“智慧政务”中的文本挖掘应用

摘要

随着大数据云计算、人工智能等技术的高速发展,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.如何利用自然语言处理和文本挖掘的方法解决网络留言平台问题是本文的关键.

针对问题一,为了更高迅速、高效率、高质量的处理网络问政平台的群众留言,建立了关于留言内容的一级标签分类模型.我们主要利用基于 SVM—EM 算法的朴素贝叶斯复合智能分类算法,对附件过滤关键字的缺失属性的估计值作为 EM 算法初始值,并计算极大似然估计完成缺失属性的填补,获取适合的最大 EM 收敛值和加速收敛,然后利用朴素贝叶斯分类算法对完整数据进行分类,得到留言内容的一级标签分类.最后使用题目所提示的 F-Score 的评估方法进行留言训练和留言判别试验.最后我们将附件 2 分成了法制政策、城乡建设、民生环境三个一级标签分类.

针对问题二,题目要求我们根据附件 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.首先我们对留言信息数据进行了预处理,减少了数据量.接着采用基于 ICTCLAS 的 Ansj 关键词提取技术,对留言关键词以及相应的权重按照权重由高到低进行人工分析,得到具有跟踪研究意义的话题系列.分析所有话题可能包括的内涵以及在此话题中具有代表性的关键词,为每一个话题建立关键词向量,作为话题规则,从而进一步对所有留言进行话题匹配,完成分类.之后对热点话题的几个较典型的特征建立热点指标,通过评价指标进行热度计算,选出排名前 5 的热点话题.最后我们再将前五的热点话题匹配的所有留言整理在规定格式的表格,便得到了问题二的结果.

针对问题三,针对附件 4 相关部门对留言的答复意见,从答复的相关性、完整性、可解释性等角度针对答复意见的质量给出一套评价方案,并尝试实现.

关键词：贝叶斯分类算法 关键词提取技术 热度评价指标

一、问题重述

1.1 问题背景

近年来,随着微信、微博、市长信箱、阳光热线等各大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,人工处理在数据庞大的情况下,往往效率极低,不能及时提取出热点问题使群众问题得到及时地反映.不过,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.

1.2 目标任务

通过收集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.本文将利用自然语言处理和文本挖掘的方法解决以下相应的问题:

问题一: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.我们需要根据附件2给出的数据,建立关于留言内容的一级标签分类模型,同时将用F-Score对分类方法进行评价检验;

问题二:已知某一时段内群众集中反映的某一问题可称为热点问题,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.我们将根据附件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,最终得到排名前5的热点问题,和相应热点问题对应的留言信息,并将它们按照规定的格式整理成表格.

问题三:针对附件4相关部门对留言的答复意见,我们需从答复的相关性、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、问题分析

2.1 问题一的分析

针对问题一,题目要求我们根据附件2给出的数据,建立关于留言内容的一级标签分类模型.由附件1提供的内容分类三级标签体系，我们了解到首先要按照一定的划分体系对留言进行分类.

我们利用基于SVM—EM算法的朴素贝叶斯复合智能分类算法,对附件过滤关键字的缺失属性的估计值作为EM算法初始值,并计算极大似然估计完成缺失属性的填补,获取适合的最大EM收敛值和加速收敛,然后利用朴素贝叶斯分类算法对完整数据进行分类,得到留言内容的一级标签分类.最后在不同训练集规模,不同特征数量等评估指标下，使用F-Score的评估方法进行留言训练和留言判别试验.

2.2 问题二的分析

针对问题二,题目要求我们根据附件 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.

首先我们对留言信息数据进行了预处理,减少了数据量.接着采用基于 ICTCLAS 的 Ansj 关键词提取技术,对留言关键词以及相应的权重按照权重由高到低进行人工分析,得到具有跟踪研究意义的话题系列.分析所有话题可能包括的内涵以及在此话题中具有代表性的关键词,为每一个话题建立关键词向量,作为话题规则,从而进一步对所有留言进行话题匹配,完成分类.

之后对热点话题的几个较典型的特征建立热点指标,通过评价指标进行热度计算,选出排名前 5 的热点话题.最后我们再将前五的热点话题匹配的所有留言整理在规定格式的表格,便得到了问题二的结果.

2.3 问题三的分析

针对问题三,针对附件 4 相关部门对留言的答复意见,从答复的相关性、完整性、可解释性等角度针对答复意见的质量给出一套评价方案,并尝试实现.

三、模型的假设

1.假设估算目标值时属性之间条件是相互独立的;

2.假定假设空间λ中每个假设有相同的先验概率;

四、符号说明

序号	符号	意义
1	T	训练数据
2	λ	假设空间
3	$P(\lambda)$	λ 的先验概率
4	$P(T)$	待观察训练数据 T 的先验概率
5	$P(T\|\lambda)$	给定 λ 时观察数据 T 对应的概率
6	$P(\lambda\|T)$	λ 的后验概率,即给定训练数据 T 时 λ 成立的概率
7	h_{ML}	极大似然,使 $P(T\|\lambda)$ 最大的假设
8	V_{MAP}	最可能的目标值
9	$\{a_1,a_2,\cdots,a_n\}$	属性值
10	$p(a_1,a_2,\cdots,a_n\|v_j)$	$a_1,a_2,\cdots,a_n$ 的联合概率
11	v_{NB}	NB 分类器所使用的方法
12	$X_1,X_2\cdots,X_n$	原始属性
13	$\lambda=\{t_1,t_2\cdots,t_m\}$	类别属性
14	T_i, T_j	数据子集
15	T_i	记录全部为完整记录,任何属性不含缺失值
16	T_j	为不完整记录,即属性中含有一个及以上的缺失值
17	$P(X_i\|\lambda)$, $P(\lambda\|X_i)$	条件概率,后验概率
18	P_i 和 R_i	第 i 类的查准率和查全率
19	s	浏览量
20	$score$	关键词权重
21	rn , rd , hn , cn	话题的反映数,反映的天数,点击数话,题的评论数.

五、模型的建立与求解

5.1 问题一模型的建立与求解

5.1.1 模型的分析

针对传统方法和朴素贝叶斯在网络留言分类应用中的局限性.对网络留言分类的特点进行了系统分析和研究,提出了一种基于 SVM—EM 算法的朴素贝叶斯复合智能分类算法,该算法充分融合朴素贝叶斯简单高效 EM 算法对不完全数据处理的优点,将对附件过滤关键字的缺失属性的估计值作为EM算法初始值,并计算极大似然估计完成缺失属性的填补,获取适合的最大 EM 收敛值和加速收敛,然后利用朴素贝叶斯分类算法对完整数据进行分类,提高网络留言分类到达精确度和性能.

5.1.2 模型的建立

(一)朴素贝叶斯算法与求解

朴素贝叶斯分类技术以贝叶斯定理为基础,通过数据的先验概率,利用贝叶斯公式计算出其后验概率,并选择具有最大后验概率的类作为该对象所属的类.在给定训练数据T时,确定假设空间λ中的最佳假设,贝叶斯算法提供了从先验概率$P(\lambda)$以及$P(T)$和$P(T|\lambda)$计算后验概率$P(\lambda|T)$的方法,其公式为

$$p(\lambda|T)=\frac{p(T|\lambda)p(\lambda)}{p(T)} \tag{1}$$

其中$P(\lambda)$表示λ的先验概率,$P(T)$表示待观察训练数据T的先验概率,$P(T|\lambda)$表示给定λ时观察数据T对应的概率,$P(\lambda|T)$表示λ的后验概率,即给定训练数据T时λ成立的概率.假设集合λ并寻找观察数据T在其中的概率最大的假设$h\in\lambda$,称为极大后验(MAP)假设.利用贝叶斯公式计算每个待选假设的后验概率,以此来确定 MAP 假设,即

$$h_{MAP}=\arg\max_{h\in H}P(\lambda|T)=\arg\max_{h\in H}\frac{P(T|\lambda)p(\lambda)}{p(T)}=\arg\max_{h\in H}p(T|\lambda)p(\lambda) \tag{2}$$

$P(T)$是不依赖于λ的常量.在某些情况下,可假定λ中每个假设有相同的先验概率(即对 H 中任意λ_i和λ_j,$P(\lambda_i)=P(\lambda_j)$进一步简化,只需考虑$P(T|\lambda)$来寻找极大可能假设.$P(T|\lambda)$常称为给定$\lambda$时数据$T$的似然度.而使$P(T|\lambda)$最大的假设被称为极大似然$h_{ML}$,即

$$h_{ML}=\arg\max_{h\in H}p(T|\lambda) \tag{3}$$

朴素贝叶斯分类器应用的学习任务中,每个实例 x 可由属性值的合取描述,而目标函数 f(x)从某有限集合 V 中取值贝叶斯方法的新实例分类目标是在给定描述实例的属性值$\{a_1,a_2,\cdots,a_n\}$下,得到最可能的目标值V_{MAP},即

$$V_{MAP}=\arg\max_{v_j\in V}P(v_j|a_1,a_2,\cdots,a_n) \tag{4}$$

基于贝叶斯公式重写为

$$v_{MAP}=\arg\max_{v_j\in V}\frac{P(v_j|a_1,a_2,\cdots,a_n)p(v_i)}{p(a_1,a_2,\cdots,a_n)}=\arg\max_{v_j\in V}P(v_j|a_1,a_2,\cdots,a_n|v_j)p(v_i) \tag{5}$$

朴素贝叶斯 NB 分类器算法基于一个简单的假定:估算目标值时属性$\{a_1,a_2,\cdots,a_n\}$之间条件是相互独立的,则观察到$a_1,a_2,\cdots,a_n$的联合概率正好是对每个单独属性的概率乘积,为

$$p(a_1,a_2,\cdots,a_n|v_j)=\prod_i P(a_i|v_j) \tag{6}$$

将其代入式(5)中,可得到 NB 分类器所使用的方法为

$$v_{NB}=\arg\max_{v_j\in V}P(v_j)\prod_i P(a_i|v_j) \tag{7}$$

当所需的条件独立性能够被满足时,朴素贝叶斯分类器输出的(v_{NB})等于 MAP 分类.

朴素贝叶斯分类算法利用贝叶斯定理的优势,在网络留言分类中有广泛应用,是文本分类最为精确的技术.在智能文本分类技术中,通过贝叶斯分类器的“自我学习”智能技术,能有效保护信息的正常通信,过滤垃圾信息的骚扰朴素贝叶斯分类分为以下 3 个阶段:

第 1 阶段:准备工作阶段.收集大量正常留言和垃圾留言作为样本,确定特征属性,并对每个特征属性进行适当划分,然后提取留言样本中主题和信体中的字符串,建立对应的数据库分类,输出特征属性和训练样本;

第 2 阶段:分类器训练阶段.计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计,创建贝叶斯概率库统计出每个字串在网络留言中出现的概率以及在正常留言中出现的概率,然后根据公式计算出留言中含某字串则为垃圾留言的概率;

第 3 阶段:应用阶段.使用训练好的 Bayes 分类器对分类项进行分类,其输入是分类器和待分类项,输出是待分类项与类别的映射关系通过对留言样本的训练,Bayes 分类器可以自动获取垃圾留言的特征,并根据特征的变化,对网络留言进行有效分类.

(二)SVM-EM-NB 算法与求解

首先用优化训练的 SVM 训练留言集解决一个二次规划问题,使学习器得到一个全局最优解,然后把数据集分成完整集和缺失集,计算缺失属性的数据项与完整属性数据项的相关度,取相关度最大的数据项对应的属性作为缺失属性的一个估计值,此估计值作为 EM 算法的初始值,然后执行 EM 算法的两步,完成极大似然估计,用最后估计的值来完成缺失属性的填补,最后用朴素贝叶斯分类算法对完整数据集进行分类.

输入:$T=\{X_1,X_2\cdots,X_n\}$其中, $X_1,X_2\cdots,X_n$为原始属性集, $\lambda=\{t_1,t_2\cdots,t_m\}$为类别属性.

输出:样本 X 的类别

5.1.3 模型一的求解

算法主要步骤为:

步骤 1:把数据集T分为 2 个数据子集T_i和T_j, T_i中的记录全部为完整记录,任何属性不含缺失值; T_j中的记录为不完整记录,即属性中含有一个及以上的缺失值.

步骤 2:调用 EM 算法,完成缺失数据填补.

步骤 3:随机选择 4/5 的样本作为训练集,剩余 1/5 的样本作为测试集,计算训练集样本的先验概率 P(λ)

步骤 4:在假设类条件独立的情况下,根据贝叶斯公式计算条件概率$P(X_i|\lambda)$.

步骤 5:根据式(1)计算后验概率$P(\lambda|X_i)$,输出类别,求出分类准确率.

整个算法可以分为两个部分:“建立模型”与“进行预测”.

其建立模型的伪代码如下:

```
nunAttrValues 等简单的数据从本地数据结构中直接读取
构建几个关键的计数表
for(为每一个实例) {
for(每个属性){
为 numClassAndAttr 中当前类，当前属性，当前取值的单元加 1
为 attFrequencies 中当前取值单元加 1
```

预测的伪代码如下:

```
for (每一个类别) {
for(对每个属性 xj) {
for(对每个属性 xi) {
if(F(xi)小于 m){
出现的次数没有超过阀值，不予考虑
}else {
查 numC1assAndAttr 计算 F(y， xi，xj)
}
查 numC1assAndAttr 计算 F(y， xi)
计算公式中的评价函数，记录下这个类别的评价函数值
}
```

得到了各个类别上的条件概率，再带入贝叶斯公式算出最后的概率.

算法代码实现见附录.

5.2 问题二模型的建立与求解

5.2.1 模型的分析

网络留言数据量大,直接进行分析不仅耗费时间长,工作量大,而且无法及时分析热点.首先我们对留言信息数据进行了预处理,减少了数据量,对后续数据分析提供方便.

通过对留言关键词以及相应的权重按照权重由高到低进行人工分析,得到具有跟踪研究意义的话题系列.这一过程采用了传统的人工选择,是基于机器对于自然语言的理解能力受限考虑.在确定的话题系列基础上,分析所有话题可能包括的内涵以及在此话题中具有代表性的关键词,为每一个话题建立关键词向量,作为话题规则,从而进一步对所有留言进行话题匹配,完成分类.

之后对热点话题的几个较典型的几个特征建立热点指标,选出排名前 5 的热点话题.最后对于每一个留言,遍历其 keywords[m],将其中存储的关键与话题向量中的关键词进行比较,在满足一定数量的相似度后,判定留言是否属于某一话题,若属于,则存储留言时间和内容.

模型的实现整体过程见下图:

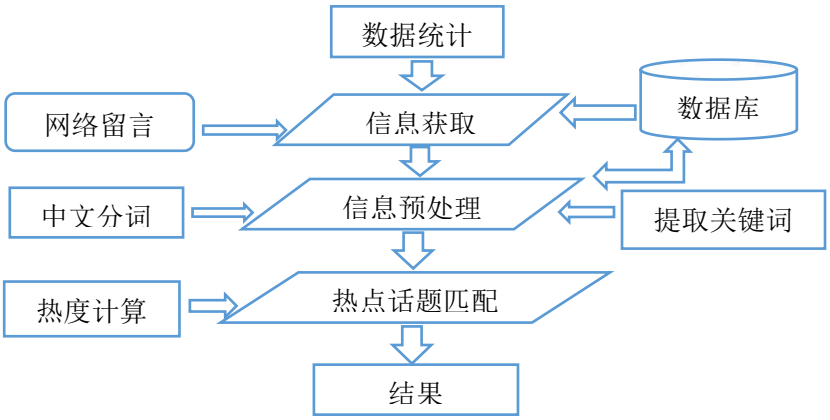

图 1—整体过程流程图

5.2.2 模型的建立

(一)关键词提取

本文采用的是基于 ICTCLAS 的 Ansj 关键词提取技术,其基本原理为依据不同词性词语的初始权重,其中标题中词权重加倍,再结合词在文中出现的位置和频率调整后,得到每个词的权重 score.由于本文需要通过关键词的热度来进一步确定当前的热点话题,故结合留言的浏览量对关键词权重作进一步改进.具体改进的公式如下:

$$\begin{cases} p = \dfrac{s_i}{\sqrt{\sum s_i^{\,2}}} \\ \text{热度}=score * p * 1000 \end{cases} \tag{8}$$

其中, s =浏览量, $score$ =关键词权重.

(二)热点话题提取:

网络留言热点话题的获取是指从大量的网络论坛留言中发现热点话题并利用关键词定位到具体的帖子的过程.本文所采用的热点话题获取方法是基于关键词热度的.热点话题获取是对所有群众留言内容中出现的关键词进行的一个统计,关键词热度越高表明越受网民的关注.这种方法所发现的话题基本能够反映出当前的热点问题,话题发现的过程如图

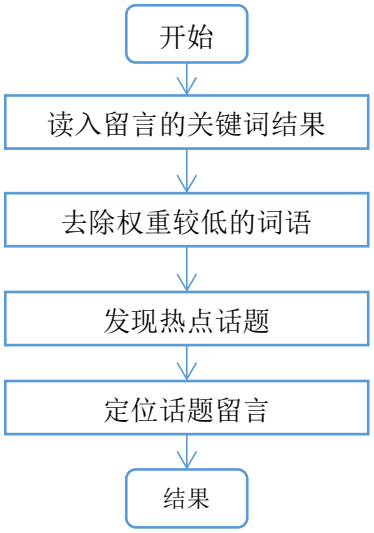

图 2—话题发现过程

(三)热度评价指标:

热点话题一般的特征:话题反映频率高、时间跨度长、反映人数多、评论人数多.

对热点话题的上述特征分别用下面几个不同的参数刻画,其中的时间单元是可以设定的一段时间,如一周、一个月等.

1. $rn(report\ number)$,表示话题在一个时间单元内的反映数.

2. $rd(report\ days)$,表示话题反映的天数.

3. $hn(hitting\ number)$,表示话题的点击数.

4. $cn(commence\ number)$,表示话题的评论数.

rn 和 rd 属于媒体关注度范畴, hn 和 cn 属于网民关注度范畴,分别是留言的点击数和跟评数的加和.

综合话题的媒体关注度和网民关注度,热点话题的关注度用下面的公式来计算:

$$\left(\frac{rn}{RN}*10+\frac{rd}{\theta}\right)+\chi\log_{hn+cn+\omega}(hn+cn) \tag{9}$$

θ 是这段时间的具体天数值,在实际应用中,这个值可以随应用的需要而设定, ω 为动态调整因子,使对数值是一个非常数 1 的变动值.此处 x 用来调整媒体关注度和网民关注度在整个公式中的比重.

5.2.3 模型的求解

根据上面建立的模型,我们提取了热点问题,并根据热点评价指标计算公式找到排名前五的热点问题,最后提取相关话题,具体步骤如下:

Step1:数据预处理,得到每个词的权重 score,按照从高到低进行人工分析.

Step2:通过热点评价指标公式进行话题热点计算,参考值对热点问题进行排名.

Step3:按照数据预处理模块中对留言信息进行的分词、关键词生成结果,依次把词语读入暂存数组 keywords[m]中,其中 keywords[m]是一个字符型的一维数组,用来暂时存放某一条留言的关键词结果.将每个话题中出现频率较高的关键词依次读入话题向量中作为话题匹配规则.依据话题匹配规则,对于每一个留言,遍历其 keywords[m],将其中存储的关键与话题向量中的关键词进行比较,在满足一定数量的相似度后,确定其是否属于需要研究的五个话题之一,若是,则将留言的发帖时间、浏览量等信息存入数据库中.

5.3 问题三模型的建立与求解

5.3.1 模型一的建立

数组匹配

我们利用前面的关键词提取,存储了答复意见中的关键词,并存储于数组 keywords[n],keywords[m]已经存储了热点问题的关键词信息,遍历其 keywords[m],将

其中存储的关键与 keywords[n]的关键词进行比较,在满足一定数量的相似度后,根据匹配到的关键词占比来评价答复意见.

5.3.2 模型一的求解

利用 JS 语言进行求解

我们这里分别抽取了 2000 份留言与答复,根据程序运行的结果,关键词的数组匹配结果达到了 92.3%,已经能说明答复与留言具有很强的相关性,以及答复的完整性和可解释性,且数据越大匹配程度越高.

自定义数组匹配代码见附录.

六、模型的检验

6.1 问题一模型的检验

实验对留言样本进行了测试,基于 Eclipse 平台,分别使用 Java 语言实现的改进的基于 EM 和朴素贝叶斯算法构建分类器,在不同训练集规模,不同特征数量等评估指标下进行留言训练和留言判别试验.

评估方法：F-Score

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{10}$$

其中 P_i 为第 i 类的查准率, R_i 为第 i 类的查全率.

实验结果如表一所示:

表 1—模型实验结果

训练留言总数/封	改进的朴素贝叶斯查全率/%	改进的朴素贝叶斯查准率	F_1/%
1000	75.2	90.3	72.6
2000	79.1	91.2	76.3
3000	83.4	92.1	79.5
4000	86.9	92.7	80.2
5000	89.6	93.2	82.6
6000	91.9	93.9	83.2
7000	93.6	96.4	86.5

该实验是基于朴素贝叶斯算法和改进的基于 SVM-EM 朴素贝叶斯复合算法进行的,从表 1 可以看出,训练集的大小对查全率及精确度都有较大的影响.一般来说,训练集中留言样本越多,网络留言分类的性能越好,稳定性越强.

实验结果表明,本文实现的改进朴素贝叶斯复合智能算法,不仅能够快速得到最优分类特征子集,而且大大提高了其学习和分类的效率和准确率,降低了分类器的错误率,具有较高的实用价值.在训练集达到 5000 之后,查全率达到了 93%以上,F-Score 特征值也越来越大,说明该特征辨别能力强,远高于朴素贝叶斯算法,这也正说明了其相对朴素贝叶斯算法的优越性.适当增加训练集样本数,改进的贝叶斯复合智能算法将在查全率

和误报率方面有更好的表现.

6.2 问题二模型的检验

对于关键词的提取:

这里我们根据关键词的权重列出了几大类热点问题的总体主题的分类,下面列出集中的三个主题的分布情况,以及每个关键词的分布概率.

图 3—主要分类的关键词分布情况与分布概率图

对于热点指标的计算与最后检索的相关留言,最后下表给出了具体排名和部分相关留言,具体情况见附录:

表 2—热点排名与所属部分留言表

热度排名	时间范围	地点/人群	具体内容	相关留言
1	2019/6/19—2020/1/26	A 市丽发新城小区	环境噪音影响居民生活	修建搅拌站噪音扰民，环境污染
2	2019/4/16—2019/10/21	A 市万科魅力之城	餐馆小区噪音油烟扰民	低层餐馆油烟扰民
3	2019/1/8—2019/7/8	西地省 A 市	A 市 58 车贷案件	严惩 58 车贷大集团诈骗保护伞
4	2019/4/19—2019/10/19	A 市地铁三号线	地铁三号线的拥堵问题	为什么没有直通松雅湖的出入口
5	2019/1/15—2019/11/11	A 市五矿万境 K9 县	开发商与施工方建房质量问题	A 市五矿万境负一楼面积缩水

6.3 问题一模型的检验

选取一定量的留言份数和答复份数,它们都有各自的关键词数组,如果匹配率逐渐升高则说明模型可靠,下表给出具体数据:

表 3—匹配度表格

匹配的份数（留言和答复）	keywords[n]	keywords[m]	匹配到的相似关键词占比/%
1000	78	80	81.2
2000	132	124	84.3
3000	187	197	87.4
4000	245	258	89.6
5000	306	312	93.8
7000	467	487	96.8

最后可以看到,当选取的份数达到一定数量后（7000）,匹配到的相似关键词占比达到了 96%,说明相关部门对留言问题的答复已经在保证相关性和可解释性的前提下做到了完整性.

七、模型的评价

7.1 模型的优点

(1)本文针对传统分类算法对网络留言识别率低的缺点,提出了采用改进的贝叶斯复合智能算法来分类网络留言.实验结果表明,复合智能算法要优于传统的朴素贝叶斯算法,而且最大程度地保持了算法的自动化和智能化,具有很强的适用性和自我学习型,突破了用朴素贝叶斯方法对大规模数据特征提取存在的局限性基于留言样本集的训练总数的增加会使该算法有更高的精确度.

(2)对热点问题进行了一系列的文本挖掘,包括留言信息获取、数据预处理、话题提取、热点问题拟合.在中文分词、关键词提取阶段,能够结合网络论坛数据特点,对相应工具进行了定制改进.在话题提取阶段,如何完成从关键词到话题匹配的过程是一大难点.

7.2 模型的缺点

本文制定了话题匹配规则,存在一定的匹配误差.

7.3 模型的推广与改进

模型可在以下两个方面进行改进:

(1)话题匹配规则不够灵活,可结合语义及自然语言处理作进一步优化;

(2)表达的信息有限,可适当增加可视化表现形式.

参考文献

[1]姜启源、谢金星、叶俊.数学模型(第三版).高等教育出版社,2003
[2]殷风景.面向网络舆情监控的热点话题发现技术研究[D].国防科学技术大学,2010.
[3]马小龙.网络留言分类中贝叶斯复合算法的应用研究[J].佛山科学技术学院学报(自

然科学版),2013,31(02):43-47+68.
[4]胡佳妮,徐蔚然,郭军等.中文文本分类中的特征选择算法研究[J].光通信研究,2005(3):44-46.
[5]HAN Jawei,KAMBER M.数据挖掘概念与技术[M].北京:机械工业出版社,2010,184-211.
[6]刘斌,黄铁军,程军等.一种新的基于统计的自动文本分类方法[J].中文信息学报,2008,16(6):18-24.
[7]朱明,王俊普.一种最优特征集的选择算法[J].计算机研究与发展,2006,35(9):803-805.
[8]邓乃扬,田英杰.数据挖掘中的新方法:支持向量机[M].北京:科学出版社,2004.
[9]巩知乐,张德贤,胡明明.一种改进的支持向量机的文本分类算法[J].计算机仿真,2009.26(7):165-168.
[10]杜树新,吴铁军.模式识别中的支持向量机方法[J].浙江大学学报:工学版,2008,37(5):521-527.
[11]曹丽娜,唐锡晋.基于主题模型的BBS话题演化趋势分析[J].管理科学学报,2014,17(11):109-121.
[12]王允.网络舆情数据获取与话题分析技术研究[D].郑州解放军信息工程大学,2010.
[13]赵旭剑,张立,李波等.网络新闻话题演化模式挖掘软件,2015,36(6):1-6.
[14]张旭,张振江,刘云.BBS舆情系统爬虫模块的研究[J]铁路计算机应用,2010,19(12):18-21.
[15]熊祖涛.基于Web文本信息抽取的微博舆情分析[D].西安:西安科技大学,2013.
[16]赵旭剑,邓思远,李波等.互联网新闻话题特征选择与构建[J].软件,2015,36(7):17-20.