第八届“泰迪杯”数据挖掘

C 题：“智慧政务”中的文本挖掘应用

目录

基于大数据技术的“智慧政务”文本挖掘应用

摘要

随着投诉治理力度的不断加强，各地市投诉都在不断减少，作为新公共管理理论发展的重要内容，当代政府回应思想是政府与公众互动的基础。地方政府回应是了解民情，获得民众支持的重要因素，也是政府责任的提现。通过大数据技术对留言分类，发现不少留言问题是可以提前控制的，针对留言问题，提前向相关部门提供预警信息可以大大降低投诉风险和一些社会矛盾的发生。针对客户留言内容中的问题展开研讨，利用大数据技术，整合客户档案，留言信息详情等数据，重点利用准确率达到 81%的逻辑回归模型，辅助以朴素贝叶斯模型和随机森林展开留言层次归类分析。其次用 sklearn 文本聚类找出热点问题，分维度考虑热点问题等级，最后对留言回复建立余弦相似度模型，准确分类客户留言信息检测答复质量优度，统筹协调互动化服务手段提前展开主动服务，从而提升群众的优良感知。

关键词：智慧政务；信息分类；文本挖掘；朴素贝叶斯

Application of Intelligent Government Text Mining Based on Big Data Technology

Abstract

With the continuous strengthening of complaint management, the number of complaints from various cities is decreasing. As an important part of the development of the new public management theory, the contemporary government response thought is the basis of the interaction between the government and the public. Local

government's response is an important factor in understanding people's feelings and gaining public support. It is also a reminder of the government's responsibility. Through the prediction and analysis of big data technology, it is found that many problems can be pre-controlled. The article classifies the number of replies and message fields in the government message board statistically, integrates customer files, accurately predicts hot issues of customers, and carries out proactive services in advance through coordinated and interactive service methods, thus effectively improving customers' good perception and solving the problems.

Key words: smart government affairs; Information classification; Text mining; Naive Bayesian

一 背景

1.1 留言分类背景

如今网络在人们的日常生活中的应用已经日益广泛，人们通过网络得到的最大的便利就是信息的快速交流。而留言板不失为一种网站普遍适用的信息交互方式。通过留言板，可以发布自己的观点，问题，相互交流，增强各个登录用户之间的交流，是有用的信息在互联网上实现快速传递，提高办事效率，因此微信 、微博、12345 政府热线、阳光热线等网络问政平台，逐渐成为了政府的左膀右臂，帮助政府了解民意，汇聚民智，凝聚民气。但是传统的问政平台受理时间长，效率低，受人为影响因素较大，督促较为困难，为了提高群众的优良感知，政府平台需要对留言进行划分整理，体现现代信息技术的优势，对留言问题进行统一，有效，准确的答复。

1.2 留言分类的目的与意义

随着信息技术的日益发展，大数据，物联网，云计算等技术逐渐走进了各行各业，走进了行政机构，相关信息系统为行政公开以及执法规范化创造了好的支撑和帮助，将大数据分析用在行政上面主要是通过挖掘行政大数据的价值，保证科学的管理、服务科学决策、推进科学发展。在行政方面正确的应用了大数据分析，更能达到顺应时代潮流的目的，保持与时代步伐一致，从而工作效率也会相应的提高。我们希望通过建模处理后用精炼的语言反映问题，并且能够得到准确，及时有效的答复意见基本反馈，完成智能政务的基本要求。

1.3 留言分类模型的主要方法

本文基于政府的实际需求，完成了政府留言信箱的分类整理。论文首先对政府邮箱留言进行文本处理，分类整理，运用了 python 数据库 jieba 分词，其次是对留言文本进行建模处理，运用 sklearl 库选出测试集和训练集数据，最后根据高斯朴素贝叶斯公式构建模型，并对模型进行优化处理。

1.4 本文的主要内容

运用自语言处理技术对留言区的内容建立模型，分类处理；挖掘热点问题并排序，做出热点问题明细表；对答复意见建立检验模型，提高答复意见的相关性，完整性，可解释性以及及时性。

1.5 本文的结构安排

首先是文本聚类，运用 jieba 分词和高斯朴素贝叶斯公式构建模型；其次要挖掘热点问题，进行人群归类，做出热点问题明细表；对于相关部门给的答复意见做出模型检验，最后根据分析给出处理建议。

二 留言文本分类

2.1 留言分类背景

随着科技的进一步发展，网络成了最普遍的交流方式，许多公众品台为了完善品台服务效果，增强群众体验度，都创建了用户留言板块，但是留言适量多，人工回复效率低，分类耗时长等问题接踵而至，为了提高留言回复效率，降低人工成本，我们对群众的留言建立分类模型，通过模型将问题详情进行分类，不仅能够及时有效的统计问题，还能提高留言回复效率，在一定程度上降低和预防了危险事件的发生，提高了群众体验优度。

2.2 关键词提取

具体方法：根据留言平台的内容中出现的关键词进行提取，将提取的关键字进行分类汇总，并且按照出现次数进行排序，从而获取到所有诉求的关键字信息。

2.2.1 通过最大正向匹配法对留言内容进行分词

最大正向匹配，通常简称为 MM 法，基本思想是以词典为依据，取词典中最长单词为第一个次取字数量的扫描串，在词典中进行扫描（为提升扫描效率，还可以跟据字数多少设计多个字典，然后根据字数分别从不同字典中进行扫描）。例如：词典中最长词为“中华人民共和国”共 7 个汉字，则最大匹配起始字数为 7 个汉字。然后逐字递减，在对应的词典中进行查找。具体流程图如图 1 所示。

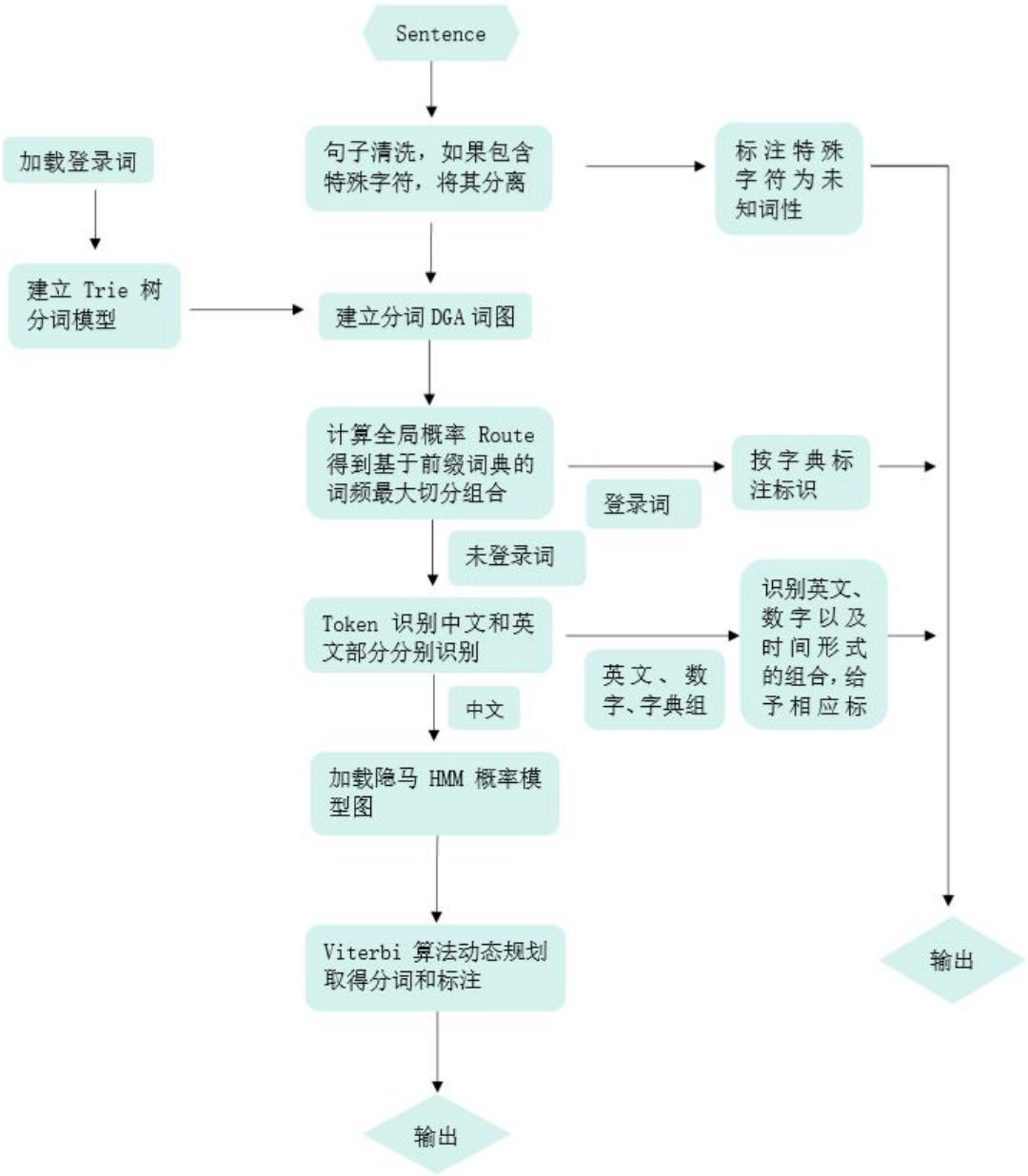

图 1 分词流程图

2.2.2 去除无用词

建立常见词库，并标注常见词库中每个词的词性。由于中文语句中影响句意的主要词性是动词和名次，所以根据 2.2.1 分词结果，标记句子中每个词的词性，只保留名次和动词即可。建立停用词库，该词库中主要包含留言内容的常见无用词以及符号，分词后的句子在停用词中循环查找，如含有停用词中的词语，则删除。2.2.1 示例的处理结果如表 2 所示。

表 2：关键词分步骤提取结果

分词	西湖	建筑	公司	占	道	施工	有	安全隐患
词性	名词	名词	名词	动词	动词	名词	名词	名词
词性筛选	西湖	建筑	公司	占	道	施工	有	安全隐患
去无用词	西湖			占	道		有	安全隐患

2.2.3 查询同义词、近义词

建立同义词近义词词库，通过数据处理匹配后将匹配成为留言类专用词汇，方便后期处理，当词汇中存在非标准的词条时，返回输入词条。如“声音大”经过词库处理后，标准留言词汇为“扰民”。

2.3 文本分类的模型建立

基于大数据挖掘技术的文本分类模型，通过分析历史留言详情，问题答复等信息，利用分布式计算，数据挖掘等技术，运用朴素贝叶斯模型和 logistics 回归模型对留言内容进行分类模型处理，选取准确率最高的模型为检验模型，分析。归纳留言反映问题，建立留言标签，划分留言等级，并展开有针对性的服务，增大问题解决的效率。

2.3.1 朴素贝叶斯算法

（1） 朴素贝叶斯算法基本原理

朴素贝叶斯方法是在贝叶斯算法的基础上进行了相应的简化，即假定给定目标值时属性之间相互条件独立。也就是说没有哪个属性变量对于决策结果来说占有着较大的比重，也没有哪个属性变量对于决策结果占有着较小的比重。

（2） 朴素贝叶斯基本方法

朴素贝叶斯分类（NBC）是以贝叶斯定理为基础并且假设特征条件之间相互独立的方法，先通过已给定的训练集，以特征词之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入 X 求出使得后验概率最大的输出 Y。朴素贝叶斯公式计算如下图所示：

$$\mathrm{P}(Y_K|X) = \frac{P(X|Y_K)P(Y_K)}{\sum_K P(X|Y = Y_K)P(Y_K)}$$

（3）朴素贝叶斯算法结果

第一阶段——准备工作阶段，这个阶段的任务是为朴素贝叶斯分类做必要的

准备，主要工作是根据具体情况确定特征属性，并对每个特征属性进行适当划分，然后由人工对一部分待分类项进行分类，形成训练样本集合。

第二阶段——分类器训练阶段，这个阶段的任务就是生成分类器，主要工作是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，并将结果记录。其输入是特征属性和训练样本，输出是分类器。

第三阶段——应用阶段。这个阶段的任务是使用分类器对待分类项进行分类，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。第四阶段——模型检验阶段，这个阶段的任务是根据朴素贝叶斯公式，不断完善模型准确率，找出分词效率最高运算方法。最终获得随机森林模型准确率为 65%。

2.3.2 随机森林

（1）随机森林算法基本原理

随机的从原始数据集中抽取 m 个子样本，在训练每个基学习器的时候随机选取 k 个特征，从这 k 个特征中选择最优特征来切分节点，从而更进一步的降低了模型的方差。

（2）随机森林算法基本方法

输入：训练数据集$\mathrm{D}=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$,样本子集的个数 T

输出：最终的强分类器 f(x)

从最开始的留言中集中随机抽取 m 个样本，得到训练集D_t，用训练集D_t训练一个 CART 决策树，该模型预测的最终类别为该样本点所到叶节点中数量最多的类别。

（3）随机森林算法结果

通过随机森林得到模型准确率为 78%。对于大数据时代的大样本情况下，随机森林的训练速度有一定的优势优势，并且训练可以高度并行化，所以我们通过随机森林的方法对留言问题进行分类，随机森林随机选择的样本子集大小 m 越小模型的方差就会越小，但是偏差会越大，所以在本案例中，我们通过交叉验证的方式来调参，从而获取一个合适的样本子集的大小。最终获得随机森林模型准确率为 78%。

2.3.3 Logistic 回归模型

（1） Logistic 回归模型基本原理

逻辑回归（Logistic Regression, LR）虽然带有‘回归’二字，但是逻辑

回归却属于分类算法。逻辑回归可以进行多分类操作，但由逻辑回归算法本身性质决定其更常用于二分类。逻辑回归假设数据服从伯努利分布，通过极大似然函数的方法，运用梯度下降或牛顿法来求解参数，来达到将数据二分类的目的。

（2） Logistic 回归模型基本方法

逻辑回归的假设函数为：$h_\theta = g(X\theta) = \frac{1}{1+e^{-X\theta}}$ ，其中 X 为样本输入$h_\theta(x)$为模型输出，θ为要求解的模型参数，设 0.5 位临界值，模型输出值$h_\theta(x)$在[0,1]区间内取值，因此可从概率角度进行解释：$h_\theta(x)$越接近于 0，则分类为 0 的概率越高；$h_\theta(x)$越接近于 1，则分类为 1 的概率越高；$h_\theta(x)$越接近于临界值 0.5，则无法判断，分类准确率会下降。

（3） Logistic 回归模型算法结果

本案例的逻辑回归模型是基于 sigmoid 函数建立的，公式为：$h(x) = -\frac{1}{1+e^{-x}}$，考虑二分类任务，其输出标记$y \in \{0,1\}$（分为正例和反例），线性回归模型产生的预测值$X = \omega^T + b$，$x$可以分为大于 0，小于 0，和等于零。我们对应上面的 Sigmoid 函数，当$x > 0$时我们判断为正例，$x < 0$判断为反例。单位阶跃函数不连续，因此不可导，这样会对后续的优化造成困难。而 Sigmoid 函数的图像与单位阶跃函数在一定程度上接近。因此用此函数代替。最后用 Logistic 模型进行检验评估，再次完善留言分类模型问题，提高模型整体的准确率。最终获得随机森林模型准确率为 80%。

2.3.4 模型结果

我们经过模型的比对，选取准确度最高的 Logistic 回归模型作为留言文本分类模型。详见表 3：模型准确度表

表 3：模型准确度表

模型	朴素贝叶斯	随机森林	Logisti 回归
准确率	65%	78%	80%

三 热点问题挖掘

3.1 热点问题挖掘基本思想

基于大数据技术的机器学习方法，挖掘出留言问题中的热点问题并定义合理的热度评价指标。本案例首先通过 K-means 算法划分聚类留言内容，实际运用 sklearn.cluster.Kmeans 方法实现聚类应用。其次通过特征工程中的 TF-IDF 算法，把原始数据转换为特征向量，自动提取出关机词汇。最后结合余弦相似性另根据现有的数据创建新的特征，计算标准的词语频率并判断留言的重复性，对相似留言进行归类合并处理，整合出热点问题排名。

3.2 热点问题实际解决办法

3.2.1 热度评价指标的定义

比较受广大群众关注或者欢迎的新闻或者信息,或指某时期引人注目的地方或问题称为热点问题。本文指某一时间段内群众集中反映的某一个问题称为热点问题。在留言案例中我们提出群众活跃度高的相似问题，共性问题，以及关注度高的问题定义为热度高的热点问题。

3.2.2 多维度考虑热点问题

（1）共性问题处理

共性问题是指相似性高，出现次数高的问题。针对此类问题我们利用 K-means 算法结合 TF-IDF 特征工程归类相似问题，结合地点人群划分做出共性问题表，具体表格见表 4： 共性问题表。

4： 共性问题表

热度排名	问题 ID	时间	地点/人群	问题描述
1	1	2019/2/21	A 市 A4 区	58 车贷非法经营，没有人处理
2	2	2019/8/19	A 市 A5 区汇金路五矿万境 K9 县	小区的好多问题没有人负责
3	3	2019/4/11	A 市金毛湾	学生的入学学校没有落实
4	4	2019/9/15	A4 区绿地海外滩小区	高铁离小区太近噪音无法消除
5	5	2016/1/8	C5 市火车站到花明楼和 A8 县火车站的铁路	铁路尽快动工
6	6	2019/1/10	西地省京港澳高速城区	噪音扰民
7	7	2019/1/16	B2 区千亿大道和机场大道	早点启用红绿灯
8	8	2019/3/26	A6 区月亮岛路	沿线架设 110kv 高压线杆对附近居民会有危险
9	9	2020/1/8	A 市三一大道	打通机场北通道，使附近高速畅通
10	10	2019/2/25	A 市长房云时代	房子出现裂缝，质量堪忧

（2）关注度排名

“关注度”一词本身，就是当下的社会民众对日常生活的一种思考、学习、交流的行为方式。关注度是汉专语词汇中的一个热词，意为关注的程度。它对所有的对象，包括人物、事件在内的种种，关注程度都属是指当下的眼前的状态和热度。我们根据群众的点赞 反对数定义 群众的关注度，结合 K-means 算法划分聚类的留言内容，归纳特定地点和人群的热点关注问题，做出表格参考，具体排名见表 5：关注度排名表。

表 5：关注度排名表

热度排名	问题 ID	关注度排名	时间	地点/人群	问题描述
1	1	2344	2019/2/21	A 市 A4 区	58 车贷非法经营，没有人处理
2	2	2097	2019/8/19	A 市 A5 区汇金路五矿万境 K9 县	小区的好多问题没有人负责
3	3	1767	2019/4/11	A 市金毛湾	学生的入学学校没有落实
4	4	669	2019/9/15	A4 区绿地海外滩小区	高铁离小区太近噪音无法消除
5	5	242	2019/6/19	A 市富绿物业丽发新城	物业收费太高，私自停住户自来水
6	6	80	2019/1/10	西地省京港澳高速城区	噪音扰民
7	7	78	2019/1/16	A 市	把和包支付作为考核任务下放给基层工作者
8	8	78	2019/3/26	A6 区月亮岛路	沿线架设 110kv 高压线杆对附近居民会有危险
9	9	66	2019/9/15	A 市三一大道	打通机场北通道，使附近高速畅通
10	10	60	2019/2/25	A 市长房云时代	房子出现裂缝，质量堪忧

（3） 群众活跃度

我们定义为同一用户在留言次数上较多为活跃用户。经过上面的排名，我们对热点问题已经有了一定的了解，一方面，如果一个用户过度的活跃，并且留言的是同一个问题，就说明这件事情对于此用户是重点问题。另一方面，从时间跨度来说，持续留言时间过长的问题没有得到合理的解决的话，可能导致群众对相关部门造成不满，形成危害社会的心理，所以，结合上述两点我们将留言重复次数较多，时间跨度较长的用户进行了归类，并且我们可以考虑把这类问题列为热点问题，避免引起人群的躁动与不满。结合附件三数据，找出活跃度靠前的 10 名用户，做出活跃度排名表，详细排名表见表 6：群众活跃度排名表。

表 6：群众活跃度排名表

热度排名	留言次数	重复留言数	留言用户ID	时间跨度
1	92	80	A00031618	2019/1/3—2019/12/31
2	7	2	A000103957	2019/3/12—2019/11/11
3	7	5	A000108466	2019/1/20—2019/12/14
4	6	4	A000107700	2019/3/1—2019/10/12
5	6	6	A000100792	2019/1/6—2019/1/30
6	4	3	A000107866	2019/1/4—2019/12/24
7	4	4	A000100793	2019/1/7—2019/5/22
8	4	3	A000100428	2019/3/28—2019/9/26
9	4	2	A000104234	2019/3/4—2019/9/10
10	4	3	A000108437	2019/7/23—2019/11/5

3.3 热点问题挖掘结果

结合共性问题表，关注度排名表以及群众活跃度排名表的综合考虑，把三个维度按 3:4:3 的比例整合出热点问题表，得出热度指数，具体见表 7：热点问题表。

表 7：热点问题表

热度排名	问题 ID	热度指数	地点/人群	问题描述
1	1	*****	A 市 A4 区	58 车贷非法经营，没有人处理
2	2	****	A 市 A5 区汇金路五矿万境 K9 县	小区的好多问题没有人负责
3	3	***	A 市	政府加快周边，交通路的建设
4	4	**	A 市金毛湾	学生的入学学校没有落实
5	5	*	A 市长房云时代	房子出现裂缝，质量堪忧

注：‘*’代表从结合共性，关注度以及群众活跃度综合考虑较高问题排名。

‘*’的多少代热点问题排名梯度

四 答复意见评价

4.1 答复意见质量评价方案

根据分词结果，建立词频 TF-IDF 词向量表示方法，首先我们根据语料库训练词向量，也就是针对文本中的每个词汇，我们均用它的向量表示，并获得每个

文本的向量表示，可以将文本中出现的词汇进行求和，利用平均以及加权求和等方式获取最后的结果。再通过有监督的文本表示，从任务中模型的隐层向量提取出来人为是对应文本表示向量，例如文本分类模型 TextCNN，根据模型不断的迭代，最终收敛到较好的效果，可以将模型的池化层拼接后的结果输出作为文本的表示向量。基于翻译任务的 Seq2Seq 模型，亦可以将 RNN 最后一个时间步的输出作为表示特征文本的向量。根据这种思想建立答复意见质量评价方案，结合答复意见评价指标，判断答复意见质量。

4.2 答复意见质量评价标准

4.2.1 相关性标准

（1）相关性标准基本思想

根据词频向量绘制向量夹角示意图，如果夹角为 0 度，意味着方向相同，线段重合；如果夹角为 90 度，意味着方向不完全相似；如果夹角为 180 度，意味着方向正好相反，也就是说，夹角越小，就代表越相似。具体流程如图 2 所示：

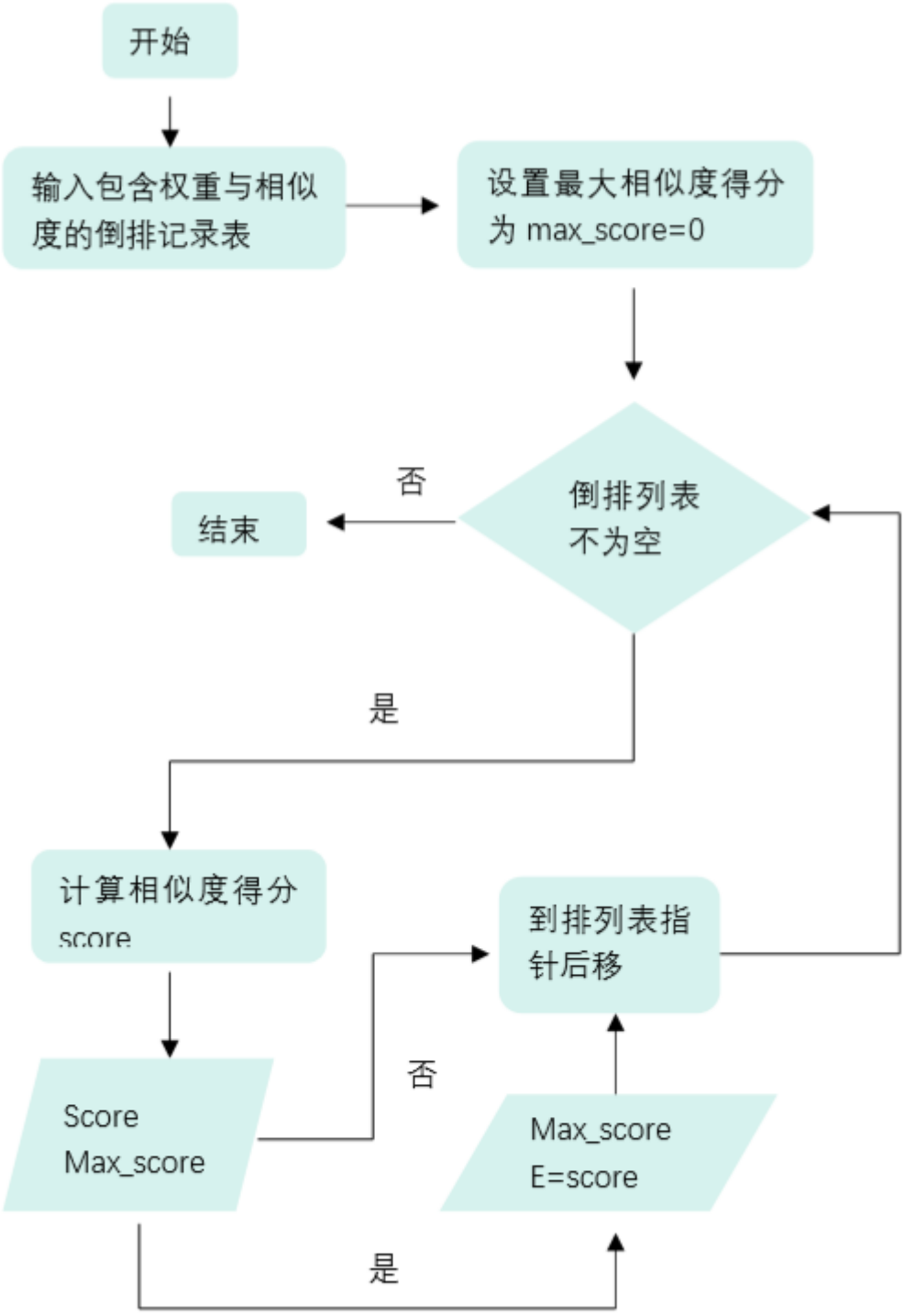

图 2 余弦相似度流程图

（2）相关性标准应用结果

通过余弦定理，计算问题 A 和答复意见 A 的余弦夹角，得到问题 A 与答复意见 A 的相关性余弦值越为 0.72，得到的余弦值接近 1，表明夹角接近 0 度，也就

代表答复的内容与留言的内容是具有一定相关性的，70%的留言答复都是与留言内容相关的。

4.2.2 完整性标准

（1）完整性标准基本思想

首先根据用户编号和答复内容中提及编号进行匹配，判断答复内容与问题是否对应，其次拆分答复意见内容，对其实行分词汇总，关键词提取等操作，再通过分词向量的比对，计算余弦值，如果余弦值接近 1 就可以判断该答复内容的完整性。

（2）完整性标准应用结果

通过 TF-IDF 词向量表示方法获取留言词频向量库和答复意见词频向量库，通过余弦定理计算得出余弦值为 0.62，说明 62%的答复意见是完整的回答了留言问题。

4.2.3 及时性标准

（1）及时性标准基本思想

对提问时间和答复时间进行做差取绝对值处理，处理结果如果在两个工作日以内，说明答复是较为及时的。

（2）及时性标准应用结果

通过计算得出时间跨度低于两个工作日的答复占比 60%，也就是说，14 天以内，相关部门答复了百分之六十的留言问题。

4.3 答复意见质量评价结果

综合答复意见的相关性，完整性，及时性三方面质量评价指标，最终得出答复意见质量处于良好等级。但是随着网络的不断发展，很多群众加入到了网络评价渠道，留言数据进一步增多，这对答复工作带来一定的困扰，通过模型的归类整合，可以提高答复质量，缓解答复意见不及时，不相关，不完整的问题。

五 总结

随着互联网和信息资源共享的迅速发展，信息检索等大量文本数据的处理需求与日俱增，文本分类成为数据分类问题中一个重要的研究课题。文本分类一般包括两个步骤：第一步，通过样本训练，利用样本和类别之间的联系，建立一个

样本分类函数；第二步，通过样本分类函数，对新文本进行分类。在文本信息分类中，常用方法有朴素贝叶斯、支持向量机、随即森林以及逻辑回归，通常来说逻辑回归算法相比其他分类算法简单高效时间复杂度低。当需要处理的数据规模很大时，对处理时效也有较高要求,可以基于朴素贝叶斯算法开发适用于大规模文本数据集在线分类的并行算法。本文提出的评价体系，既为准确综合评价文本分类系统奠定了基础,又为构建文本分类系统和制定分步实施策略提供了一个参考，未来可以在留言分析平台进行推广，提高群众的优良感知和相关部们的答复效率，进一步降低相关行业运营成本，及时有效的解决群众的问题，为相关部门的决策提供有效的数据支撑。

参考文献：

[1] 大数据分析背景下的行政应用

[2] 秦怀强 . 朴素贝叶斯分类算法浅析 . 福建质量管理

[3]胡锡衡 . 正向最大匹配法在中文分词技术中的应用 . 鞍山师范学院学报

[4] 张振亚 , 王进 , 程红梅 . 基于余弦相似性的文本空间索引方法研究

[5]孙凯,孟祥武 . 基于半监督 K-means 的主动学习聚类算法 . 中国科技论文在线