第八届“泰迪杯”

全国数据挖掘挑战赛

——

C 题：“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。

智慧政务是解决社会问题的有效方法。随着智慧政务的逐渐落地、智慧应用的日趋多元化,加强对于智慧政务的发展、提供基于数据的精准化规划指导变得更加重要。在此背景下，本文围绕利用自然语言处理和文本挖掘的方法解决智慧政务中出现的处理数据问题同时，希望结合研究中对所学知识的运用,通过讨论,归纳出该数据与研究方法的优势及应用，从而为新数据和文本挖掘技术在研究中的推广起到促进作用。

关键词：智慧政务；文本数据；自然语言处理技术；大数据；分类

Abstract

In recent years, with wechat, microblog, mayor's mailbox, sunshine hotline and other online political platforms gradually becoming an important channel for the government to understand the public opinion, gather people's wisdom and gather people's spirit, the amount of text data related to various social situations and public opinions has been increasing, which has brought great challenges to the work of relevant departments that used to rely mainly on human 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 smart government system based on natural language processing technology has become a new trend of social governance innovation and development, which has a great role in promoting the management level and efficiency of the government.

Smart government is an effective way to solve social problems. With the gradual implementation of smart government and the diversification of smart applications, it becomes more and more important to strengthen the development of smart government and provide accurate planning guidance based on data. In this context, this paper focuses on the use of natural language processing and text mining methods to solve the problem of data processing in smart government. At the same time, it hopes to combine the use of the knowledge learned in the research, through discussion, summarize the advantages and applications of the data and research methods, so as to promote the promotion of new data and text mining technology in the research.

Key words: smart government; text data; natural language processing technology; big data; classification

目录

第一章 绪论

本章主要介绍了研究问题提出的背景依据,即智慧政务已经受到了高度的关注，并且对准确化的数据处理提出了更高的要求。阐明了研究开展的目的，一方面，用数据说话，为决策提供科学依据，以优化决策过程，做到科学决策、科学实施、科学管理；另一方面，充分利用大数据时代公众的参与热情，确保问题的界定、政策的制定与实施等环节符合民意。

1.1 研究背景

从概念的来源来看，与智慧城市一样，智慧政务具有舶来词中国化的特征。在英文中，并没有与其意义完全相同的学术概念。与之相近的概念有:Smart government, Intelligent govermment, Ubiquitous govermment(简称 u-Goverment)以及 government3.0 等。尽管这些概念的表述形式有所不同，但其内涵和外延却极为相近”。本文对智慧政务作如下界定:智慧政务是利用大数据、云计算、物联网、移动互联网等技术，在电子政务的基础上，通过数据共享、整体协同、智慧管理等，实现公共服务从全能型转向智慧型、服务型，因此，智慧政务是电子政务发展到高级阶段的必然产物。

智慧政务是互联网+线下处理的方式，是可以实现政务服务高效化，数据实时化，响应及时化的政务工具，可以简单、高效的解决百姓的各类问题，数据统计清晰明了化。而“政府大数据”是最权威、最具公共服务价值的数据，比如人口、教育、交通、环境、公共资源、公共安全等。能够把这块的数据开发好、利用好，对增强公共服务的针对性，提高公共治理的精准性，提高工作效率和公众满意度具有不可估量的作用大数据时代，问题的复杂性与不确定性使得政府难以再依靠传统的治理经验解决问题，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。在

此情况下，政府必须充分利用大数据原理进行科学决策。从问题的界定到政策的制定、执行、评估等各个阶段，都需要数据处理。

1.2 研究目的

- 提高政府管理水平

通过处理互联网上收集到的数据，并对文本数据做相关处理，多方数据联动，实时、精确、智能输出客观评估结果，加强政务数据的获取、组织、分析、决策。

- 提高政府的决策效率

多部委，全方位数据共享，提高决策的科学性和精准性，提高政府预测预警能力以及应急响应能力，节约决策成本。

- 提高政府公共服务水平

借助大数据，还能逐步实现立体化、多层次、全方位的电子政务公共服务体系，推进信息公开，促进网上办事实时受理、部门协同办理、反馈网上统一查询等服
务功能，加快推进智能化电子政务服务和移动政务服务新模式的初步应用，不断
拓展个性化服务，进一步增强政府与社会、老百姓直接的双向互动、同步交流。

- 提高城市管理水平

通过充分利用大数据的各类资源，实现统一协调的管理信息整合，发挥城市网格化管理效用，达到最大程度的共享应用，以提升城市和社区的服务质量、提高服务能力、加强服务管理，创建服务型社会，使城市管理工作和社区服务水平迈上更高的台阶。

- 促进信息服务业发展

对从民众收集到的文本进行挖掘，建立基于自然语言处理技术的智慧政务系统，能够促进信息服务业的发展，足不出户也能解决问题。

1.3 研究意义

“党的十九大报告明确提出，要转变政府职能，深化简政放权，创新监管方式，增强政府公信力和执行力，建设人民满意的服务型政府。互联网+政务服务”就以简政放权、放管结合、优化服务为核心，创新践行“互联网+”思维，以权力清单为基础，以数据共享和流程优化为重点，以大数据、云计算、移动互联网、物联网等新兴技术为支撑，以增强人民群众获得感为落脚点，开启了从“群众跑腿”到互联网“数据跑腿”的服务新模式，实现了由“政府端菜”向“群众点餐”的转变。事实充分证明，“互联网+政务服务”已经成为政府职能转变的新动力、建设服务型政府的重要路径、“放管服”改革的基本依托、推动释放市场潜力活力的新增长极和供给侧结构性改革的有力杠杆。

大数据时代智慧政务建设的意义：

(1)推进多方力量协同的社会治理智慧政务作为社会治理模式的一种,是建立在新的有效政务需求之上的。智慧政务将改变过去由政府主导的局面，智慧政务的建设将会融合政府、企业、社区等多种力量进行社会治理。

(2)推动政府职能的转型升级。随着社会发展与科技的进步，公民及.一些社会组织利用智慧政务平台，发布自己的公共需求，其需求也将会从各种渠道得以满足，形成自我管理与自我服务。只有当需求超过自治能力时，才会向政务部门求助。因此，智慧政务将推动政府向有限政府、服务型政府转变。

(3)推进政务自身的发展。在大数据时代，智慧政务建设将利用大数据等技术，及时采集数据，对公民的需求进行预测;同时用户也可以通过智慧政务获取信息、对政务部门]进行评价，对智慧政务实现动态监督，最终使信息化与政务相结合，实现社会的有效治理。

1.4 思路与框架

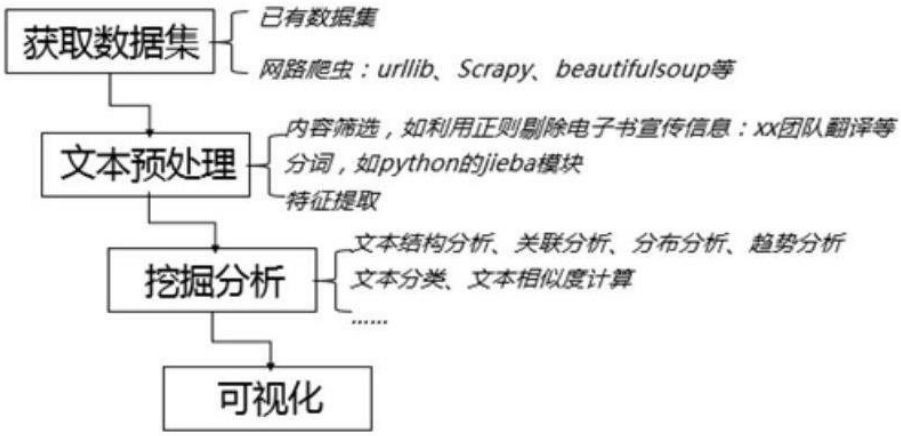


图 1-1 文本挖掘基本流程

1.4.1 文本挖掘的基本流程

- 收集数据

✧ 数据集。

✧ 抓取。这个是 Python 做得最好的事情，优秀的包有很多，比如 scrapy，beautifulsoup 等等。

- 预处理（对这里的高质量讨论结果的修改，下面的顺序仅限英文）

1.去掉抓来的数据中不需要的部分，比如 HTML TAG，只保留文本。结合 beautifulsoup 和正则表达式就可以了。pattern.web 也有相关功能。

2.处理编码问题。没错，即使是英文也需要处理编码问题！由于 Python2 的历史原因，不得不在编程的时候自己处理。英文也存在 unicode 和 utf-8 转换的问题，中文以及其他语言就更不用提了。这里有一个讨论，可以参考，当然网上也有很多方案，找到一个适用于自己的最好。

3.将文档分割成句子。

4.将句子分割成词。专业的叫法是 tokenize。

5.拼写错误纠正。pyenchant 可以帮你!

6.POS Tagging。nltk 是不二选择，还可以使用 pattern。

7.去掉标点符号。使用正则表达式就可以。

8.去掉长度过小的单词。len<3 的是通常选择。

9.去掉 non-alpha 词。同样，可以用正则表达式完成\W 就可以。

10.转换成小写。

11.去掉停用词。Matthew L.Jockers 提供了一份比机器学习和自然语言处理中常用的停词表更长的停词表。中文的停词表可以参考这个。

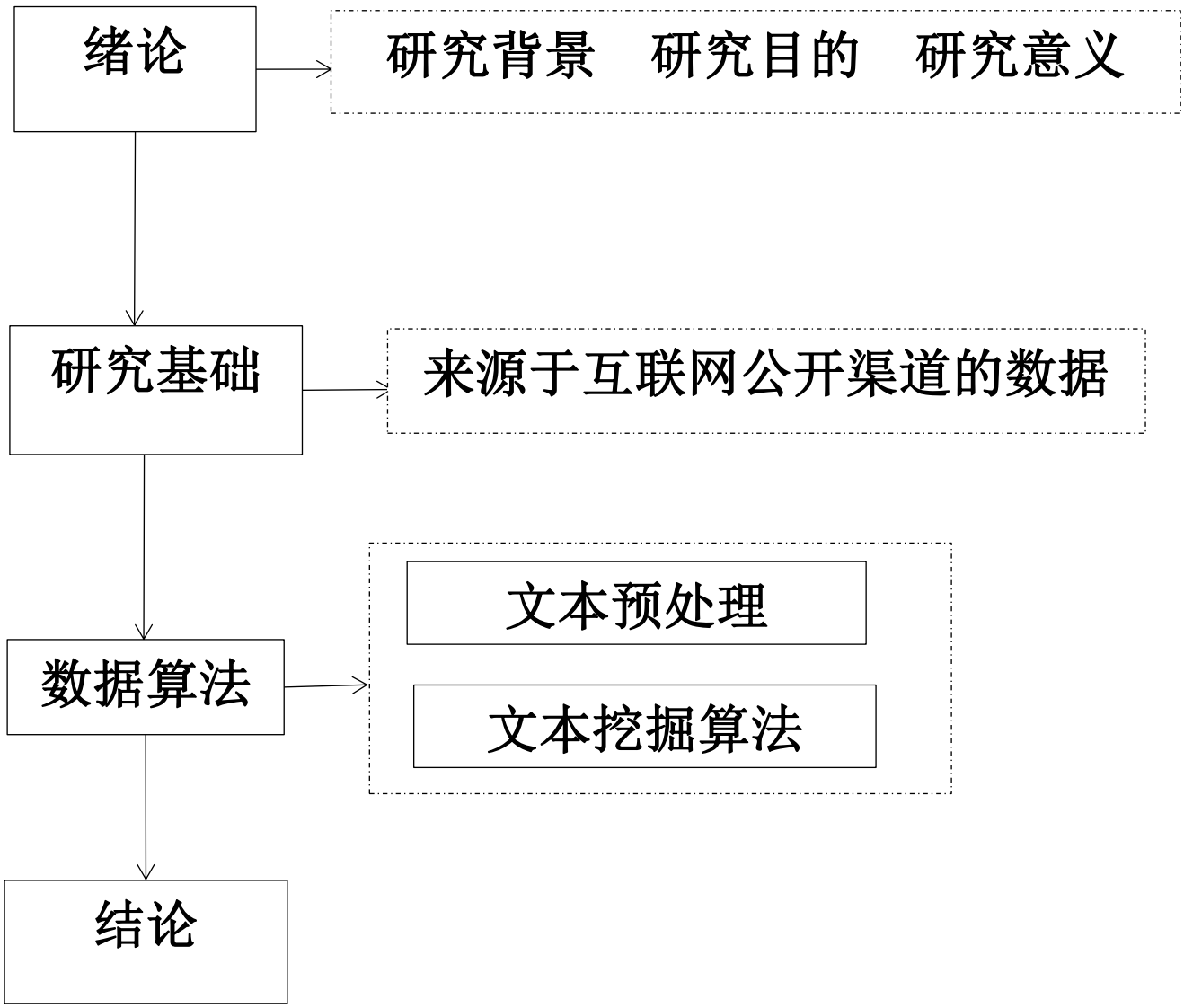
12.lemmatization/stemming。nltk 里面提供了好多种方式，用 wordnet 的方式，不会出现把词过分精简，导致词丢掉原型的结果，如果实在不行，也用 snowball 吧，用 porter，porter 的结果可能会不知道是什么词。MBSP 也有相关功能。

13.重新去掉长度过小的词。是的，再来一遍。

14.重新去停词。上面这两部完全是为了更干净。

15.到这里拿到的基本上是非常干净的文本了。如果还有进一步需求，还可以根据 POS 的结果继续选择某一种或者几种词性的词。

1.4.2 研究框架



第二章 分析方法与过程

2.1 基于 F-Score 的特征评价准则

F-Score 是度量特征在不同类别间的区分度的一种指标，F-Score 值越大，代表该特征在不同类别之间的区分度越强。假设 X_k 代表数据集中的样本(k=1,2,.... N).n+为正类样本的数量,n 为负类样本的数量，则数据集中第 i 个特征的 F-Score 可由

$$F_i=\frac{\left(\bar{x}_i^{(+)}-\bar{x}_i\right)^2+\left(\bar{x}_i^{(-)}-\bar{x}_i\right)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)}+\bar{x}_i^{(+)}\right)^2+\frac{1}{n_--1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)}-\bar{x}_i^{(-)}\right)^2}$$

计算得到,式中:口 $_i$ 表示该特征在整个样本集上的平均值，口 $^{(+)}{}_i$ 和表示该特征在正类样本上的平均值,口 $^{(-)}{}_i$ 表示该特征在负类样本上的平均值。$x^{(+)}{}_{k,i}$ 表示第 k 个正类样本在第 i 个特征上的值，$x^{(-)}{}_{k,i}$ 表示第 k 个负类样本在第 i 个特征上的值。

2.2 算法性能的评价指标

准确率(Accuracy)、召回率(Recall Rate)是信息检索和统计学分类

$$Accuracy=\frac{TP+TN}{TP+TN-FP+FN}$$

$$Recall=\frac{TP}{TP-FN}$$

$$F1=\frac{2TP}{2TP+FP-FN}$$

领域中常用的两个度量指标，F_1值(F1-Measure)能够综合考虑这两个指标因此本文采用以上三个指标来度量算法的性能，计算方法如以下公式所示:

其中 TP 代表正确分类为正样本的样本数量,TN 代表正确分类为负样本的样本数量，FP 代表错误分类为正样本的样本数量,FN 代表错误分类为负样本的样本数量。

2.3 算法描述

假设样本数据集为 $D=(X_1,X_2....,X_N)$，原始特征集合为 $F=(f_1,f_2，...f_d)$，基于 F-Score 的特征选择方法具体描述如下:

输入:样本数据集 $D=(X_1,X_2....,X_N)$,其中 $X_i=(X^1{}_i,X^2{}_{i,}...,X^d{}_i)$,特征集合为 $F=(f_1,f_2,...，f_d)$.

输出:最优特征子集 S.

Step1：初始化:F←-.“包含所有特征的初始集合”;已选特征子集 S← ϕ；

Step2：特征初始评估:对于每一个特征 f∈F,基于 F-Score 统计特性对其进行评价，然后根据每个特征的评价值进行降序排序，构成的特征集合记做 F’；

Step3：第一个特征的选择找到评价值排序最靠前的特征 f_{max},

Step4：取 F 中的下一-个特征，如果为空，则算法停止，否则执行下一步.

Step5:以分类器的分类效果为判据 J,从候选集合 F’中以评价值排序为基准选择特征。假设当前已选特征子集 St,当前的判据值为 Jt,依序从候选集合 F’选中特征 f_i 加入 St,即 $S_{t+1}=S_t+\{f_i\}$,加入之后所得判据值为 J_{t+1}.若 $J_{t+1}<Jt$,则从 S_{t+1} 中删除 f_i,并返回上一步否则，保持加入 f_i 后的特征子集，并更新判据值，返回上一步；

Step6：最终所选的特征集合 S 即为最优特征子集。

2.4 问题一：群众留言分类

2.4.1 算法过程

```
import  os
import  os.path
import  codecs
filePaths=[]
fileContents=[]
for  root,dirs,files  in  os.walk('Users/pc/Desktop/附件 2'):
        for  name  in  files:
                filePath=os.path.join(root,name)
                filePaths.append(filePath)
                f=codecs.open(filePath,'r','utf-8')
                fileContent=f.read()
                f.close()
                fileContents.append(fileContent)

import  pandas
corpos=pandas.DataFrame({
        'filePath':filePaths,
        'fileContent':fileContents
})
import  jieba
segments=[]
filePaths=[]

for  index,row  in  corpos.iterrows():
        filePath=row['filePath']
        fileContent=row['fileContent']

segs=jieba.cut(fileContent)
for  seg  in  segs:
                segments.append(seg)
```

```
                filePaths.append(filePath)
segmentDataFrame=pandas.DataFrame({
        'segment':segments,
        'filePath':filePaths
})

print(segmentDataFrame)
```

2.5 问题二：热点问题挖掘

2.5.1 算法过程

```
import  os
import  os.path
import  codecs
filePaths=[]
fileContents=[]
for  root,dirs,files  in  os.walk('C:\\Users\\asus\\Desktop\\2020.txt
.txt'):
        for  name  in  files:
                filePath=os.path.join(root,name)
                filePaths.append(filePath)
                f=codecs.open(filePath,'r','utf-8')
                fileContent=f.read()
                f.close()
                fileContents.append(fileContent)

import  pandas
corpos=pandas.DataFrame({
        'filePath':filePaths,
        'fileContent':fileContents
})
```

```
import  jieba
segments=[]
filePaths=[]
for  index,row  in  corpos.iterrows():
        filePath=row['filePath']
        fileContent=row['fileContent']
        segs=jieba.cut(fileContent)
        for  seg  in  segs:
                segments.append(seg)
                filePaths.append(filePath)
segmentDataFrame=pandas.DataFrame({
        'segment':segments,
        'filePath':filePaths
})

import  numpy
segStat=segmentDataFrame.groupby(
        by='segment'
)['segment'].agg({'计数':numpy.size}).reset_index().sort(
        columns=['计数']
        ascending=False
)

stopwords=pandas.read_csv(
"路径.txt",
encoding='utf-8',
index_col=False
)
fSegStat=segStat[
~segStat.segment.isin(stopwords.stopword)
]

from  wordcloud  import  WordCloud
import  matplotlib.pyplot  as  plt
```

```
wordcloud=WordCloud(
                font_path='字体路径\simhei.ttf',
                background_color='black'
)
words=fSegStat.set_index('segment').to_dict()
wordcloud.fit_words(words['计数'])
plt.imshow(wordcloud)
plt.close()
```

2.6 问题三：答复意见的评价

由于数据的局限性，本文未对中文开放域对话系统回复质量评价进行深入的探索，但通过本文的实验和讨论，认为可以在中文语料中进行对比实验，只是目前还没有发现非常适合此项任务的大规模中文语料，故本文未针对中文开放域对话系统回复质量评价进行进一步的实验和讨论。

通过一系列的尝试和实验，本文为意见答复系统提供了评价思路，且在实验的过程中构造了有针对性的数据集，通过对深度学习在开放域对话系统回复质量评价中的尝试，也拓宽了现有的评价思路，由此可见，本文对于意见答复系统的发展是有进步意义的。

对话系统任务自产生以来一直都是自然语言处理研究领域的热点问题，这是因为对话系统的实现与其他自然语言处理任务有很大区别，传统的自然语言处理任务的目标主要是通过一些算法和技术来挖掘或学习现有数据中的信息，目前也取得了非常大的成效，语义分析，分词，语法结构分析等任务都已经达到了很高的完成度。而对话系统与这些任务最大的差别，就在于对话生成实际模拟的是人类智慧中最高级的表达形式：演绎。对于人类来说，归纳和演绎是人类智慧的最高表现形式，二者的根本区别在于归纳是一个可以通过学习和不断地训练得到的一种“肌肉记忆”，而演绎则是人类大脑经过复杂的运算和感受，通过将环境对本体的影响进行“内化”之后输出的带有强烈个人色彩的表达形式。所以从传统的自然语言处理任务到对话系统任

务的这段旅途，可以看作是人类在人工智能之路上里程碑式的成就。

在众多生成模型出现的现状下，对话系统回复质量评价就显得格外重要。本论文的研究就着眼于对话系统回复质量评价，通过对不同类型对话系统回复内容的分析和研究，本研究设计了多个非常有针对性的实验，对目前对话系统回复质量评价的方法进行了总结和分析，也大胆的对对话系统回复质量评价的发展进行了探索和尝试。本文通过对任务型对话系统的特点分析，构造了合理的指标和实验过程，并通过不同的特征抽取证明了方法的可行性，在实验中还注意保证了数据来源的多样性，以此来保证得到的实验结果具有一定的普适性。

第三章 结论与讨论

3.1 数据与研究方法的讨论推广

在开展智慧政务研究过程中，本文采用了互联网数据这一新的数据源，同时综合运用了自然语言处理和文本挖掘的方法等多种研究方法由于该数据与研究方法在过去的同类研究中使用较少，本文在研究中也不断进行探索改进，经过实证研究发现，该数据源与研究方法能较好的满足研究需要，下面将就研究中采用的数据及研究方法进行讨论，归纳总结其价值，期望对未来其他相关研究能起到促进作用。

在研究方法上，本文采用了以文本挖掘技术为核心的研究方法，在前人相关研究中采用的方法基础上进行改进和优化，从而一定程度上提升了方法的全面性、精确性和综合性：

（1）全面性。在文本预处理上，过去很多文本挖掘都是基于已有的通用分词词库，在特定行业的文本处理中表现一般，难以识别相关术语，但传统的新词库构建往往需要大量人工识别和标注才能完成，本研究充分利用学术文献中关键词作为新的专业词源，使用一种相对简单的方法完成了专业词库的构建，既减少了工作量，又能更全面的覆盖待挖掘文本的信息；在应用关注度、城市关联度的计算中，类似的指标测度一般是根据词语出现频率、移动窗格中词语的共现性来测算，本研究在此基础上将文本影响力、汉语写作中段落与句子之间的关联性差异纳入权重设定的考虑范畴，提升了权重计算考虑因素的全面性。

（2）精确性。在智慧政务热点的提取上，过去相关研究多采用词频计算的方法，将词频高的词语作为关键词，虽然科学合理，但把每个词语孤立看待，
未从整体和词语关联角度分析，精确性可能存在不足，而本研究运用基于图论的

Ｔｅｘｔｒａｎｋ作为关键词权重计算的依据，能较充分的考虑词语在整篇文章中的地位，相比词频的方法在精确性上有所提升。在智慧政务内容及应用分类的分析上，类似的研究中的做法仅根据共现关系来判断联系，同样缺乏对整个语料文本系统　综合的判断，而Ｗｏｒｄ２ＶｅＣ将每个词语进行向量化表达，对于度量词语之间的内容相关、词语含义的交叠更加的精确，从方法层面上来说精确性有所提升。

（3）综合性。过去的学术论文文本数据的研究方法均较单一，仅利用文本分析计算某个指标进行说明，缺乏可视化及空间分析，而本文在研究中运用了自然语言处理技术，基于此得出结论和建议，更综合的运用研究方法，达到了更全面的研究效果。

参考文献

[1]赵玎，陈贵梧.从电子政务到智慧政务:范式转变、关键问题及政府应对策略[].情报杂志,2013(01):204-207.
[2]多淑金，郭梅.我国智慧政务建设的问题与对策[J]保定学院学报,2015(9):38-43.
[3]于冠一,陈卫东,王倩.电子政务演化模式与智慧政务结构分析[J].中国行政管理,2016(2):22-26.
[4]秦彩杰,管强.一种基于 F-Score 的特征选择方法[J].宜宾学院学报,2018,18(06):4-8.