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。

针对问题 1，我们基于引入综合评价指标的模糊综合评价模型，建立关于留言内容的一级标签分类模型。首先对题目所给数据进行数据清洗，即保留下含有重要信息的特征，而去除掉一些无关的特征。同时，我们利用神经网络对数据中存在缺损值项进行删除，进而保证数据录入与运行的完整性。由此，我们便得到了一个能够很清晰反应综合满意度的数据集。接着，我们引入了一个十分重要的参数：综合评价指标——平均星级 AS。为得到他和经过数据清洗后的数据集的关系，我们建立了模糊综合评价模型，并绘制 AS 时序图。

针对问题 2，我们利用问题一的模型，对热点问题进行挖掘，并得出了五个热点问题：

针对问题三，首先，我们建立 AS-模糊综合评价模型。本模型的重点是建立评论特征提取模型。首先，在对评论文本进行数据预处理之后，我们进行评论文本的数值化。而对于不符合这种结构的语言，我们基于潜在语义分析，对其进行词频矩阵降秩。从而，语义相近词语的相关性得到增强，反之亦然。最后，我们对词频矩阵进行 Logistic 回归，再利用回归后的数据进行聚类分析，可以得到 6 种聚集形式，即得到了等级分布。接着，我们开始建立评论-评级评分模型，我们运用支持向量机模型，绘制评论等级-评分等级分布的散点分布图。由该图可清晰的看出评论和评级之间的最佳测量度。而对于该散点图来说，在这个二维的空间中，随着横坐标增大，纵坐标增大，则该产品所代表的潜在成功就越高。因而该模型也能反映出潜在成功和潜在失败的关系。接着，我们引入相关系数模型。为探索特定星级与评论种类之间的关系，我们通过定性和定量两方面进行深入了解，在定性分析过程中，我们应用卡方检验得出卡方的值大于临界值，表明星级对评论有显著的影响。为定量分析，我们引入相关性分析，绘制相关系数列联表，可以得出：特定星级会引起更多的评论。同理，对于评论中的关键词和整体特征的关系，我们也利用相关分析模型，并发现一些评论中的关键词和评级有着密不可分的关系。

关键词：综合评价指标 AS 潜在语义分析 聚类分析 Logistic 回归分析

1.问题背景

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。Hu 等人提出了一个系统完整的评论挖掘过程，包括产品特征挖掘、主观句定位、情感分析、极性判定以及结果显示等部分，并在其论文中阐述了商品特征挖掘的方法.popescul5 等人将改进的 PMI(point mutual information)值引入特征词简直，查全率略有降低而查准率有了显著的提升。Scaffidi 等人开发了一种新的检索系统 red opal，它只识别单个词和二词短语作为特征，并对每个产品的每个特征都进行打分，输出结果按照产品特征的打分综合排序，该方法用打分的方式来区分特征的重要程度，但是在识别特征的过程中依然没有用更有效的剪枝来剔除冗余特征。我们发现，截至目前，少有研究深入考虑影响因素之间相互存在的制约关系。因此，基于题目背景，本文通过对各个特征关系建立模型得到其确定的制约关系，进而建立了一个反应满意度变化的评价-评级评分模型。

2.问题重述

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。 附件给出了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。请利用自然语言处理和文本挖掘的方法解决下面的问题。

3.问题假设

1.假设在数据剔除和清洗过程中数据的完整性和反映出的整体趋势保持不变；

2.假设用户在进行评价和评级时对产品的喜好程度是统一的。

3.假设综合评价指标随时间的跳动情况不影响其整体走向的趋势。

4.假设卡方检验中的简化处理不会对卡方的求值有所影响；

5.假设皮尔逊相关系数，在数据组数 n 较小时，即使相关系数的波动较大，也并不影响整体的相关性，以此保证相关性分析的进行。

4.问题求解步骤

1.针对问题一，我们使用 F-Score 对分类方法进行评价，建立关于留言内容的一级标签分类模型。第一步，我们进行模糊综合评价，需进行数据清洗以及构造平均指标；第二步，我们绘制平均指标时序图；第三步，得到产品满意度趋势，以大致分类。从而得出问题二的答案。

2.针对问题三。我们从以下五个方面入手。

第一，我们建立评价-评级评分模型，这一步我们需建立评论特征提取模型，包括进行数据获取、生成词频矩阵、并基于潜在语义分析对词频矩阵降秩。

第二，我们需把握上升或下降的趋势，这里我们引入 P 参数进行表达。

第三，我们需评估潜在成功或潜在失败，在这里，我们得到分布散点图。

第四，我们要利用热点反作用的影响进行增加满意度，这里我们为验证此表达的成立，建立了相关系数模型。

第五，我们需利用关键词与热点具有较强相关性的影响，在这里为验证表达的成立，我们建立典型相关性分析。

5.问题模型的建立与解决

5.1 问题一

5.1.1 问题分析

问题一是对附件 2 中提供的三种数据材料进行预处理，以消除和合理处理不包含实际意义较小的数据。使用数学证据来识别，描述和支持有意义的定量和定向模式，从而有利于以便后续将群众留言分派至相应的职能部门处理。

5.1.2 基于神经网络的数据清洗及缺项有效剔除、插值

1.数据预处理

数据预处理是数据挖掘过程中的关键步骤。高质量的数据可以改善数据挖掘模型。数据质量可以快速有效地挖掘出有价值的信息。大数据环境中的数据结构越来越多，复杂且不断增加的数据量，并受到噪声数据，数据丢失和不一致的影响，通过上述方法收集的在线评论数据源通常是不规则且嘈杂的，不能直接用于挖掘在线评论信息首先需要对在线评论进行预处理。在线评论是非结构化的文本信息，中文在线注释的预处理主要包括垃圾邮件过滤，中文分词，属性扩展，减少属性，删除无用的单词以及词性标记。

在线商品评论质量特征	细分特征	特征描述
文本特征[7]	语法特征[9]	专有名词、数字、情态动词、感叹词等
	语义特征[10]	情感极性、评价对象等
	文体特征[11]	词汇和结构属性
元数据特征[8]	商品评论发布者与读者之间的互动属性[9]	评论“有用”票数、评论发表时间、商品评分等

图 1 数据剔除原理

2.神经网络模型的建立

基于数据的处理和分析，在神经网络模型中，根据神经网络的互连方式的不同，可以将神经网络分为前馈神经网络，反馈神经网络和自组织神经网络。网络的含

义图如图 2-1 至图 2-3。

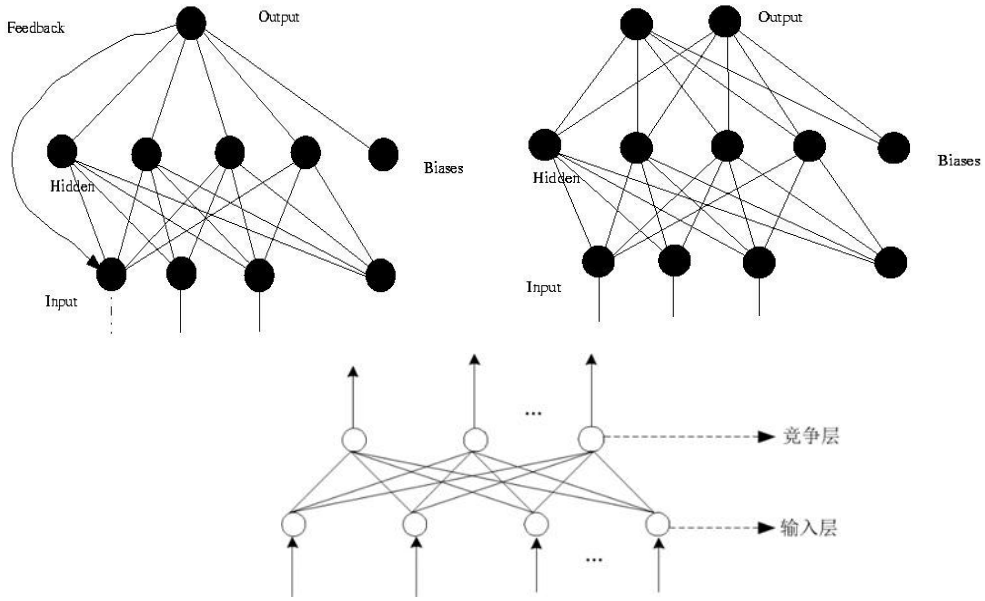


图 2 神经网络分析图

此问题中使用 BP 神经网络，因为在神经元网络模型中，具有反馈的神经网络模型具有更强的校正能力。 为了提高该模型的准确性，该问题使用前馈神经网络作为模型构建的基础。 如图所示：

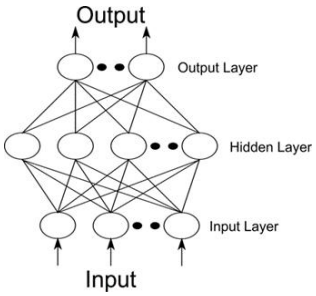


图 3 神经网络模型建立

3. 基于遗传算法的模型求解

步骤 1：计算神经元输入和输出

分别定义每个输入层和输出层的数据，在计算每个层中神经元的输入和输出后，我们可以得到：

（1）隐藏层输入向量为：

$$hi_h(k)=\sum_{i=0}^{7} w_{hi}x_i(k) \qquad h=1,2,\cdots,13$$

（2）隐藏层输出向量为：

$$ho_h(k)=\mathrm{f}(hi_h(k)) \qquad h=1,2,\cdots,13$$

（3）输出层的输入向量为：

$$yi_o(k)=\sum_{h=0}^{13} w_{oh}ho_h(k) \qquad o=1$$

（4）输出层输出向量为：

$$yo_o(k)=\mathrm{f}(yi_o(k)) \qquad o=1$$

步骤 2：通过误差函数计算输出层中每个神经元的偏导数
在获得输入和输出函数之后，需要通过数学运算获得输出层中每个神经元的误差函数的偏导数。 具体解决方案是：

$$\frac{\partial e}{\partial w_{oh}}=\frac{\partial e}{\partial yi_o}\frac{\partial yi_o}{\partial w_{oh}}$$

其中，偏导数可以通过计算得到：

$$\frac{\partial e}{\partial yi_o}=\frac{\partial(\frac{1}{2}\sum_{o=1}^{1}(d_o(k)-yo_o(k)))^2}{\partial yi_o}$$
$$=-(d_o(k)-yo_o(k))yo_o'(k)$$
$$=-(d_o(k)-yo_o(k))\mathrm{f}'(yi_o(k))\triangleq-\delta_o(k)$$

$$\frac{\partial yi_o(k)}{\partial w_{oh}}=\frac{\partial(\sum_{h}^{7}w_{oh}ho_h(k))}{\partial w_{oh}}=ho_h(k)$$

步骤 3：计算隐藏层中每个神经元的误差函数的偏函数
在计算隐藏层中每个神经元的误差函数的偏导数时，需要使用隐藏层与输出层的连接权重，输出层的函数以及隐藏层的函数来计算. 其具体计算为：

$$\frac{\partial e}{\partial w_{oh}}=\frac{\partial e}{\partial yi_o}\frac{\partial yi_o}{\partial w_{oh}}=-\delta_o(k)ho_h(k)$$

$$\frac{\partial e}{\partial w_{hi}}=\frac{\partial e}{\partial hi_h(k)}\frac{\partial hi_h(k)}{\partial w_{hi}}$$

$$\frac{\partial hi_h(k)}{\partial w_{hi}}=\frac{\partial(\sum_{i=0}^{7}w_{hi}x_i(k))}{\partial w_{hi}}=x_i(k)$$

$$\frac{\partial e}{\partial hi_h(k)}=\frac{\partial(\frac{1}{2}\sum_{o=1}^{1}((d_o(k)-\mathrm{f}(\sum_{h=0}^{7}w_{ho}ho_h(k))^2))}{\partial ho_h(k)}\frac{\partial ho_h(k)}{\partial hi_h(k)}$$

步骤 4：使用输出层修改连接权重
通过使用每个神经元的输出来修改连接权重，特定的解决方案计算为：

$$\Delta w_{oh}(k)=-\mu\frac{\partial e}{\partial w_{oh}}=\mu\delta_o(k)ho_h(k)$$

$$w_{oh}^{N+1}=w_{oh}^{N}+\mu\delta_o(k)ho_h(k)$$

步骤 5：使用输入层修改连接权重

通过使用每个神经元输入来修改连接权重，特定的解决方案计算为：

$$\Delta w_{hi}(k) = -\mu \frac{\partial e}{\partial w_{hi}} = \delta_h(k) x_i(k)$$

$$w_{hi}^{N+1} = w_{hi}^{N} + \mu \delta_h(k) x_i(k)$$

步骤 6：计算全局误差

$$E = \frac{1}{2m} \sum_{k=1}^{m} \sum_{o=1}^{1} (d_o(k) - y_o(k))^2$$

步骤 7：判断

在判断中，需要将准确性与最大预算数进行比较。 如果预算的准确性和数量不符合要求，则需要循环执行上述步骤，直到满足要求为止。

4.图标结果

调整此问题中建立的神经网络模型后，该问题中获得的神经网络回归图如下图所示：

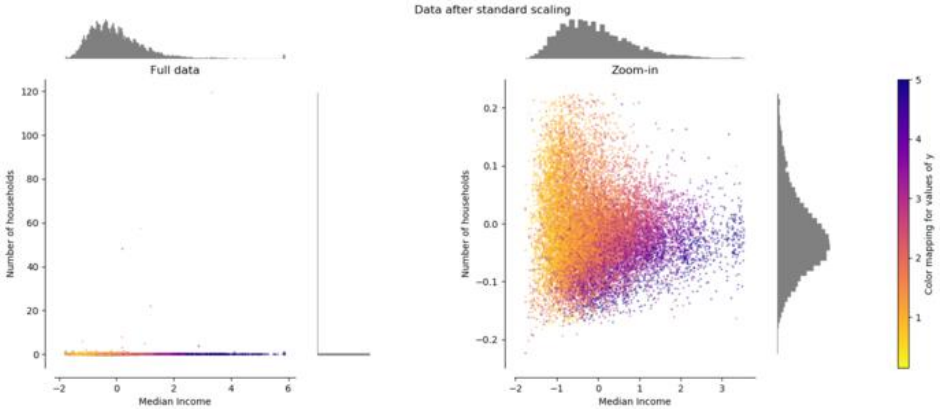


图 4 神经网络回归分析

上图可以通过回归分析获得，散布分布图的位置可以如上所述获得。

5.基于 Levenberg-Marquardt 算法的模型改进

为了提高模型预测的准确性和计算速度，本节使用另一种算法进行 BP 神经网络模型求解，该算法致力于减少残差并减少迭代次数。 因此，本文介绍了用于模型求解的 Levenberg-Marquardt 算法。

基于数值优化的 lm 算法不仅使用目标函数的一阶导数信息，而且还使用目标函数的二阶导数信息。 lm 算法的迭代公式为：

$$\mathrm{X}_{k+1} = \mathrm{X}_k - (\mathrm{J}^{\mathrm{T}}{}_k \mathrm{J}_k + \mu I)^{-1} * \mathrm{J}_k * \mathrm{F}(\mathrm{X}_k)$$

其中，J_k 是包含网络误差对权重和阈值的一阶导数的雅可比矩阵，I 是单位矩阵，μ 是阻尼因子。 LM 算法根据迭代结果动态调整阻尼系数，以使误差函数的值在每次迭代时都减小。 它是梯度下降法和牛顿法的结合，收敛速度更快。

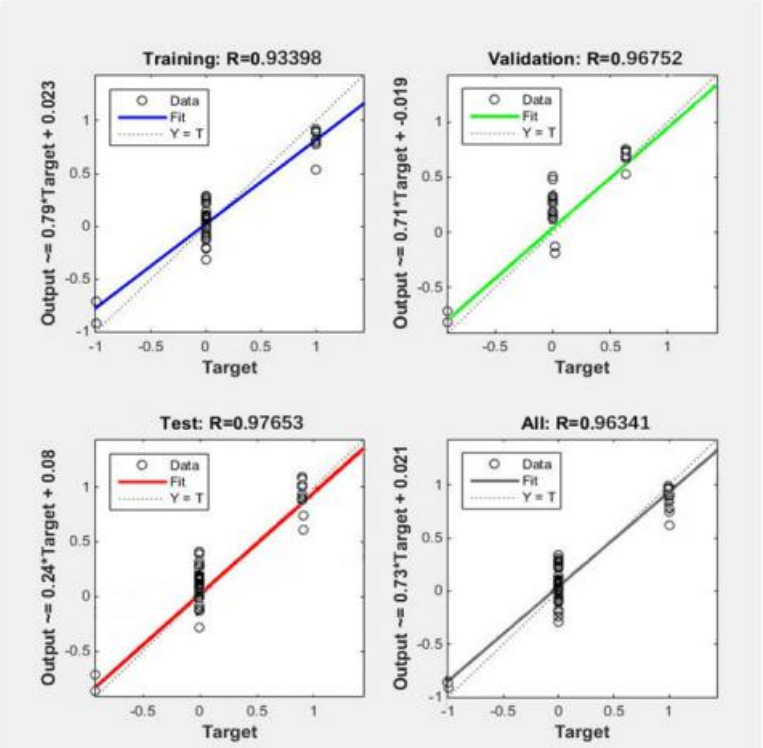


图 5 残差图

根据上述步骤求解改进的 BP 神经网络模型，并根据相应的更准确的权重和阈值获得残差图。 可以看出，残差比以前更收敛。 并且残值也较小，表明该问题基于 Levenberg-Marquardt 算法的 BP 神经网络模型比基于遗传算法的 BP 神经网络模型更为准确。由此我们得出剔除好的数据

5.1.3 结论

由 AS 时序图我们得出了平台用户反馈的趋势，即在时间序列中，我们得出的一级分类标签是明显的。由此我们得出一级分类标签。

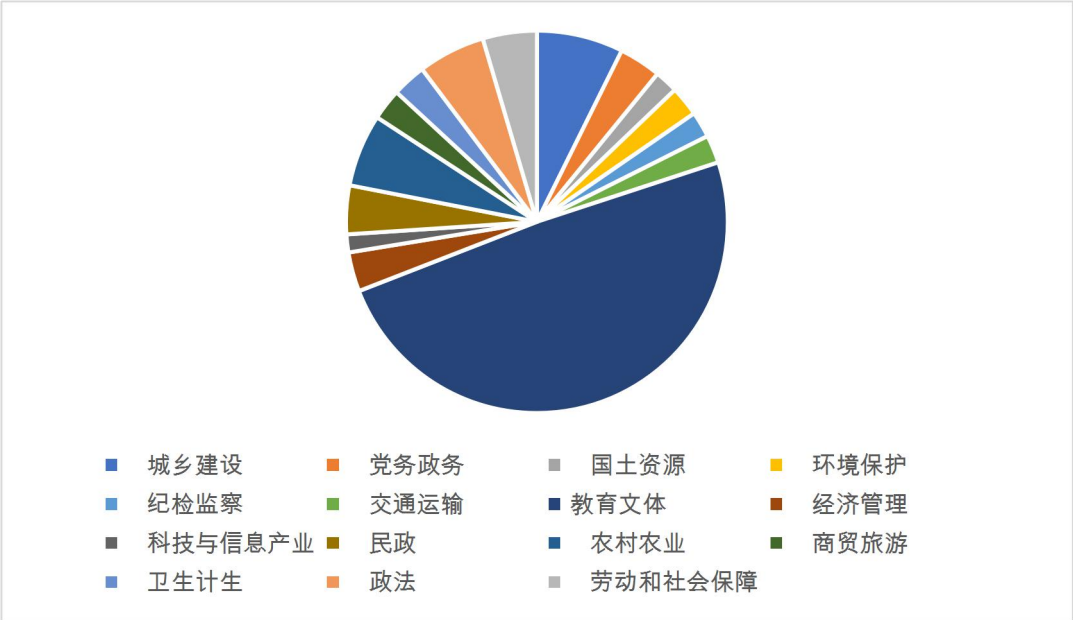


图 6 结果示意图

5.2 问题二

5.2.1 问题重述

某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。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。

5.2.2 构造评价星级指标

1 模糊综合评价法介绍

模糊综合评价法是综合考虑与评价对象有关的多个模糊因素，进行评价并给出最终评价等级结果的方法。这种综合评估方法具有更好的性能，对被评估对象的模糊属性的特征进行处理，并将定性处理和定量处理相结合。该方法的基本原理：首先确定评价指标，评论集和权重集，使用模糊变换，使用隶属度表示因素之间的隶属度，形成一个模糊矩阵，经计算符合性后，获得评价对象。由于模糊综合评估方法是在评估过程中逐个对象进行评估的，因此评估对象的评估不受集合中其他对象的影响，评估结果会影响评估对象。而值对象是唯一的。

2 构造平均星级指标

对于元数据特征，我们可以通过构造综合评价指标——平均星级 AS 来实现。

在引入 AS 之前，我们根据题目中所给数据，找出其中较为重要的列，对其进行定义以下变量，并对变量进行进一步分析和确立其中的关系。

1）定义一级标签指标 S

S 表示的是所给数据中的一列，我们以此来表示提交留言的人的满意度。因而我们可以很明显的知道，S 指标越大，AS 指标则相应地越大。

2）定义有效性指标 E

其代表了评价文本的是否有效。

$$E=\begin{cases} k_2 \times \dfrac{n_{help}}{n_{total}} & \text{if } n_{total} \neq 0 \\ 0 & \text{if } n_{total} \equiv 0 \end{cases}$$

3）定义基数指标 T

T 指标基数越大，则关注评论的基数就越大，则评论的热度也就越大。这决定了我们在问题二中筛选出热点问题.

$$T = k_3 \times n_{total}$$

$$k_3 = \begin{cases} -k_3 & \text{if } s = 1 \text{ or } 2 \\ 0 & \text{if } s = 3 \\ k_3 & \text{if } s = 4 \text{ or } 5 \end{cases}$$

而 k3 的取值取决于 s 的取值。当 s=1 or 2 时，表示客服对产品不满意，此时的 T 指标对综合指标的影响为负值，当 s=3 时，T 指标对综合指标是没有影响的，当 S=4 or 5 时，T 指标综合指标的影响偏大，此时 k3 为正值.

4）定义重要程度指标 VT

在这里，我们引用 0-1 定义，将 1 视作点赞超过 10 的问题，表示此评论具有一定的代表性，而此时我们可以乘以一个恰当的系数如(1.5)，从而把乘积这个整体当做一个 0-1 权值变量，进而影响整个评论的重要程度，从而对整个综合指标产生影响。

$$VT=\begin{cases} k_4(k_4>1) & \text{if } vine\equiv 1 \\ 1 & \text{if } vine\equiv 0 \end{cases}$$

5）定义次数指标 VP

此变量为核实实际需求。

$$VP=\begin{cases} 1 & \text{if verified purchase=Y} \\ k_5(k_5<1) & \text{if verified purchase=N} \end{cases}$$

为此，我们对于以上变量进行建立模型：

$$AS=VT\times VP\times(k_1\times S+E+T)$$

通过在 MATLAB 中编写程序，我们对 AS 绘制时序图，得到下图：

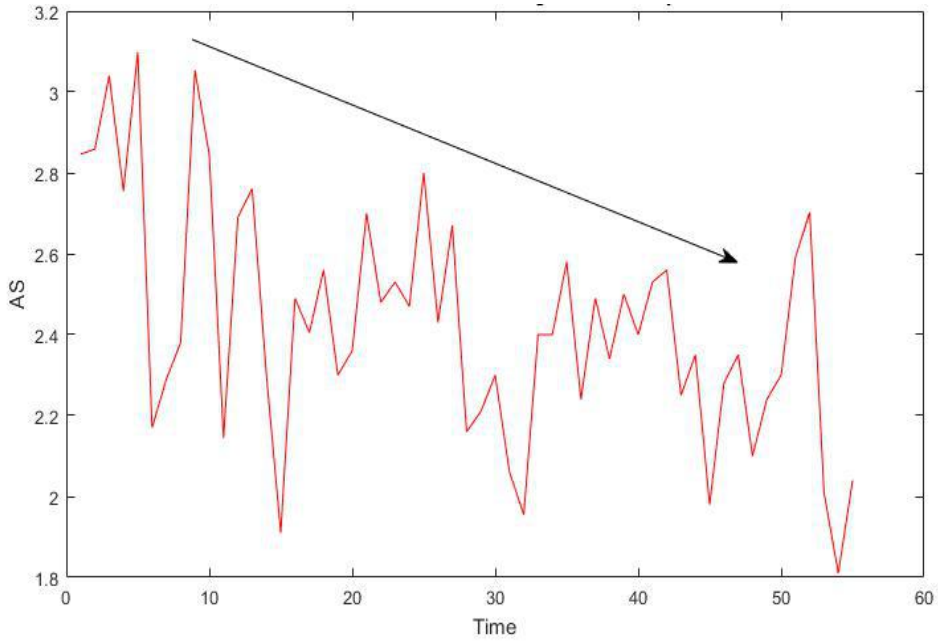


图 7　AS 时序图

5.2.3 结论

通过时序图我们分析得出热点问题随时间变化情况，由此我们得出热点问题留言明细表，并保存为附件 ““热点问题留言明细表.xls”。

5.3 问题三

5.3.1 问题分析

在这个问题中，我们需要针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对 答复意见的质量给出一套评价方案，并尝试实现。

5.3.2 建立评论特征提取模型

1.数据获取

我们基于题目所给数据，在附件 4 中获取评论文本，并对其进行文本预处理。

2.生成词频矩阵

本文通过分词、词性标注、词汇约减，来统计词汇频率，从而构造词频矩阵，即进行评论文本的数值化。

对于 m 个文本，用到了 n 个特征词，则可表示为矩阵：

$$X_{\mathrm{m\times n}}=(word1,word2,\cdots,wordm)^{T}=\begin{Bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{Bmatrix}$$，其中 w_{ij} 表示词语 j 在文本 i 中的频次。

3.基于潜在语义分析对词频矩阵降秩

潜在语义分析认为自然语言中词汇在文档中的出现遵从某种潜在的语义结构，同时文档是由词语组成的，词语也需要置于文档中从而被理解，对于出现不符合这种结构的语言表达我们统称为噪声。在自然语言处理中，应当需要发现这种潜在语义结构，并剔除这种噪声的影响，而潜在语义分析则是利用奇异值分解对噪声进行过滤和去除，当较小的奇异值被忽略后，噪声得以消减，这时，词语间的潜在语义关系得以显现。我们在 MATLAB 中进行数据处理。首先，进行截断奇异值分解，对自然语音的噪声进行预处理，从而获得降秩后的矩阵。第二部，对文档词频矩阵 $X_{\mathrm{m\times n}}$ 进行奇异值分解，并得到近似矩阵 X_k ，

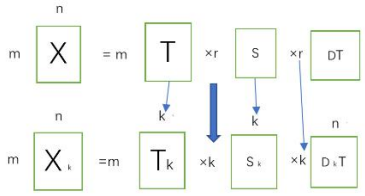


图 8 奇异值分解思路

其中，T、S、D 指的是分解的矩阵;T k、S k、D k 指的是相应的截断后的矩阵名称;v、k 是截断前后相应矩阵的维数。X k 称为截断的奇异值分解式，去除噪声的过程中，为保证不破坏句子结构中潜在的语义结构。其中 k 的取值应满足如下条件:

$$\sum_{i=1}^{k} s_i / \sum_{\mathrm{j=1}}^{\mathrm{r}} \mathrm{s_j} \geq 0.85 \text{ , k<r}$$

结合数据评论的具体情况，我们确定 k 值为 140，这时截取的奇异值占总数的 85.9% 。经过逆运算，我们得到一个 1201 ×194 的近似矩阵。

通过降秩，我们发现，新矩阵不再是一个分布不密集的矩阵，同时很好的保留了原矩阵的大部分信息。同时，语义相近词语的相关性得到增强，而语义无关词语的相关性减弱，这也会便于我们后期的一个分析处理。

5.3.3 有序 Logistic 回归分析

有序 Logistic 回归模型是用来分析 1 个有许多分类因变量与多个自变量之间的关联，本题利用该模型，探究商品评论与评论词语之间的关系，进而找到影响有用性的主成分。

在本文中，评论中点赞数和反对数的数值越大，表明评论有用性程度越高。而评论有用性数值代表不同的有用性等级，我们建立的有序 Logistic 模型如下:

$$g(P_r(Y \le i|x))=\ln\frac{P_r(Y \le i|x)}{P_r(Y>i|x)}=\ln\frac{P_r(Y \le i|x)}{1-P_r(Y>i|x)}$$

$$=\ln\frac{\Phi_1(x)+\Phi_2(x)+\cdots+\Phi_i(x)}{1-(\Phi_1(x)+\Phi_2(x)+\cdots+\Phi_i(x))}=\alpha_i+\beta' x \ , i=1,2,\cdots,k$$

其中，Y 是因变量，本研究中代表等级为 0，1，…，5；x 是自变量，代表各个词语所构成的向量，称为模型的协变量；$\Phi_i(x)$ 表示给定 x 下属于等级 i 的概率；α_i 为常数项，β 是待估计参数向量，β' 是其转置向量；$P_r(Y \le i|x)$ 代表在 x 的条件下，Y 小于等于 x 的概率；$g(P_r(Y \le i|x))$ 是其 Logistic 变换在这里，我们做出假设，令其为协变量的线性函数。我们将因变量设为评论的有用性投票数量，自变量设为主成分分析结果中各主成分，考虑有序 Logistic 回归。本文利用 SPSS 软件，对数据进行有序 Logistic 回归分析，结果见表。

0.206	0.025	85.663	0.000
0.050	0.068	15.260	0.000
0.063	0.034	3.992	0.001
0.024	0.035	5.225	0.005
-0.038	0.012	3.244	0.006
-0.064	0.046	6.555	0.012
0.025	0.052	4.573	0.015
0.024	0.046	11.457	0.003
0.034	0.078	12.457	0.008
0.048	0.068	5.445	0.006
0.130	0.041	3.449	0.001
-0.120	0.024	4.789	0.025
-0.240	0.031	3.145	0.012
0.244	0.024	3.797	0.013
-0.001	0.064	4.577	0.015
-0.024	0.243	3.457	0.001

图 9 分析结果

5.3.4 评估与时间之间的关系以及声誉与时间之间的关系

$$MP=\frac{\sum_{i=[n/2]}^{n} AS_i-\sum_{i=1}^{[n/2]} AS_i}{n[n/2]}$$，即 $$P=\begin{cases}1 & , MP \ge 0\\ 0 & , MP<0\end{cases}$$

由此我们绘制散点图

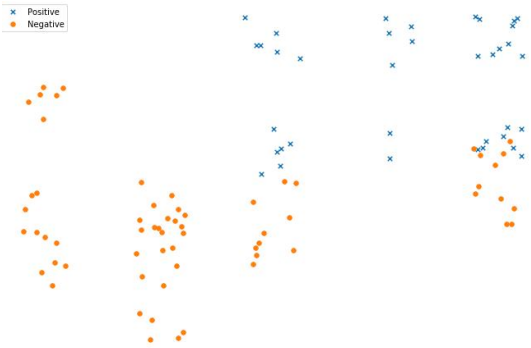


图 10 等级分布散点分布图

5.3.5 建立相关系数矩阵

根据前面问题的解答，我们得知热点问题也被分成了 6 个等级，根据方差分析中的列联表问题，由此我们可以构建 5*6 的列联表。

为此，我们对数据进行方差分析如下

Star rating	Comment text type						
	1	2	3	4	5	6	Total
1	1203	956	654	700	10	20	3543
2	843	625	560	213	32	23	2296
3	621	521	820	1210	421	123	3716
4	231	321	561	1521	861	321	3816
5	24	60	125	357	1125	1304	2995
Total	2922	2483	2720	4001	2449	1791	16366

图 11 列联表

5.3.6 建立相关系数模型

卡方检验是用途非常广的一种假设检验方法，它在分类资料统计推断中的应用，包括：两个率或两个构成比比较的卡方检验；多个率或多个构成比比较的卡方检验以及分类资料的相关分析等。而实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，如果卡方值越大，二者偏差程度越大；反之，二者偏差越小；若两个值完全相等时，卡方值就为 0，表明理论值完全符合。在此，我们用卡方分析来检验列连表中变量之间是否存在显著性差异，或者用于检验变量之间是否独立。

为此，我们首先对模型进行简化假设，我们假定行变量和列变量是独立的。且满足：

$$H_0 : X_1 = X_2 = \cdots = X_j$$

$$H_1 : X_1, X_2, \cdots, X_j \text{ is not all equal}$$

接下来，我们对变量进行定义，对于一个实际频数 f_{ij}，我们假设它的期望频数为 e_{ij}，而落入第 i 行和第 j 列的概率是总频数的个数 n 乘以该实际频数 f_{ij}，即：

$$e_{ij} = n \cdot (\frac{r_i}{n}) \cdot (\frac{c_i}{n}) = \frac{r_i c_j}{n}$$

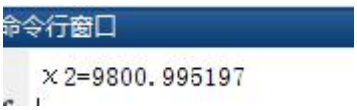
为此，我们得到了卡方的计算公式为

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$ 其中，其自由度为 $(r-1)(c-1)$

其中，f_{ij} 表示列联表中第 i 行第 j 列类别的实际频数；e_{ij} 表示列联表中第 i 行第 j 列类别的期望频数。

为此，我们对卡方进行决策，据已有的资料显示，我们可以根据显著性水平 α 和自由度 $(r-1)(c-1)$ 查出临界值 χ_α^2，而若 $\chi^2 \geq \chi_\alpha^2$，拒接 H_0，若 $\chi^2 < \chi_\alpha^2$，接受 H_0。

为此，我们设计 MATLAB 算法以实现卡方检测，最后，我们得到的输出结果如下图所示：



而我们知道，若卡方的值大于临界值，我们则需拒绝原假设，这就从定性角度和我们解释了显著的影响。

为定量分析特定星级对评论的种类是如何影响的，我们在此进行相关性分析。

关系数是最早由统计学家卡尔·皮尔逊设计的统计指标，是研究变量之间线性相关程度的量，一般用字母 r 表示，本文则采用皮尔逊相关系数进行分析。

$$\rho_{XY}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}}$$

由此，我们绘制相关系数的大小、正负得到影响的结果，我们将该结果以列联表的形式展示：

coff	comment 1	comment 2	comment 3	comment 4	comment 5	comment 6
star 1	0.921	0.516	0.765	0.214	-0.851	-0.796
star 2	0.854	0.754	0.462	-0.756	-0.758	-0.845
star 3	0.452	-0.124	0.856	0.756	-0.564	-0.344
star 4	0.576	0.311	0.861	0.921	0.214	-0.642
star 5	-0.962	-0.612	-0.214	0.264	0.854	0.942

star 1 与 comment 1 共同作用下的值为 0.921,说明评论的重要程度与点赞程度正向相关。同理，star 5 与 comment 6 共同作用下的值为 0.942，说明反对与重要程度成反比。而此相关性分析的结果也定性、直观的说明了点赞数、反对数对整个数据的影响

5.3.7 典型相关性分析

典型相关分析，是对互协方差矩阵的一种理解，是利用综合变量对之间的相关关系来反映两组指标之间的整体相关性的多元统计分析方法。它的基本原理是：为了从总体上把握两组指标之间的相关关系，分别在两组变量中提取有代表性的两个综合变量 U1 和 V1（分别为两个变量组中各变量的线性组合），利用这两个综合变量之间的相关关系来反映两组指标之间的整体相关性。

而在本题中，我们利用典型相关分析，以此直观的判断和分析这些特定的情感词与点赞反对数之间的关系。

1.建立典型相关分析模型

在典型相关性分析中，我们首先定义原始变量相关系数的矩阵为

$$R=\begin{pmatrix}R_{11} & R_{12}\\ R_{21} & R_{22}\end{pmatrix}$$

我们将 A 设为重要程度的系数矩阵：

$$A=\begin{bmatrix}a_1 & a_2 & \cdots & a_r\end{bmatrix}_{p\times r}=\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1r}\\ a_{12} & a_{22} & \cdots & a_{2r}\\ \vdots & \vdots & & \vdots\\ a_{p1} & a_{p1} & \cdots & a_{pr}\end{bmatrix}$$

同理，我们将 B 设为不重要程度的系数矩阵：

$$B=[b_1\ b_2\ \cdots\ b_r]_{p\times r}=\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1r} \\ b_{12} & b_{22} & \cdots & b_{2r} \\ \vdots & \vdots & & \vdots \\ b_{q1} & b_{q2} & \cdots & b_{qr} \end{bmatrix}$$

为此，我们求解变量重要程度和不重要程度出现次数的协方差：

$$\operatorname{cov}(x_i,u_j)=\operatorname{cov}(x_i,\sum_{k=1}^{p}a_{kj}x_k)=\sum_{k=1}^{p}a_{kj}\operatorname{cov}(x_i,x_k)$$

从而，我们就可以求出 x_i 与 u_j 的相关系数：

$$\rho\left(x_i,u_j\right)=\sum_{k=1}^{p}a_{kj}\operatorname{cov}(x_i,x_k)/\sqrt{D(x_i)}$$

为此，我们在 MATLAB 程序中进行相关系数的计算，并得到以下相关系数组成的矩阵：

0.2140　　0.8540
0.3260　　0.7650
0.4230　　0.6440
0.7510　　0.3420
0.9640　　0.2340　　，

其中，第一列表示的是“重要程度”，第二列表示的是“不重要程度”我们在此基础上，绘制相关系数的变化趋势，得到如下结果图：

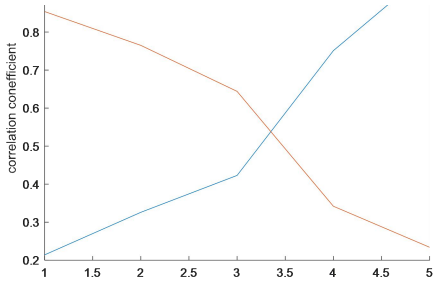


图 12　典型相关分析结果图

根据文本处理模型中，我们可以根据其占所有评论的权重，求出两组单词的占比图，并绘制成饼状图方便直观查阅：

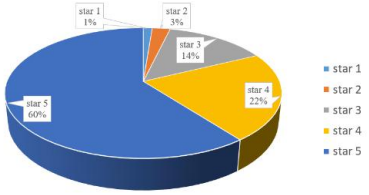


图 13　重要程度占比图

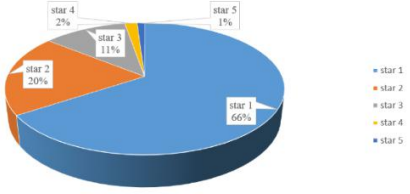


图 14　不重要程度占比图

由该饼状图可知，我们建立的评价模型可以很好的对重要的评论进行筛选，由此我们可以很好地利用该模型进行处理.

5.3.8 结论

等级分布的散点分布图可定性的分析测量度，而该散点图也表明了潜在成功（和潜在失败。我们用综合评价指标 0-1 化后的 P 值，反映了整体声誉的变化趋势，当 P 为 1 时，表示整体声誉在上升，P 为 0 时，表示整体声誉在下降。同时我们得到答复时间的快慢会引起更多这种类型评论，由此我们得到评论中的关键词和重要程度具有较强的相关性。

6.模型优缺点

6.1 优点：

1. 本文巧用综合评价模型，对大量的数据进行清洗预处理，即适当的进行剔除和补项，构建综合评价指标——平均星级 AS，以达到刻画用户满意度随时间的变化关系，这对后续模型的建立有层层递进的作用，而 AS 时序图又能将大量的数据和时间建立联系，进一步可以间接解决其他用户的反馈和时间的关系，这对后续题目的建立有很重要的帮助。
2. 本文的亮点在于建立评价-评级评分模型，在建立模型之前，我们对于评价的处理也是十分新颖。我们基于潜在语义分析，利用奇异值分解来达到对词频矩阵降秩的目的，这一步是十分重要的，就相当于我们对语义相近的相关性增强，从而方便我们进行有序 Logistic 回归分析，来得到对于整体评价的一个分类。进而，我们利用支持向量机模型来构建该评价-评级评分模型。
3. 本文在对星级与评论建立关系时，我们首先利用卡方检测定性分析星级与评论之间的关系，这一步是非常重要的。在得到卡方值后，我们更加印证了自己的观点，接着，我们应用皮尔逊相关系数进行相关性分析，得到相关性列联表，直观生动的反映出了星级和评论之间的关系。这一步从定性到定量的分析，符合自然界中对事物追根溯源的过程，也便于读者理解。
4. 在建立评价中的关键词与各变量之间的关系时，我们大胆的将重要程度和不重要程度作为研究对象，建立相关分析模型，对二者进行相关性分析并得到相关系数矩阵，而相关系数是可以反映评价中的关键词与哪些变量有着不可或缺的关系，由此我们就可以得到，评价中的关键词和评级具有较强的相关性，这一步让我们对于整个的分析有个一个更好的印证。

6.2 缺点：

1. 综合评价计算复杂，且依赖指标权重矢量的确定主观性；
2. 在相关性分析中，在 n 较小时，相关系数接近于 1 的程度与数据组数 n 相关，这易给人一种假象；
3. 在探索评价关键词和其他变量的关系时，由于事件原因，我们只分析了两个单词。这其实对于整体探索来说，是由偏差的。

7.参考文献

[1] 杨桂元,黄己立.数学建模[M].合肥:中国科技大学出版社，2008.8;

[2] 刘来福等.数学模型与数学建模[M].北京：北京师范大学出版社，1997；

[3] 沈继红.数学建模[M].哈尔滨：哈尔滨工程大学出版社，1996；

[4] 林齐宁．决策分析[M].北京：北京邮电大学出版社 2003.2；