数据中有效信息; 2.做分词准备; 3.统计高频词并选取特征词。

（二）群众留言问题：

长期以来，网络问政平台的留言都是政府与人民群众进行交流的一个重要平台和工具。在当前通信和网络都非常方便的情况下，群众反映问题的形式是多样化的，在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（参考附件 1 提供的内容分类三级标签体系）对留言进行分类，以便后续将群众留言分派给相应的职能部门进行处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。本文只要对附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

我们通常使用 F-Score 对分类方法进行评价：

自然语言处理在群众留言文本挖掘中的应用：留言文本种类繁多，包括留言主题对应的文本。这些数据是一个巨大的知识库，如果能有效利用这些数据，不仅具有巨大的商业价值，也具有很高的社会价值。而利用自然语言处理技术对这些数据进行挖掘，可以有效的利用这些数据，构建出各种应用方式。

随着人工智能技术的兴起，各行各业都积极的引入人工智能技术，各个行业都掀起了利用人工智能技术进行变革的热潮。而自然语言处理作为人工智能领域的一个重要研究领域，是计算机科学和人工智能领域科研人员的一个重要研究方向。早期的自然语言处理方法更多的使用基于统计学的模型，而自从深度学习技术在图像领域取得突出成就后，自然语言处理领域也迅速引入的深度学习方法，在应用方面，目前自然语言处理的应用范围已经非常广泛。主要包括：

（1）拼写检查。对用户输入的文本进行分析，对拼写错误进行提示，改善用户体验。

（2）情感分析。通过对用户评论的感情色彩进行判断，得出用户对产品的感受，有助于后续改善产品。

（3）机器翻译。目前各种机器翻译软件可以进行数十种语言之间的互相翻译，一是靠的海量的语料库，二就是基于深度学习的自然语言处理算法了

（4）搜索引擎。搜索引擎涉及的技术非常的多，但是自然语言处理绝对是其中比较重要的一部分，搜索引擎利用自然语言处理技术分析用户的搜索内容，从而返回最匹配的结果。

本文我们可以利用自然语言的优点来实现如下：

（1）留言问题智能管理

群众的留言问题多，如果依靠人工对留言信息进行分析提取，这样不仅费时也费力。效果也得不到保障。而利用自然语言处理技术，可以实现自动化的留言信息抽取，节约时间及人力成本，同时也能保障效果。

（2）智能的留言分类

以海量的群众留言为基础，通过自然语言技术对这些海量留言进行挖掘和学习进行留言分类。结合每一个群众的留言内容，根据关键词进行准确的分类。

(3)个性化回复

个性化回复是以群众的留言内容为基础，提取出留言信息，从词库中找到合适的回复的句子。从而准确的提取关键信息进而回复给群众。

（三）热点问题的分析与挖掘

（1）热点问题的分析:

某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。我们要对附件 3 某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，并给出排名前 5 的热点问题。给出相应热点问题对应的留言信息。

附件 3 给出了 4326 条群众的留言问题， 主要有留言编号，留言用户，留言主题，留言时间，留言详情，反对数和点赞数。在处理问题的过程中，我们首先对留言时间按升序排列，2017 年的数据只有一条，会发现，群众留言内容存在不一致性，本文我们采用 LDA 模型来分析。

发掘用户的留言主题前，先要理解主题就是文本的中心思想。所以在留言的潜在主题前，需要把文本中有价值的关键性特征词提取出来，作为可观测变量来进一步研究分析其概率分布特征状况。对于文本集中出现留言主题数目的概率分布，很容易凸显出某个潜在主题的内容，那可能需要重点关注这个关键主题，它可能是某一时刻群众留言文本集的热点。

（2）热点问题的挖掘：

LDA 主题模型的构建： LDA 主题模型属于无监督算法，可以自发的得出文本中的主题。LDA 主题模型的本质是贝叶斯概率模型，模型包含特征词、主题、文档三层结构并构成三层贝叶斯概率模型。LDA 模型假设是主题是以某个概率将文档集合中的文档组合而成的，而若干特征词可构成一个主题。

（五）答复意见的评价：

从答复的相关性看，人工受限、工作时间有限。传统人工客服回复方式，人手有限、工作效率低、信息协作共享能力差。除外，由于客服不能提供 24 小时全天候服务，导致有些留言被遗漏，不能得到及时回复。

从答复的完整性看，缺乏统一的服务标准，不能显示智慧服务的快捷化和专业化。附件 4 中可筛选出部分留言只回复收悉，并没有给出详细的解答。同时答复意见格式不统一。

从答复的可解释性看，意见缺乏专业性。由于服务过程无法监督监控，缺乏考核评价机制，以至于答复意见出现不专业，对度把握无法控制，部分问题搪塞掩饰过去。

根据以上出现的问题，提出此套评价方案：规范服务标准，培训和指导相关工作人员，提高工作效率；开发机器人客服，智能识别群众留言内容匹配相关业务知识库，实时答复；如若机器人客服无法答复，转接到相关业务部门人工答复。既可提升服务效率，还可增强群众体验感。

四、结论

本文根据群众问政留言记录数据，使用 F-Score 对分类方法进行评价，建立关于留言内容的一级标签分类模型。通过 LDA 模型的文本挖掘方法，对相应数据进行预处理，挖掘留言的热点问题。从各个维度分析相关部门对留言的答复意见，设计一套完整答复意见的评价机制。

五、参考文献

[1] 高源 . 自然语言处理发展与应用概述 [J].中国新通信 ,2019(02):117-118.

[2] 刘小安 . 卷积神经网络在自然语言处理中的应用研究综述 [A]. 中国计算机用户协会网络应用分会 . 中国计算机用户协会网络应用分会 2017 年第二十一届网络新技术与应用年会论文集 [C]. 中国计算机用户协会网络应用分会 : 北京联合大学北京市信息服务工程重点实验室 ,2017:5.

[3] 王灿辉 , 张敏 , 马少平 . 自然语言处理在信息检索中的应用综述 [J]. 中文信息学报 ,2007(02):35-45.

[4] 唐晓波，房小可. 基于文本聚类与 LDA 相融合的微博主题检索模型研究［J］. 情报理论与实践，2013 (8): 85 －90.

六、代码

```
import pandas as pd
import re
import jieba
from gensim import corpora, similarities, models
#from pandas import to_datetime

data = pd.read_excel('F:\\C 题全部数据\\附件 2.xlsx')
data.shape


data_dup = data['留言主题'].drop_duplicates()
data_qumin = data_dup.apply(lambda x: re.sub('x', '', x))
data_cut = data_qumin.apply(lambda x: jieba.lcut(x))

stopWords = pd.read_csv('stopword.txt', encoding='GB18030', sep='hahaha',
header=None)
stopWords = ['≮', '≯', '≠', '≮', ' ', '会', '月', '日', ' - '] + list(stopWords.iloc[:, 0])

data_after_stop = data_cut.apply(lambda x: [i for i in x if i not in stopWords])
dictionary = corpora.Dictionary(data_after_stop)
corpus_vector = [dictionary.doc2bow(word) for word in data_after_stop]
#for doc_vec in corpus_vector:
    #    print(doc_vec)
     # corpora.MmCorpus.serialize('corpus_vector.mm', corpus_vector)
dictionary.token2id
dictionary.get(969)
corpus_vector
model = models.TfidfModel(corpus_vector)
tfidf = model[corpus_vector]
for doc_tfidf in tfidf:
       print(doc_tfidf)
```