“智慧政务”中的文本挖掘应用

摘要：随着科技的发展，大部分人都选择用微博，微信，或者邮箱等工具给相关政府部门进行网络留言提出建议。但由于各类社情民意相关的文本数量剧增，导致相关部门的工作量增大。所以我们通过对群众留言分析和热点问题的处理，政府部门能够更有效的管理政务。本文是针对附件给出的群众问政留言记录及相关部门对群众留言的答复意见，进行数据挖掘与分析。为了保证评论数据挖掘分析的质量和全面性，我们对数据进行预处理——留言 “去空、去重”、停用词过滤和分类等。之后，本文主要进行三个方面的数据挖掘分析工作：一 方面是根据群众留言来分类；一方面是根据热点问题以及热点问题留言的挖掘；另一方面相关部门对留言答复意见的质量评价方案。

关键字：数据挖掘　　预处理　　网络留言

Data mining analysis based on network message

Abstract: With the development of science and technology, most people choose to use micro-blog, WeChat, or mailbox and other tools to the relevant government departments to leave a message to make suggestions. However, due to the sharp increase in the number of texts related to various social and public opinion, the workload of the relevant departments has increased. So through the analysis of mass messages and the treatment of hot issues, government departments can manage government affairs more effectively. This paper is for the attachment given the mass question political message record and the relevant departments of the response to the mass message, construct a model, data mining and analysis. In order to ensure the quality and comprehensiveness of the review data mining analysis, we pre-process the data - the message "de-emptying, de-heavy", de-icing and classifying words, etc. After that, this paper mainly carries on three aspects of data mining analysis work: on the one hand, according to the mass message to classify, on the other hand according to the hot issues and hot issues of the excavation of messages, on the other hand, the relevant departments to the message reply to the quality evaluation program.

Keywords: Data Mining Pre-processing Network Message

目录

1.挖掘目标 1

2.分析方法与过程 1

2.1.总体流程 1

2.2 具体步骤 2

步骤一：数据预处理 2

步骤二：TF-IDF 算法 3

步骤三 ：文本表示 4

步骤四 答复意见评价 5

3.结论 6

4.参考文献 7

1.挖掘目标

本次建模针对网络问政平台的群众留言，采用特征特征向量法对文本数据进行分析，来达到以下两个目标：

（1）、通过文本数据挖掘方法来对网络平台的群众留言进行处理，把相似问题的留言划分为一类，并定义一个合理的评价指标来把热点留言做一个排序。文本挖掘技术大大降低了工作人员的工作量和差错率，同时又能提高工作效率。

（2）、对群众的留言数据进行分析，有效的了解了人民群众的想法、意见和需求，是实现政府了解民意、汇聚民智、凝聚民气的重要渠道。

2.分析方法与过程

2.1. 总体流程

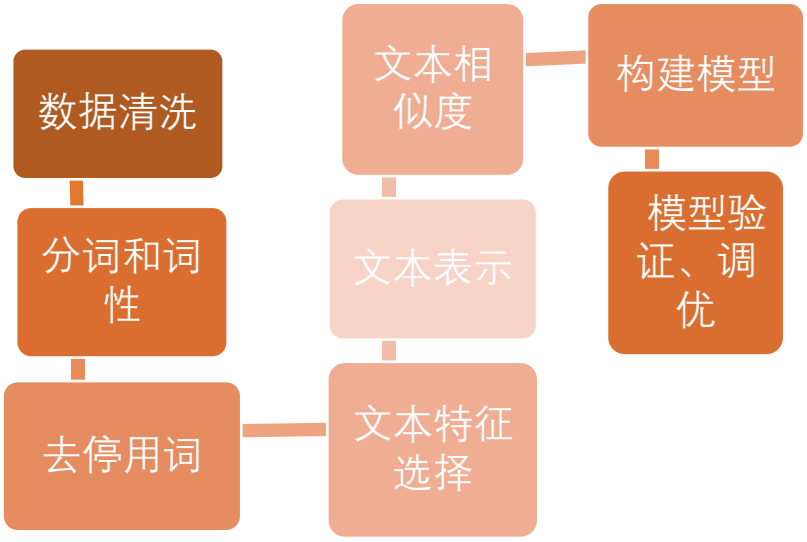


图 1　总体流程图

本用例主要包括以下几个步骤：

步骤一：对文本进行预处理，留言的数量大。第一步：要“去空、去重”；

第二步进行停用词过滤，去除掉留言中无意义的词，如：的，了，啦，哈哈。

步骤二：TF-ID 算法

文本的数学处理，对文本的特征进行选择，用数学方法选取最具分类信息的特征。常用的方法有：文档频率、信息增益、卡方校验和相互信息等。利用 TF-IDF 方法对特征进行权重对比：一个词的重要度与在类别内的词频成正比，与所有类别出现的次数成反比。

步骤三：对文本进行表示：词向量作为文本的基本结构——词的模型。良好的词向量可以达到语义相近的词在词向量空间里聚集在一起，这对后续的文本分类，文本聚类等等操作提供了便利。

步骤四：答复意见评价：针对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

2.2 具体步骤

步骤一：数据预处理

有些数据文本并不能用数据挖掘技术。即使可用也需要建立在对数据文本的数据进行预处理的基础之上。如果要对留言数据进行情感分析，就必须先将文本数据进行预处理，转化为结构化的数据。

1) “去重”、“去空”

对于存储了全部留言记录的 pdf 文件，每一句留言数据代表了一个留言文本，但却难免会出现两个完全一样的文本和一些空行。本文首先进行“去重”、“去空”的预处理工作，从而起到了降低有关部门的工作量。 在导入这些留言文本时，要同时进行是否为空的判断，只导入不为空的文本，从而过滤掉了空白文本。以下是“去重”的代码程序

```
commentsList2.Add(commentsList [0]);

  for(int i=1; i< commentsList.Count; i++)

{

   IsRepeated = false;

    For (int j=0; j<1; j++)
```

在导入这些留言文本时，要同时进行是否为空的判断，只导入不为空的文本，从而过滤掉了空白文本。以下是去空的程序：

```
StreamReader sr = new StreamReader
      String line
      While(line = sr.ReaderLine() != null)
{
          If(line.ToString() !=" ")
{
          commentList.Add(line.ToString());
          }}
```

2）停用词过滤:
评论文本在经过去重、去空后，并非所有的剩下的词语都可以作为特征词，里面还有一些包含的信息量很低甚至没有信息量的词语，需要将它们清洗掉，否则将会影响部门的对留言的答复。在问题的背景中，，用处理自然语言处理技术之前会自动过滤掉某些字或词，这些字或词即被 称为 StopWords（停用词）。 本文采用了“词性+停用词表”的过滤方法。停用词词性都用 StopwordPropsList 表示，然后对每个分词后的文本进行遍历扫描，把对应词性的词语全部过滤掉。部分停用词词性列表如下所示：

;

```
StopwordPropsList.Add(“p”);

StopwordPropsList.Add(“pde”);

StopwordPropsList.Add(“ple”);
```

步骤二：TF-IDF 算法

文本的数学处理，对文本的特征进行选择，用数学方法选取最具分类信息的特征对群众进行分类。常用的方法有：文档频率、信息增益、卡方校验和相互信息等。利用 TF-IDF 方法对特征进行权重对比：一个词的重要度与在类别内的词频成正比，与所有类别出现的次数成反比。

TF-IDF 算法步骤

第一步，计算词频

$$词频(TF) = 某个词在文章中的出现次数$$

考虑到文章有长短之分，为了便于不同文章的比较，进行"词频"标准化。

$$词频(TF) = \frac{某个词在文章中的出现次数}{文章的总词数}$$

第二步，计算逆文档频率：

这时，需要一个语料库（corpus），用来模拟语言的使用环境。

$$逆文档频率(IDF) = \log\left(\frac{语料库的文档总数}{包含该词的文档数 + 1}\right)$$

如果一个词越常见，那么分母就越大，逆文档频率就越小越接近 0。分母之所以要加 1，是为了避免分母为 0（即所有文档都不包含该词）。log 表示对得到的值取对数。

第三步，计算 TF-IDF：

$$TF - IDF = 词频(TF) \times 逆文档频率(IDF)$$

步骤三 ：文本表示

本步骤主要是结合步骤一、二的文本处理之后，通过 word2vec 方法对文本进行词向量的表示，再利用主题特征方法 LDA、LSI 等表示词向量的主题特征，便于后面文本分类和聚类。

1）one-hot 表示

one-hot 简称读热向量编码，也是文本表示中特征工程中最常用的方法。它的步骤如下：

1.构造文本分词后的 map，每个分词是一个 bit，bit 值=0 or 1

2.每个分词的文本表示成该分词的比特位是 1,其余位为 0 的矩阵表示。

比如“A 市西湖建筑集团占道施工有安全隐患”这个留言主题，在经过上述几个步骤进行文本处理后，已变为“A 市 西湖 建筑 集团 占道 施工 安全 隐患 ”

可以构造一个词典，{ “A 市”：1,“西湖”：2,“建筑”：3,“集团”：4,“占道”：5,“施工”,“安全”,“隐患”}

每个词典索引对应着比特位，那么利用 One-hot 表示为：

A 市：[1,0,0,0,0,0,0,0,0],西湖：[0,1,0,0,0,0,0,0,0]…等等，以此类推

2）共现矩阵

共现矩阵顾名思义就是共同出现的意思，词文档的共现矩阵主要用于发现主题(topic)，用于主题模型，如 LSA。

局域窗中的 word-word 共现矩阵可以挖掘语法和语义信息，**例如：**

-A 市西湖建筑集团占道施工有安全隐患

-A 市在水一方大厦人为烂尾多年，安全隐患严重

以上两句话，设置滑窗为 2，可以得到一个词典，{“A 市西湖”,“西湖建筑”,“建筑集团”,“集团占道”,“占道施工”,“施工安全”,“安全隐患,“A 市水一方”,“水一方大厦”,“大厦人为”,“人为烂尾”,“烂尾多年”, …}

我们可以便得到一个共现矩阵。中间的每个格子表示的是行和列组成的词组在词典中共同出现的次数，也就体现了共现的特性。

3）存在的问题和不足

word2vec 模型的问题在于词语的多义性，导致主题可能存在一定的偏差。比如这个“交通”词语，常见的词组有一些交通工具如“小汽车”、“摩托车”、“火车”等，但在留言中，也可能出现“交通银行”等词会使主题的判断出现偏差，对于 word2vec 模型来说，它倾向于将所有概念做归一化的平滑处理，所以会得到一个最终的表现形式。

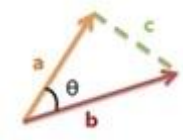
步骤四 答复意见评价

本步骤主要是针对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。

1）相关性：答复意见内容是否与问题相关

对于相关性，我们将利用余弦相似度算法对问题与留言进行文本相似度量化

余弦相似度算法



$$\cos\theta = \frac{a \cdot b}{||a||||b||}$$

余弦值的范围在[-1,1]之间，值越趋近于 1，代表两个向量的方向越趋近于 0°，他们的方向更加一致，相应的相似度也越高。需要指出的是，在文本相似度判定中，因为文本特征向量定义的特殊性，其余弦值范围为[0,1]，即向量夹角越趋向于 90°，则两个词向量越不相似。

由于在经过之前的步骤后，产生的表征文本特征的两个向量的长度是不同的，因此需要对上述向量进行处理。我们可以采用剔除掉向量中不重要的词汇，从而使得两个向量长度保持一致，目前主要依靠经验设定一些关键词来处理，但是其准确率不可保证；当然也可以利用归并向量，并根据原向量是否在新向量（归并后的向量）存在，若存在则以该词汇的词频来表征，若不存在则该节点置为 0，示例如下：

文本 1: 我/是/王中国/

文本 1: 我们/是/王家人/

Vector: 我/是/王家人/我们/
Vector1 = (1, 1, 1, 0)
Vector2 = (0, 1, 1, 1)
上述“/”为采用 IK 分词，智能切分后的间隔符，则归并后的向量如 Vector 所示，对齐后的向量分别为 Vector1 和 Vector2。之后则根据两向量的余弦值确定相似度。
2）完整性：是否满足某种规范
对于这一部分，首先我们可以通过构造一篇答复模板，这个模板有一些固定的格式，比如开头有“网友 XXX 您好！您的留言已收悉。现将有关情况回复如下”，结尾有“感谢您对我们工作的支持、理解与监督!”及答复日期等。然后利用（1）中的余弦相似度方法，对完整性给出一个量化指标。
3）可解释性：答复意见中内容的相关解释
这一部分，一方面我们认为可以进行抽样调查，每月随机抽取 1/5 的回复拿出来进行可解释性评价，而最终得到的 x 是以年为单位，也就是说，在每月评价后再利用统计方法算出年度 x 值，主要考虑用均值、方差两项统计值。而每月的 1/5 中，可以对解决问题的进度，以及执行效率、论据透明度、逻辑性四个方面请专业人士进行人工量化，再进一步得到一个较为合理的可解释性指标。

于是，我们对上述三个指标进行综合分析，最终评价分数=40%*r+30%*a+30%*x（r 表示内容与问题的相关程度，a 表示完整性指标，x 表示可解释性）

3.结论

总结本次比赛，我们根据附件表中留言数据的特点，对数据主要进行了四步操作。步骤一中，对文本进行预处理，主要是“去重”、“去空”，接着用 jieba 库分词，最后对停用词进行过滤。

步骤二中，利用 TF-IDF 算法，对文本特征进行选择，用数学方法选取最具分类信息的特征对群众留言进行分类，完成子任务 1——文本分类。

步骤三中，文本表示主要通过 word2vec 方法中的对文本进行词向量的 one-hot 表

示，再利用共现矩阵表示词向量的主题特征，这样便可以从中发现留言中的热点问题，完成子任务 2——热点问题挖掘。最后还对 word2vec 的不足进行了说明。

步骤四中，针对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，并给出了一个量化指标方法，，完成子任务 3——答复意见评价
本次留言数据挖掘分析的过程中，进行了大量的自然语言处理技术，通过特征向量的方式，基本完成了本问题的三个子任务。

4.参考文献

百度百科、CSDN