“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。

对于问题 1，通过利用 excel 对去重后的留言主题进行分类筛选，通过对关键字的筛选来分出某条留言主题是属于二级分类和三级分类中的哪一类，先筛选一级标签里面的一级分类，再接着筛选留言主题，最后进行整合。

对于问题 2，利用网络传播热度指数(R)[1] 计算公式进行热度计算，先对留言主题进行文本筛选，关键词包含筛选，筛选结果进行技术统计，筛选结果数值越大，热度指数越高。

对于问题 3，先对留言时间和答复时间进行时间差的计算，计算出时间差后，使用函数，使用公式进行计算，最后进行排序，数值越小则评价越好。

关键词：去重、热度指数、文本筛选、统计。

“Application of text Mining ”in Intelligent Government Affairs

Abstract

In recent years, as online political inquiry platforms such as Wechat, Weibo, Mayor's mailbox and Sunshine Hotline have gradually become important channels for the government to understand public opinion, gather people's wisdom and rally people's morale, the amount of text data related to all kinds of social conditions and public opinions has been rising continuously, which has brought great challenges to the work of relevant departments, which mainly rely on manual to divide messages and sort out hot spots.

At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of an intelligent government system based on natural language processing technology has become a new trend in the innovative development of social governance. it plays a great role in improving the management level and administration efficiency of the government.

For question 1, through the use of excel to classify and screen the deduplicated message topics, and through the screening of keywords to sort out which category a message topic belongs to in

the second-level classification and the third-level classification, first filter the first-level classification in the first-level label, then screen the message topic, and finally integrate it.

For question 2, the heat index (R) formula is used to calculate the heat. First, the text of the message topic is filtered, the keywords are screened, and the screening results are technical statistics. The larger the value of the screening result is, the higher the heat index is.

For question 3, first calculate the time difference between the message time and the reply time, calculate the time difference, use, function, use, formula, and finally sort it. The smaller the value, the better the evaluation.

Keywords: weight removal, heat index, text screening, statistics.

目录

1、挖掘目标

（1）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至相应的职能部门处理。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。

(2) 对热点问题进行分类, 有助于相关部门进行有针对性的处理, 提升服务效率

（3）提供评价方法。

2、数据方法与过程

总体流程图

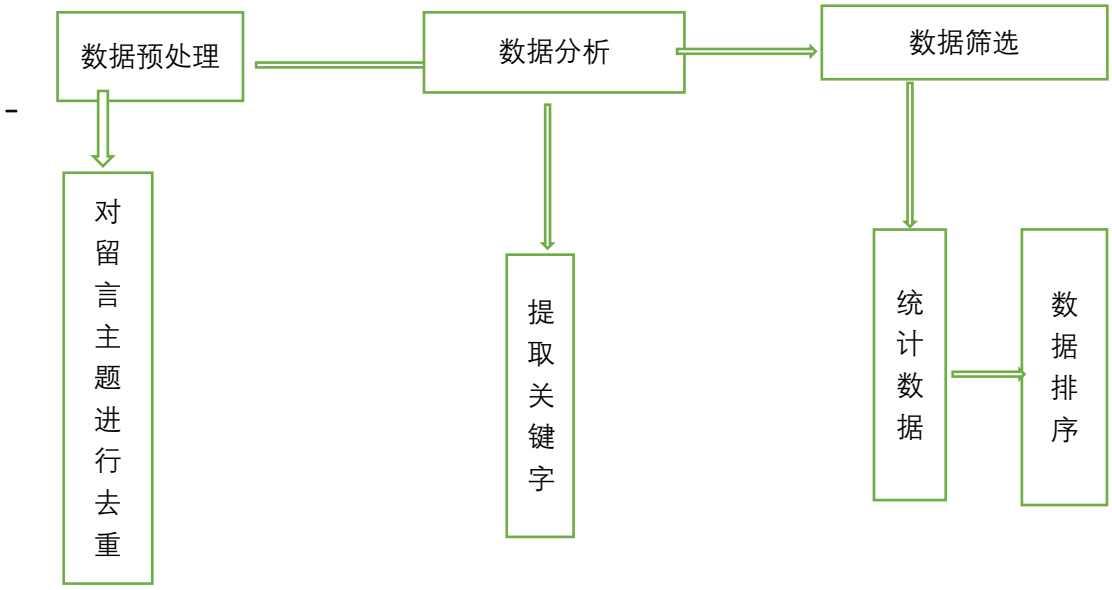

图 1: 总体流程图

本用例主要包括以下步骤：

步骤一：数据预处理，在题目中所给出的数据中，出现了多个重复的留言主题数据，在原数据的基础上进行去重处理。

步骤二：数据分析，观察数据的年份走向趋势，结合图像，预测接下来的趋势。

步骤三：数据筛选，利用文本对关键词进行筛选，选择文本筛选包含关键词筛选，对结果进行统计分类情况

2.1 问题 1 分析方法与过程

留言分类，先对数据进行去重处理，，进行筛选操作，筛选留言主题内容，包含关键词筛选，最后进行结果统计。

2.2 问题 2 分析方法与过程

热点问题挖掘，参考网络传播热度指数(R)，对数据进行筛选计算，通过团队合作，把统计结果结合在一起。

2.3 问题 3 分析方法与过程

根据留言时间和答复时间计算出两者的时间差，并以以分钟为单位，再将时间差的数据进行降序排序，根据计算结果，推出：用时越短，评价越高。

3、结果分析

3.1 问题 1 结果分析

留言分好类，可以让政府的工作效率提高，按照原数据分好的一级分类和附件 1 所给的二级分类和三级分类，在附件 2 中进行二级分类中所含的关键词进行筛选，例如一级分类中的教育文体，其二级分类的关键词有：考试、失学、辍学、体育、文化、文物等，筛选出关键词后，在用同样的方法进行三级分类，所筛选出的结果就属于对应的分类结果。

3 .2 问题 2 结果分析

及时发现热点问题，及时解决百姓所需。在留言主题进行实习、打工、纳税、道路、工作、扰民等关键字筛选，在统计其留言条数，进行排序。

3.3 问题 3 结果分析

评价方案可以提高工作人员的积极性，提高工作效率，评价的标准是以回答的时间快慢来决定。如果答复时间与留言时间相差很远，有可能会导致错过解决该问题的最佳时间。

4、结论

对留言信息进行研究，可以减少人工的工作量，提高政府的工作效

率，更好的为人民服务，解决人民困扰的问题，可以赢得民心，有利于社会的发展。掌握热点问题，观察问题发展的趋势轴向，预测热点问题，时刻关注民生所向。

5、参考文献·

【1】网络传播热度指数(R)计算公式
R=Y1xb1+Y2xb2+ Y3xb3+ Y4xb4
r=(+-1)x100
X1=新闻条数 x0.189+电子报刊条数 X0.175
+客户端条数 x0.187+微信公众号条数 x0.182
+政务条数 x0.167+外媒条数 x0.100
合成系数
X2=微博条数 x0.319+论坛条数 x0.355
+博客条数 x0.326
X3=视频条数
X4=其他网站条数
a1=1.05
a2=1.001
标准化公式参数
a3=1.1
a4=1.005
b1=40%
b2= 45%
权重系数
b3=5%
b4=10%