网络问政平台挖掘多属性综合评价方法

摘要

本文旨在基于来自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见等信息，通过数据特征工程、数学建模、数据分析、自然语言处理和文本挖掘的方法，解决群众留言的多分类问题，结合语言模型进行文本分析，探讨如何提高问政信息处理效率的指导性问题。

针对任务一，建立了 word2vec 模型对附件 2 中的数据进行文本特征选择，用数学建模的思想选取文本分类特征，并构建算法模型，建立关于留言内容的一级标签分类模型，并使用分类模型的评估方法—F-Score 对模型精准率和召回率进行评分，通过评分不断进行模型优化，目的是增大 F-Sore 的数值。Word2vec 工具对词向量的嵌入解决性能良好，挖掘得到了比较有意义的关键字，提高了效率，降低了差错率，减少了工作量。

针对任务二，使用 LsiModel 模型算法，将语料库计算出 Tfidf 值，获取词典 token2id 的特征数，通过 doc2bow 计算测试数据的稀疏向量，求得测试数据与样本数据的相似度。根据这种模型的算法能有效地有针对性地处理和提升服务效率，进行有效地热点问题的分类。

针对任务三，针对相关部门对留言的答复意见，分析答复的相关性、完整性、可解释性等多个角度，从答复的多个性能对答复意见的质量给出一套评价方案，并尝试通过实践进行实现。

关键词：互联网、网络问政、分类、信息处理效率

Abstract

The purpose of this paper is to source from the Internet based the crowd asked zheng message records, and the relevant departments for some replies the opinions and other information, through the data characteristics of the engineering, mathematical modeling, data analysis, natural language processing and text mining, the method of resolve the message the multiple classification problems, language model for text analysis, to explore how to improve administrative guidance information processing efficiency question.

Word2vec model was built for a mission, to the data in attachment 2 text feature selection, select text classification feature, with mathematical modeling thought and algorithm model was constructed, level 1 label on the message content classification model is set up, and use the evaluation method of classification model - F - Score rating precision rate and recall rate of the model, through the Score of model optimization, the goal is to increase the value of F - Sore.

Word2vec tool embedded word vector solution performance is good, mining more meaningful keywords, improve the efficiency, reduce the error rate, reduce the workload.

For task two, we use lsimodel model algorithm to calculate the tfif value of the corpus, obtain the number of characteristics of the dictionary token2id, calculate the sparse vector of the test data through doc2bow, and obtain the similarity between the test data and the sample data. The algorithm based on this model can effectively deal with and improve the service efficiency, and effectively classify the hot issues.

In view of task 3, the author analyzes the relevance, completeness and interpretability of the replies to the comments from relevant departments, and gives a set of evaluation scheme for the quality of the replies from the aspects of multiple performances of the replies, and tries to realize it through practice.

Key words: Internet, network politics, classification,

information processing efficiency

目录

一、 问题分析

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，人工分类无法细致，人工根据经验处理，无法完善地高效地处理多条信息。利用文本分类模型，对民意信息进行特征选择和分类，有利于提高电子政务系统的检索效率和分类速度。

问题给出的数据中，附件 1 是关于三级的分类，它们是包含与被包含的关系，详细描绘了城乡建设，党务政务，国土资源，环境保护，纪检监察，交通运输，经济管理，科技与信息产业，民政，农村农业，商贸旅游，卫生计生，政法，教育文体，劳动和社会保障这 15 个方面对应的分类指标和关键字，附件 2 给出的是留言的数据 9211 条留言的内容，包括了留言主题，留言详情，一级标签等数据。附件 3 新增了反对数和点赞数两项指标，附件 4 增加了答复详情和答复时间两个指标。

问题所给的任务一要求根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。标签体系是多个类型问题的分类标准，以便快速将问题分类到同一出发点类型中。在实际分类中，要注意文本的选择以及标签建立的依靠。

任务二要求利用附件 3 将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。通过阅读国内外文献我们了解到，在机器学习的过程中数据挖掘的聚类算法，根据数据特性的不同可以采用不同的度量方法，在这里我用了马氏距离的建模方。因此我们对已有的留言进行了一定的归类区分，来构建分析模型。

任务三要求结合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。

二、数据准备

2.1 样本数据提取

样本数据提取最核心的一步是利用 python 读取 Excel 指定的列并将内容保存为 txt 内容。首先需要下载一个 python 模块，指定文件的路径放置 Excel 表格。通过另存在的 txt 的数据，首先进行分词处理，在为短词的基础上，用数学建模的思想选取文本分类特征，构建样本数据词库。

2.2 删除与分析无关数据

对于样本数据提取后，数据杂乱无章，其中包含许多的无关数据，即无用数据，实时数据，这都不利于问题的分析，所以应当在样本数据提取后进行无关数据的删除与分析。首先进行数据真实性判断，即通过 excel 来观察所给数据的整体趋势，利用拟合技术多种手段实现验证数据的真实性。然后进行数据分析，对于异常值进行处理即对于 NAN 数据或奇异点，采取方式为基于拉依达准则的数据异常值处理。在删除无关数据后进行分析，为何会出现这些数据。得出结论服务与任务一即任务二。

2.3 构造分析需要的指标

在进行数据的筛选剔除无关等分析后，需要对所取得数据进行一个分类分析指标。网络问政平台的数据详细描绘了城乡建设，党务政务，国土资源，环境保护，纪检监察，交通运输，经济管理，科技与信息产业，民政，农村农业，商贸旅游，卫生计生，政法，教育文体，劳动和社会保障这 15 个方面对应的分类指标和关键字，更有留言的数据内容，包括了留言主题，留言详情，一级标签等数据。探究政务机构本身因素以及多方面的分类与留言，通过统计学和计量学的方式构建指标，精确探究因素的影响程度，才可进一步分析从而得出结论提出建议。在众数据特征提取可以分三类即时域特征，频域特征，时频域特征。其中时域特征包括：最大值，最小值，方差，标准差等。而频域特征包括：光谱分布以及 FFT 系数平方之和等多种。第三类的时频域特征可以包括小波分解，小波包分解重构等。最后即依据问题数据特征构建分析所需要的指标。

2.4 标准化处理

在数据分析之前，我们常需要将数据标准化。利用标准化后的数据方可进行数据分析。数据标准化通常来说也就是将统计数据的指数化。数据同趋化处理和无量纲化处理是数据标准化处理主要两个方面。通过标准化处理，原始数据均可转换为无量纲化指标测评值，各指标值处于同一个数量级别上，最后可以进行综合测评分析。

三、模型假设

为了便于问题的研究，对题目中某些条件进行简化及合理假设。

1） 附件 2 中有留言主题和留言详情，假设群众留言是留言主题且对问题无差异。

四、任务一

4.1 对数据进行文本特征选择

问题附件 2 给出的是留言的数据 9211 条留言的内容，包括了留言主题，留言详情，一级标签等数据。而问题给出的附件 1 是关于三级的分类，它们是包含与被包含的关系，详细描绘了城乡建设，党务政务，国土资源，环境保护，纪检监察，交通运输，经济管理，科技与信息产业，民政，农村农业，商贸旅游，卫生计生，政法，教育文体，劳动和社会保障这 15 个方面对应的分类指标和关键字。通过三级分类以及留言的分析，可以得出文本特征。通过准备好的数据集，包括加载数据集和执行基本预处理，然后把数据集分为训练集和验证集。在此上基础以数学建模思想进行归总分析，从而进行文本特征选择

4.2 建立关于留言内容的一级标签分类模型

在建立一级标签分类模型的情况下，应当先进行特征工程。在进行这一步，原始数据会被转换为特征向量，同时也会根据现有的数据创建新的特征。在现存方式从数据集中选出重要的特征，有以下几种方式：即计数向量作为特征、TF-IDF 向量作为特征、单个词语级别、多个词语级别（N-Gram）、词性级别、词嵌入作为特征、基于文本/NLP 的特征、主题模型作为特征。在网络问政平台中，三级标签通过留言内容数据创建新特征得出二级特征。在二级特征的情况下运用同样的方法即可依靠留言从而建立一级标签分类模型。

4.3 模型优化得出分析结论

关于留言内容的一级标签分类模型。设置标签的意义在于更好地分类问题，服务于社会。用户在使用网络问政平台的同时会因为各种各样不同的问法而产生不同的问题，但归根其起始点，都是从同一标签层层众化而开始的。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

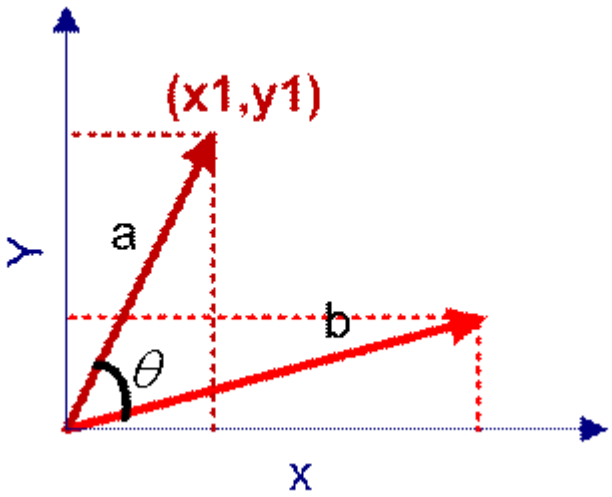
高等问题。由于三级标签过于细扩，其中会出现数据过于复杂与工作量大而产生误差以及难以解决问题。一级标签的出现，将问题更加扩大化，在同一大类内进行分析更可以高效快速解决人们日常使用网络问政平台的需求。

五、任务二

5.1 相似性统计

根据某一时间段内群众集中反映的某一问题可称为热点问题，需要相似性处理，在数据挖掘中要计算相似度，进行聚类分析和协同过滤，我了解到了余弦相似度，余弦相似度是通过计算两个向量余弦值来评估他们的相似度，衡量的是多维空间中两个个点之间的余弦值。并将在 python 数据挖掘中用向量相似度的方法进行相似性统计。余弦值越趋近于 1，代表两个向量的方向越接近，也就代表两个事件是相关的。

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}}$$



5.2 热度统计

先定义一个热点问题的标准并运用文本挖掘的方法去挖掘热点问题，使用Counter 算法来跟踪出现的次数并进行统计。

5.3 热词统计

在中文的文本挖掘中，对热点挖掘中的热词进行统计事关重要，运用 jieba 分词并进行词频统计。

六、任务三

6.1 留言的答复意见性能分类

在问题附件 4 中包含相关部门对留言的答复意见。可将答复意见性能分为答复相关性、答复完整性、答复可解释性。在分类分析之前，先对数据进行 excel 表格数据读取，转化为 txt 格式。提取不全面数据，分析原因进行剔除。答复从此三方面出发进行分类可以全面地了解答复地多样。相关性得出答复的重合率，在标签基础的情况下，通过相关性的分析可以了解答复的热点重要程度。答复的完整性是一个答复是否具有使用与解决问题的价值判断。当答复的完整性过低时他将作为干扰数据而被剔除。答复的可解释性是现实生活解决问题的重要依据。它直接关系到答复是否应当存在，凭什么存在的重要基础。

6.2 分析模型的构建

数据分析模型有很多，在数据分析中，如果将数据进行分类就能够更好的分析。分类分析是将一些未知类别的部分放进我们已经分好类别中的其中某一类；或者将对一些数据进行分析，把这些数据归纳到接近这一程度的类别，并按接近这一程度对观测对象给出合理的分类。

6.3 结合性能分析得出评价方案

在当前互联网社会中，网络问政的兴起是网络民主政治进步的标志之一。他

是一种崭新的双向交流互动工具，公众意愿表达的途径在慢慢拓宽。通过留言分析我们不难看见，民众逐渐适应网络问政模式，而网络问政平台无疑是将权力还给社会的有效途径。而通过回复时间我们不难发现网络问政平台系统存在回复时间差距问题。而这根本上是民众提问分类层度不够导致的回复不及时。但这也只存在于个别方面，并不是大趋势。如果可以通过数学分析执行树解决这个问题，将会打打提高网络问政平台的多属性使用。

从民众的心理来看的话，问题回复率降低并不会改变他们通过网络问政平台提出问题与建议的想法。而这大部分是相对固化的，而不是实际使用过程中的感受。而归结于问题答复的形式较为落后。因此，解决这个问题，可以从答复提问的立体化问政体系出发改变，创新问政平台提问答复的互动模式，是一个有效的解决途径。

从答复相关性、答复完整性、答复可解释性三个方向出发，即可知道目前网络问政平台的单调性，效率性并算不上完善，解决这两个方面，网络问政平台或许会迎来一个不一样的缤纷世界。

七、参考文献

[1]李翠霞 林楠 浅析文本挖掘技术 郑州大学软件学院

[2]蒋良孝 蔡之华 文本挖掘及其应用 中国地质大学计算机科学与技术系