“智慧政务”中的文本挖掘应用

摘　要：在互联网技术飞速发展的今天。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。与此同时，大量的与社情民意有关的数据给一直以来依靠人工来提取热点问题的相关工作带来极大的挑战。因此，运用网络文本分析和数据挖掘技术对政务问题的研究具有重大的意义。

本文参照附件 1 提供的内容分类三级标签体系。对数据进行分类，建立关于留言内容的一级标签分类模型。将附件 3 中的数据分为一级分类、二级分类与三级分类。首先通过特殊字符处理与正则表达式等方法将数据进行数据清洗，再主要利用 jieba 中文分词工具对社情民意反映信息进行分词与词性处理，之后通过去停用词来过滤无意义的词，并通过 TF-IDF 算法处理文件数据，将所有词安 TF-IDF 值大小进行排序，出现次数最多就是该文本中的代表关键词，提取出每个文本中的前五的关键词。这里采用 K-means 算法将留言问题分类，由 K-means 产生一个聚类中心，利用 Knn 算法将留言进行分类。之后再使用模型的方法得到分类的结果且使用 F-Score 算法对其进行评价。

同样采用问题一中的方法对数据行进预处理并加以改进，先提取出热词，再提取热点问题将留言中关键词出现频率作为热度评价指标，利用词频梯度和平滑算法提取热词，进一步通过先找出一些候选的话题词组再利用Attention的思想，从候选词组中找出一个包含的词语更加重要的词组，作为输出话题,再由话题提取算法得出热点问题。对找出的一些候选的热点话题进行筛选，对比候选词组的表现能力分数和出现的频率两方面得出真正的热点话题，再按其热度指数对话题进行排名。

关键词：TF-IDF；K-means 文本聚类；中文分词；Knn 算法

The thesis title

Abstract: Today with the rapid development of Internet technology. Wechat, Weibo, mayor's mailbox, sunshine hotline and other online political platforms have gradually become an important channel for the government to understand public opinion, gather people's wisdom and gather people's spirit. At the same time, a large number of data related to social situation and public opinion have brought great challenges to the work of manually extracting hot issues. Therefore, the use of network text analysis and data mining technology is of great significance to the study of government affairs.

This paper refers to the three-level label system of content classification provided in Annex 1. Classify the data and establish the first level label classification model about the message content. The data in Annex 3 are divided into primary, secondary and tertiary categories. First, the data is cleaned by special character processing and regular expression, then the Chinese word segmentation tool of Jieba is mainly used to segment and part of speech processing the social situation and public opinion information, then the meaningless words are filtered by removing the stop words, and TF-IDF is used to filter the meaningless words The algorithm processes the file data, sorts the TF-IDF values of all words, and the most frequent occurrence is the representative keywords in the text, and extracts the first five keywords in each text. In this paper, K-means algorithm is used to classify the message problem, K-means generates a cluster center, and KNN algorithm is used to classify the message. Then we use the model method to get the classification results and use the F-score algorithm to evaluate it.

In the same way, the method in question 1 is used to preprocess and improve the data. First, the hot words are extracted, then the hot words are extracted. Then, the frequency of the keywords in the message is taken as the heat evaluation index, and the hot words are extracted by using the word frequency gradient and smoothing algorithm. Further, by finding out some candidate topic phrases first, and then using the idea of attention, an included word is found from the candidate phrases As an output topic, the more important phrase of a language is then extracted from the topic extraction algorithm to get the hot topic. Some hot topics are screened, and the real hot topics are obtained by comparing the performance scores and frequency of candidate phrases. Then the hot topics are ranked according to their heat index.

Key words: TF-IDF; K-means text clustering; Chinese word segmentation; KNN algorithm

目录

挖掘目标

本次建模目标是利用收集自互联网公开来源的群众问政留言记录，利用jieba中文分词对群众留言进行分类、K-means聚类的方法，实现以下三个目标：

（1）利用jieba分词将以文本表达的留言数据转化为计算机可识别的结构化信息，结合聚类结果，对留言进行一级标签分类。

（2）根据某一时间段内反映特定地点或特定人群问题的分类，定义合理的热度评价指标，排列选出排名前五的热点问题。

（3）根据相关部门对留言的答复意见的结果，从答复的相关性、完整性和可解释性等角度对答复意见做一个评价。

1. 分析方法与过程

1.1. 总体流程

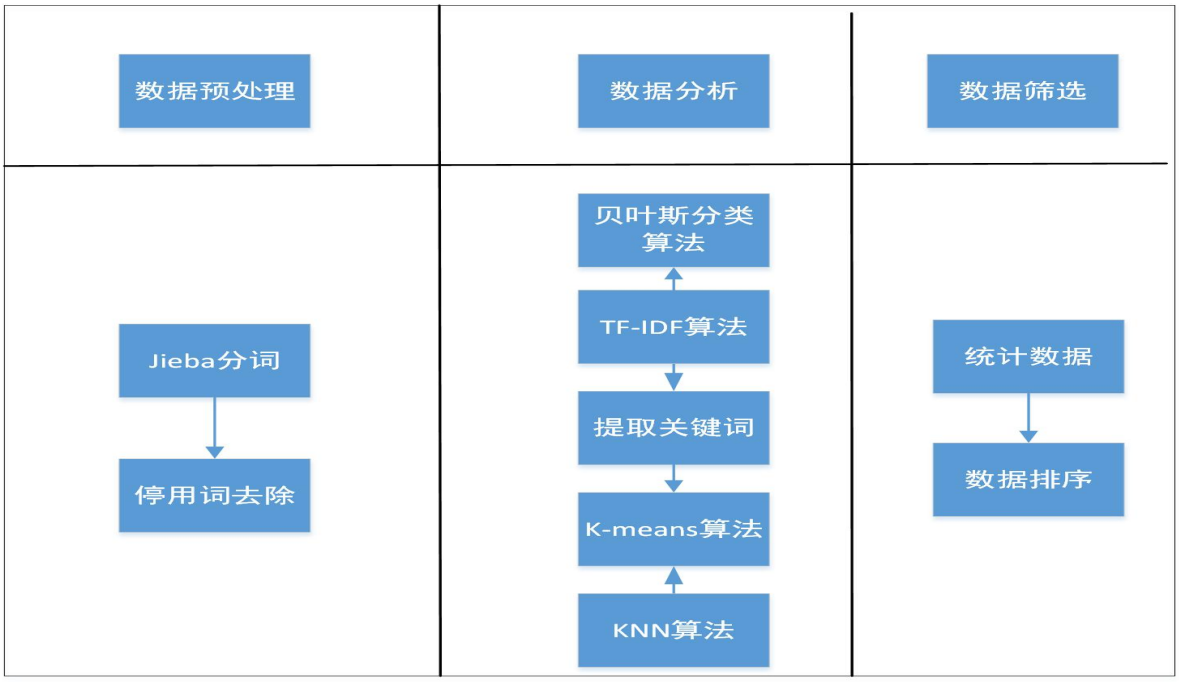

图 1 总流程图

本用例主要包括如下步骤：

步骤一：数据预处理，附件提供的原始数据上进行中文分词，再去除停用词。

步骤二：数据分析，在对留言进行分词后，把这些词转化为向量，以供数据挖掘分析使用。这次使用 TF-IDF 算法，每个留言信息的关键词，把留言信息转化为权重向量，再使用 K-means 对留言进行聚类，再利用 Knn 算法找出与各中心点相似的元素，根据数量多的判定类型。

步骤三：筛选数据，数据预处理好后，产生多个类别，根据每个类别里分类后得到的相同类型

的留言多少，进行排序。

2.2 分析方法与过程

2.2.1 数据预处理

2.2.1.1 对留言简单归类

在附件 2 所给的留言信息中有的留言彼此非常相似，在现实中有的留言信息也会彼此包含，甚至基本一致。考虑到后续操作的简单性，我们对留言信息进行简单的分类时把有着包含关系和相似关系的留言信息进行归类。

2.2.1.2 中文分词

在对留言信息分析之前，需要将附件中以中文文本为表达方式的文本信息换位计算机能够识别的结构化信息，这里采用了 python 的结巴（jieba）分词库进行分词。结巴采用了基于前缀词典实现的高效词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)，同时采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。在文本分词后会产生许多没有意义的词或者符号，需要通过一个停用词表删去这些词和符号。

2.2.1.3 对属性约简 TF-IDF 算法

分词后出现频率高的词对留言的内容影响大，出现频率低的词对留言的影响较小，所以出现频率低的这些词可以约简。TF-IDF 算法原理如下：

（1）计算词频，即 TF(Term Frequency)

词频（TF）= 关键字在文本中出现的频率

$$\text{公式：}\ \mathrm{TF}=\frac{\text{某一个关键词在文本中出现的次数}}{\text{文本总词数}} \quad （1）$$

（2）计算逆向文件频率，即 IDF(Inverse Document Frequency)

逆向文件频率 (IDF) ：总文件数目除以包涵有某一特征词的文件，就等得到该词语的 IDF。当词语的 IDF 越小时，说明词语不能被很好区分类别，反之能够被很好的区别类别。在 IDF 的计算中需要一个语料库，建立一个模拟环境。

$$\text{公式：}\ \mathrm{IDF}=\log\left(\frac{\text{语料库的文本总数}}{\text{包含改词的文本数}+1}\right) \quad （2）$$

（3）计算 TF-IDF，即 TF-IDF=TF * IDF

$$\text{公式：}\ \mathrm{TF}-\mathrm{IDF}=\text{词频（TF）}*\text{逆文档频率（IDF）} \quad （3）$$

TF-IDF 值越大表示词语在文中出现次数越多，也表示更重要。将所有词按照 TF-IDF 值大小进

行排序，出现次数最多就是该文本中的代表关键词，提取出每个文本中的前五的关键词。

2.2.1.4 确定权重向量

生成权重向量的具体步骤如下：

第一步：通过使用 TF-IDF 算法来找出每条留言描述的前 5 个关键词；

第二步：把每条留言描述所提取的 5 个关键词，合并成一个集合，计算每条留言描述对于这个集合中词的词频，如果没有则记为 0；

第三步：生成每条留言描述的 TF-IDF 权重向量，计算公式如下：

TF-IDF=词频（TF）×逆文档频率（IDF） （4）

2.2.2 问题 1

2.2.2.1 解题思路

在附件二中选取百分之八十的数据作为训练数据，首先对数据进行提取和清洗，通过结巴分词得到结构化的数据，再除去停用词得到新的分词。采用 TF-IDF 算法对分词进行词频计算，得到能代表每一条留言的关键词，通过 K-means 算法把留言分成七个类别，即城乡建设、环境保护、交通运输、教育文本、劳动和社会保障、商贸旅游和卫生计生七个一级标签。产生七个聚类中心后，通过 knn 算法对附件二中剩余的百分之二十的测试数据进行分类，通过 $\mathrm{F}-Score$ 算法对分类结果进行检验。

2.2.2.2 K-means 算法

生成留言问题的 TF-IDF 权重向量后，根据每个留言的 TF-IDF 权重向量，对留言进行分类。这里采用 K-means 算法将留言问题分类。

K-means 聚类的原理如下：

（1）样本集合：

$\mathrm{X}=\{x_1,x_2,...,x_N\}$，N 为样本总数

（2）k 个聚类的初始中心点为 $\{z_1,z_2,...,z_K\}$，$\mathrm{n_i}$ 代表第 i 类的样本数

$$\mathrm{Zi}=\frac{1}{\mathrm{n}}\sum_{\mathrm{x}\in\mathrm{ai}}\mathrm{x} \quad (5)$$

（3）定义 k 个类别为 $\{a_1,a_2,...,a_K\}$

（4）定义目标收敛函数如下所示，表示样本点到各自所述类别中心的距离的平方和

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} d_{ij}(z_i, x_j) \quad (6)$$

K-means 聚类算法一般步骤为：

1、从$\{z_1, z_2, ..., z_K\}$中随机取 K 个元素作为初始聚类中心点。

2、计算剩下的样本点到各类中心点的距离值，依次投入到距离最小的别类中去。

3、根据聚类结果，重新计算新的聚类中心$\{z_{1(y)}, z_{2(y)}, ..., z_{K(y)}\}$，y 表示第 y 次迭代。

4、将所的样本安新的中心重新聚类。

5、重复上述步骤，直至聚类中心不再变化。

6、输出结果。

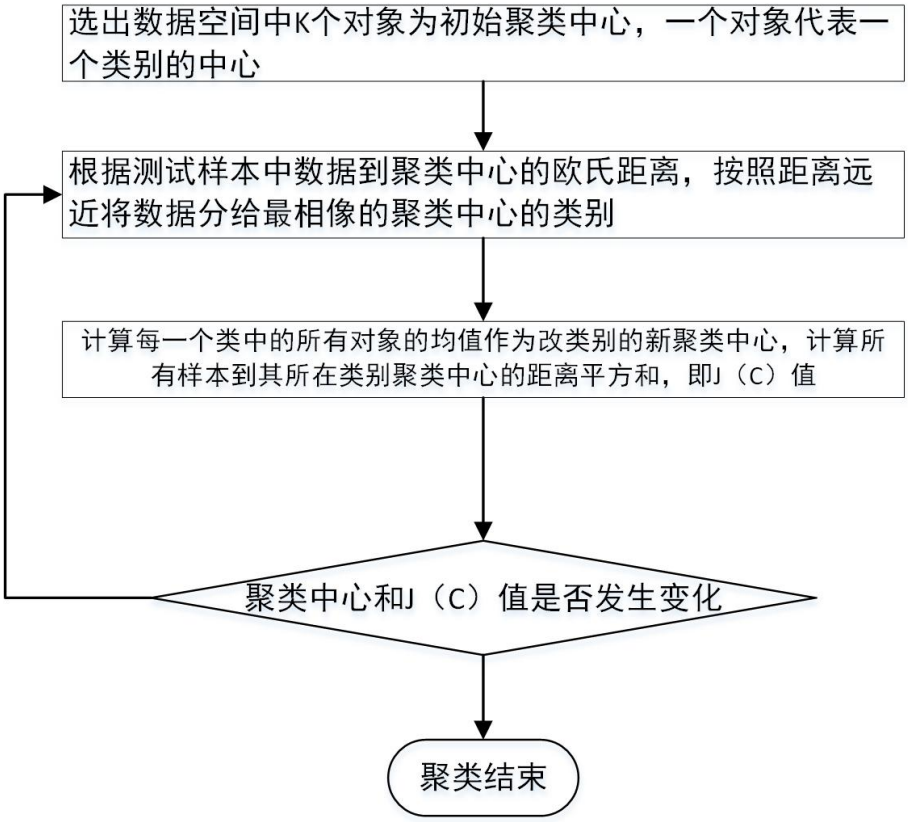

图 2 聚类流程图

2.2.2.3 Knn 最邻近分类算法

Knn 算是通过计算测试样本到训练样本的距离进行排序，选出距离最小的聚类中心点，并将其中的 K 个样本点进行投票分类， 具体算法步骤如下：具体算法步骤如下：（1）对训练样本和测试样本进行分词处理以及去除停用词后，采用 TF - IDF 算法形成权

重向量的集合。

（2）余弦相似度是通过训练样本和测试样本中特征词的向量集合来计算的，公式如下：

$$Sim(d,d_i)=\frac{\sum_{j=1}^{n}x_j*y_{ij}}{\sqrt{(\sum_{j=1}^{n}x_j{}^2)(\sum_{j=1}^{n}x_{ij}{}^2)}} \quad （8）$$

（3）依次计算新文本中每类的权重，公式如下：

$$W（d,C_j）=\sum_{d_i\in KNN}sim(d,d_i)y(d_i,c_j) \quad （9）$$

其中，d 为新文本的特征向量；可从公式 8 中得到公式 $sim(d,d_i)$；

$$y(d,C_j)=\begin{cases}1,d_i\in C_j\\0,d_i\in C_j\end{cases} \quad （10）$$

公式（10）中 $y(d_i,C_j)$ 为类别的指示函数。

（4）比较每一个类的权重，按权重大小判断文本的类别。

2.2.2.4 F-Score 模型检验

将分类好的留言数据得到一级标签和附件二中给出的一级标签进行比较，计算出正确率。计算好的数据带入 F-Score 模型进行检验，公式如下：

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

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的全差率。

2.2.3 问题 2

2.2.3.1 解题思路

根据对网上热点问题的分析和研究，发现热点问题存在以下几个特点：（1）广泛性：热点问题可能存在于生活中的任何一个方面。（2）复杂性：民众的留言建议中可能存在大量的干扰因素，分析过程中需要人为的辅助判断。（3）突发性：短时间内某一问题的留言建议突然爆发，快速增长，不可被预见。（4）集中性：某一问题在短时间内反复出现，且投诉人和地区集中。

针对以上特点，需要在问题 1 中的数据处理的基础上加以改进，先提取出热词，再提取热点问题，方法的整体流程图如下：

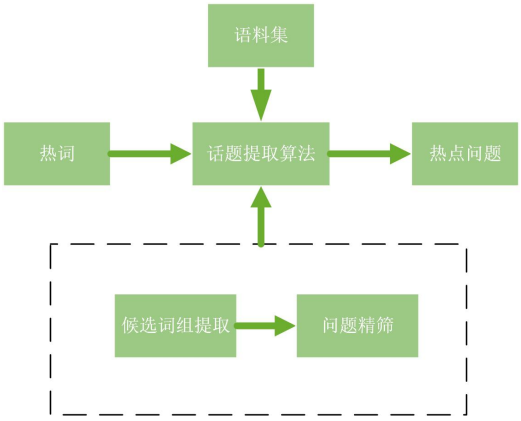

图 3 热点问题挖掘流程图

2.2.3.2 热词提取

利用词频梯度和平滑算法提取热词，热词会受到很多干扰因素的影响，如下图所示：

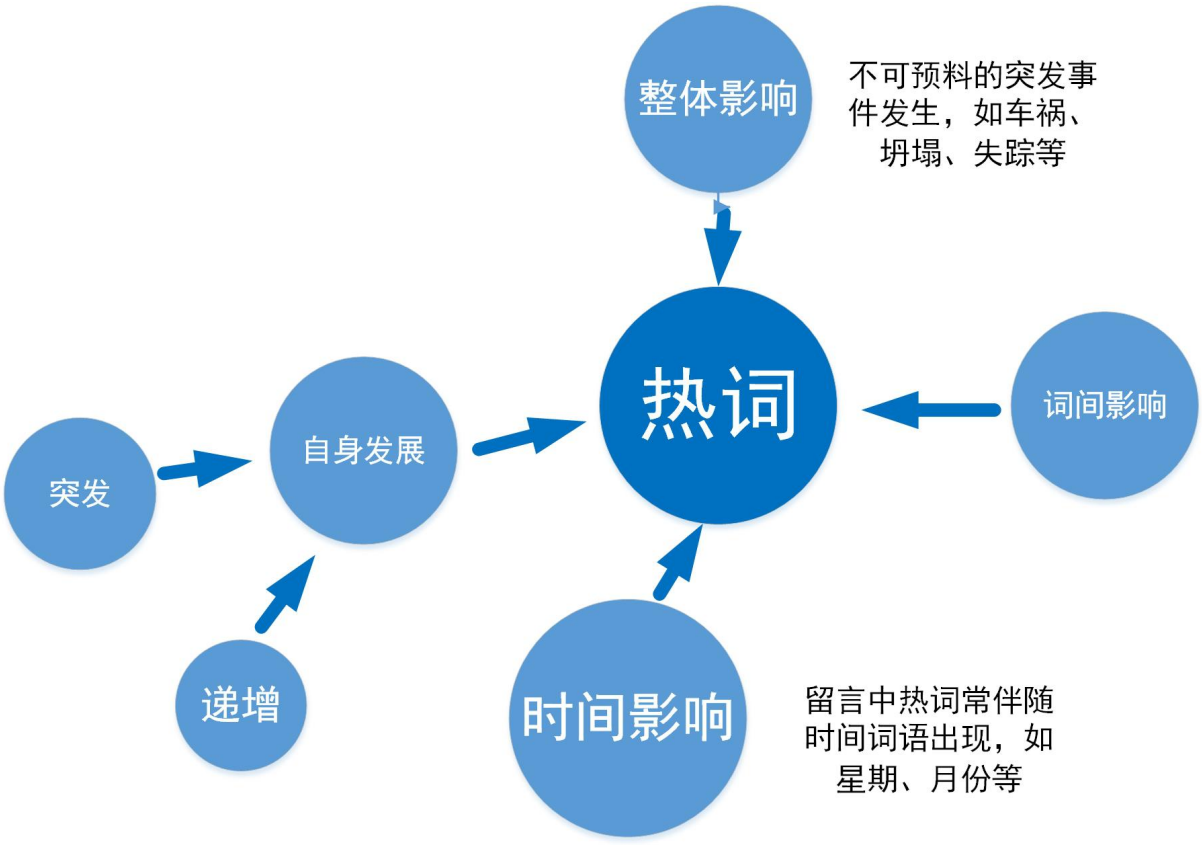

图 4 干扰因素

整体影响：因为一些不可预料的事件产生，留言数据会有一个随时间的不确定波动。

词间影响：受一些地名、人名或者特殊词语影响，使一些不相干词成为热词。

时间影响：从一些周期性时间词的影响，使得“周一”“早上”“明天”等词变成热词。

自身影响：一些事件发生导致相关词的突发性、递增性等的增长。

由于这些词干扰，我们需要对热词进行筛选

1. 数据预处理：同问题 1 处理方法对文本进行分词、去除停用词等工作。

2. 梯度：词频增量的主要衡量指标

$$S(w_i)=\frac{F(w_i,T_j)}{F(w_i,T_1,T_{2,\dots},T_j)} \quad (11)$$

其中，w_i 表示某一词语，T_j 表示时间窗口，$F(w_i,T_j)$ 表示词语 w_i 在时间窗口 T_j 的出现频数。$S(w_j)$ 表示某一词语目前的热度分数。

3. 贝叶斯平均：一种利用外部信息来评估总体均值的方法，尤指一种已存在的信息。

$$\bar{x}=\frac{C*m+\sum_{i=1}^{n}x_i}{n+C} \quad (12)$$

用户评分排名为例子，参与评分的总人数偏少时会造成结果的分数不够客观有效。如果再增加一些外部数量，假设外部有人数为C 的用户也参加了评分，都评了分数得到平均分 m。将得到评分加入原来的评分中，两者求平均分，这时得到的结果就会比单单计算原来用户得到的结果更加客观具有更高的价值。贝特斯评价的结果是受评分人数的影响的，当人数偏少时得到的结果偏向平均分，当人数增多时，结果就会更偏向于算数平均。

4. 热度分数计算：通过贝叶斯平均数值对梯度分数进行改修。

$$\text{修正分数}=\text{平均分}+\frac{\text{词频}}{\text{词频}+\text{平均词频}}\times\left(\text{梯度分数}-\text{平均分}\right) \quad (13)$$

其中，其中，平均词频是公式（12）中的C，平均分公式（12）中的m。

5. 共现模型：通过进一步筛一些互为共现词的热词。

共现词的发现由 word2vector 的方法计算，通过对热词的共现词的筛选，避免信息的重复，能够找到更重要的热词。

6. 时间序列分析：词频的时间序列分析，可以知道热词的平均时间出现密度，可以筛选出一些更具有价值的热词。

2.2.3.3 话题提取

提取出热词，进步提取出热题，步骤如下：

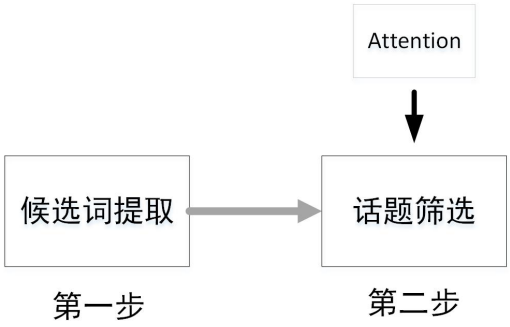

图 5 话题提取步骤

（1）内部聚合度——互信息

信息熵是用来衡量一个随机变量出现的期望值，一个变量的信息熵越大，表示其可能的出现的状态越多，越不确定，也即信息量越大。

$$H = -\sum_{i-1}^{n} p_i \log p_i \quad (14)$$

互信息可以说明两个随机变量之间的关系强弱。定义如下：

$$I(X;Y) = \int_X \int_Y P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)} \quad (15)$$

公式（14）（15）变形可得：

$$I(X;Y) = H(Y) - H(Y|X) \quad (16)$$

公式（16）为Y的不确定性。X已经知道的情况下，Y的不确定性变成条件熵。Y的不确定性减小的量由X引入得知。这个量越大，表示X的出现，使得Y出现的不确定性减小，即Y有很大的概率出现，也代表了X、Y的关系越亲密。

（2）话题筛选

某一热词被挑选出来后，在不同的语义语境中代表不同的意义，表达的信息也不同。比如留言编号 188416 提取出来的词是“卫生”、“乡村卫生室”、“卫生局”、“卫生许可证”、“执业许可证”、“无证行政”、“乡村”，但是“乡村卫生室”、“卫生许可证”、“执业许可证”等中的“许可证”、“卫生”、“乡村”等词在其他情况下也经常出现，指向性不强。所以这里需要对这些候选的话题进行筛选。

如何找出重要的词语是筛选的关键和主要目的，其方法和思想核心与 Attention 机制一样。比如：“许可证”和“卫生”比“乡村”和“卫生”搭配、搭配更好，包涵信息更多，更合适。

Attention 通过 TF - IDF 算法来确认，通俗来说就是各个词组在词组中表现出来的特异性。热词候选词组能力分数 S 表达公式如下：

$$\text{Score（s）}=\frac{\sum_{i=1}^{N}\frac{Corpus\left(w^{h},w_{j}\right)}{Corpus(w_{j})}}{N} \quad （17）$$

其中，N 表示候选词组中的词语数量和候选词组中包含的第 i 个词语，Corpus (w) 表示在此语料库中含有相关语料词语 w 。

当词组出现频率越高，出现次数越多，也就代表事件越重要。总体来说，我们过对候选词组的表现能力分数和出现的频率的共同抉择，筛选出真正的热点问题。

2. 结论

实现热点问题的文本挖掘具有很大的现实意义，能够让群众的建议得以有效呈现在相关工作人员面前，同时能够过滤到很多重复和无意义的留言建议，节约大量时间和人力。热点问题的自动挖掘避免了很多重复无疑的筛选工作，又能使群众关心的问题被重视和发现，能够得到及时的回复和解决。

3. 参考文献

[1]鞠冬彬,赵宪佳.一种改进的最邻近分类算法[J].信息通信,2018(12):5-7.

[2]刘培磊,唐晋韬,王挺,谢松县,岳大鹏,刘海池.基于词向量语义聚类的微博热点挖掘方法[J].计算机工程与科学,2018,40(02):313-319.

[3]姜玉坤. 舆情热点信息挖掘技术的研究与应用[D].天津大学,2017.

[4]魏会建. 基于属性约简和属性加权的朴素贝叶斯分类算法的研究[D].吉林大学,2014.

[5]郑瑞娟,张仰森.基于概念的 Web 文本分类方法及实现[J].北京信息科技大学学报(自然科学版),2013,28(02):77-81.

[6]时志芳. 移动投诉信息中热点问题的自动发现与分析[D].北京邮电大学,2013.

[7] 黄敏.网络舆情热点挖掘算法研究与实现[J].安徽大学学报(自然科学版),2012,36(06):67-72.