第八届“泰迪杯”数据挖掘挑战赛论文

——C 题：“智慧政务”中的文本挖掘应用

目录

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。

本文根据比赛所给出的，来自互联网公开来源的群众问政留言记录，针对比赛的三个问题，运用机器学习、深度学习、自然语言处理、文本挖掘等算法，基于 Python 语言，尝试实现对群众留言的人工智能处理与评价，对留言反映的问题类型进行分类，并制定了留言及回复的评价标准，推动提高政府的管理水平和施政效率。

关键词：智慧政务、文本挖掘、自然语言处理、大数据、云计算、人工智能

1. 挖掘目标

(一) 群众留言分类

在处理网络问政平台的群众留言时，首先需要按照给定的划分体系对留言进行分类，贴上标签，以便后续将群众留言分派至相应的职能部门处理。

本题的目标是通过建立文本挖掘的分类模型，对已有留言样本进行学习，构建合理的泛化能力强的词库与分类器对留言进行分类，并使用 macro F_1-Score 对分类器进行评价：$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$，其中$P_i$为第i类的查准率，$R_i$为第i类的查全率。

(二) 热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题，如“A4 小区多位业主反映小区楼下烧烤店深夜经营导致噪音扰民”就可总结成一个热点问题。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。

本题旨在通过综合考察同一留言主题集中在某段时间出现的次数、留言的支持数与反对数，利用命名实体识别、文本聚类、文本相似度分析等方法，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，定义一个合理的热度评价指标，并给出热度在前 5 名的热点话题的内容及支持与反对情况。

(三)　答复意见的评价

留言的回复是考察政府各部门解决问题能力及与群众沟通能力的不二选择。本题针对留言样本中相关部门对留言的答复，从答复的相关性、及时性、可解释性、易读性、完成度等角度，为答复意见的质量指定一套评价方案。

2. 群众留言分类

该问题为经典文本分类问题，我们采用机器学习常用的分类器集成，完成最终模型的搭建。具体的分析步骤为：

1、 数据清洗（包括分词、转为词向量、降维等操作）；

2、 基础模型搭建与调参；

3、 集成模型搭建。

2.1 数据清洗

根据主办方提供的数据，我们首先做了基本的观察与处理：

1、 共 9210 条数据，无缺失数据；

2、 进行了去除空格符、换行符、制表符的操作；

3、 留言普遍涉及网址及电话内容，因此有较多对文本分类无意义的英文与数字字符；

4、 文本部分由“留言主题”、“留言详情”两列组成，我们分别针对两列单独及将两列文本拼接起来制作了分类器，比较后发现将两列文本拼接后得到的分类器效果较好，因此生成了“总内容”列，用于模型训练；

5、 一级标签有“城乡建设”、“环境保护”、“交通运输”、“教育文体”、“劳动和社会保障”、“商贸旅游”、“卫生计生”共 7 类；

6、 数据轻微不平衡，“环境保护”、“交通运输”、“卫生计生”三类数据较少，数据分布如下表：

一级标签	数据量
城乡建设	2009
劳动和社会保障	1969
教育文体	1589
商贸旅游	1215
环境保护	938
卫生计生	877
交通运输	613

（表 1 原始数据分布）

2.1.1 数据增强

由于数据轻微不平衡，为了提高模型性能，使模型更好拟合，我们对“环境保护”、“交通运输”、“卫生计生”三类数据量小于 1000 的数据进行 EDA 数据增强。EDA 数据增强部分依赖 synonyms 库，对原始文本做了同义词替换、随机插入、随机替换、随机删除 4 步操作后，生成新的近似数据。EDA 部分的函数定义在“eda_func.py”文件中。

我们分别对“环境保护”、“交通运输”、“卫生计生”三类数据进行了 1 倍、1 倍、2 倍的增强，使每类标签的数据量都在 1000 至 2000 的范围内。增强后总共 12328 条数据，分布如下：

一级标签	数据量
城乡建设	2009
劳动和社会保障	1969
教育文体	1589
商贸旅游	1215
环境保护	1898
卫生计生	1783
交通运输	1866

（表 2 EDA 增强后的数据分布）

2.1.2 词库构建

我们的词库基于 THUOCL：清华大学开放中文词库，以及平时在做文本处理时获得的词库。由于留言内容涉及许多道路、小区、地铁站的名称，一般的分词词库欠缺这些专有名称，于是我们通过爬虫的方式，在网上收集了许多城市的道路名与小区名，将其加入原始词库，构建了“dict_update.txt”作为新词库。基于该词库下的分词结果更好更有意义。

2.1.3 去停用词与分词

观察到留言文本中有较多对分类无意义的英文与数字字符，但也并不是所有的数字与字母都无用，比如 4S 店、wifi 等词语就对文本类别有很大的提示。所以，考虑到网址和电话的字符特性，我们首先将 2 位以上的连续数字与 5 位以上的连续英文字符去掉。

考虑到英文、数字、标点符号等情况，我们根据分词的结果多次尝试你，最终对“哈工大停用词表”进行了些许修改，构建了专门的停用词表，去除了文本中的停用词与标点符号。

分词部分运用中文分词第三方库 jieba 的精确模式进行了针对“总内容”一列的分词，分词结果由空格符隔开。并且，考虑到分类器的后续测试与使用，我们将 jieba 分词的结果组成词典并保存，接下来进行 TD-IDF 词向量构建和分类器使用时可传入 TfidfVectorizer 的 vocabulary 参数。

2.1.4 词向量与词典的构建

训练与测试样本输入分类器时，需要以向量的形式。因此，我们首先将 7 类“一级标签”转为 0 至 6 的标签编码。接下来，基于 TF-IDF 算法，我们运用 sklearn 模块中的 feature_extraction.text 进行特征抽取，将分词后的文本数据转为反映词频与词语重要性的词向量。考虑到使模型更好地优化与拟合，降低维数，不应保留整个语料库，应剔除罕见单词的影响，于是我们仅保留了词频最高的 5000 个词。

后续两题的词库、词典构建与分词操作也与第一题类似，因此不做重复阐述。

2.2 基础模型搭建与调参

考虑到速度、性能等问题，我们决定使用基础机器学习分类器和集成方法，最终性能已经较好，因此没有尝试 LSTM 等深度学习分类器。模型学习与调参均依赖 sklearn 模块，采用的基础机器学习分类器有多项式朴素贝叶斯、KNN、逻辑回归、SVM 分类器、XGBoost、多层感知器分类器、Gradient Boosting 分类器、随机森林。其中，为避免由于维度过高而导致 SVM 分类器性能差，对 SVM 的训练集与测试集采用了 SVD 降维至 120 维。

接下来对这些分类器进行调参，以 5 折交叉验证与 f1_macro 指标为评价标准。其中，由于 XGBoost 能优化的参数较多，对其采用了随机搜索（Random Search）进行调参，其余分类器用网格搜索（Grid Search）调参。经过调参后，各分类器的最优性能如下：

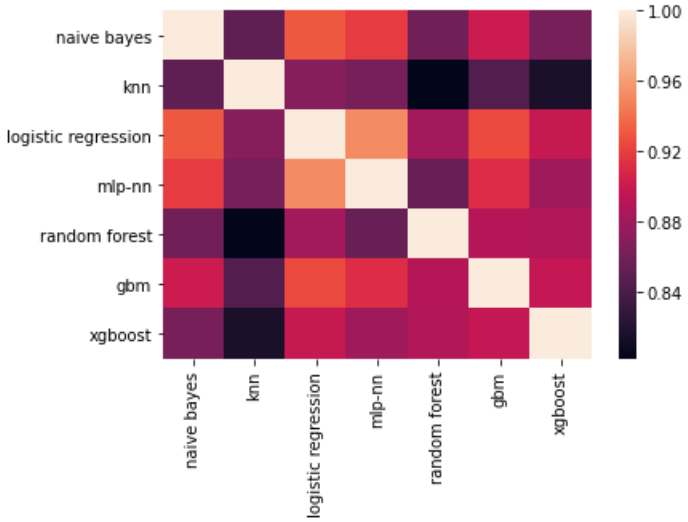
分类器	f1_macro
多项式朴素贝叶斯	0.915
KNN	0.855
逻辑回归	0. 929
SVM	0.906
XGBoost	0.886
多层感知器	0. 918
Gradient Boosting	0. 913
随机森林	0. 856

（表 3 基础分类器在测试集上的 F1 值）

2.3 集成模型与结果分析

为进一步优化，获得更好的结果，我们将基础模型集成，构建最终的分类器。其中，由于针对 SVM 分类器做的 SVD 降维无法实现，用原始为降维数据进行训练时 f1_macro 仅能达到 0.729，因此决定舍去 SVM 分类器。

最终由朴素贝叶斯、KNN、逻辑回归、XGBoost、多层感知器分类器、Gradient Boosting 分类器、随机森林 7 个基础模型，采用投票法构建集成模型。集成模型性能优秀，在测试集的 F1 值达到 0.930。通过 seaborn 库绘制了各基础分类器相关性热力图（如图 1）。



（图 1 集成模型中各分类器相关性热力图）

3. 热点问题挖掘

对于该无监督学习问题，我们的思路是：对每一条留言的文本，考虑时间的因素留言时间在与之相差 90 天内的一一比对其文本的词向量的余弦相似度，超过阈值的认为是相似主题并归类。由于数据质量很好，无缺失及异常数据，因此不需要做过多清洗，可直接构建模型。

3.1 隐狄利克雷模型

LDA(Latent Dirichlet Allocation) 采用词袋模型，利用贝叶斯思想，找到主题分布和词分布。 LDA 假设文档主题的先验分布是 Dirichlet 分布，即对于任一文档dd, 其主题分布θ_d为：

$$\theta_d = Dirichlet(\vec{\alpha}),$$

其中，α为分布的超参数，是一个K维向量. LDA 假设主题中词的先验分布是 Dirichlet 分布，即对于任一主题k, 其词分布β_k为：

$$\beta_k = Dirichlet(\vec{\eta}),$$

其中，η为分布的超参数，是一个V维向量。V代表词汇表里所有词的个数。 对于数据中任一一篇文档d中的第n个词，我们可以从主题分布θ_d中得到它的主题编号z_{dn}的分布为:

$$z_{dn} = multi(\theta_d),$$

而对于该主题编号，得到我们看到的词ω_{dn}的概率分布为：

$$\omega_{dn} = multi(\beta_{zdn}),$$

若有M个文档主题的 Dirichlet 分布，而对应的数据有M个主题编号的多项分布，这样$(\alpha \to \theta_d \to \overrightarrow{z_\alpha})$就组成了 Dirichlet-multi 共轭.

如果在第d个文档中，第k个主题的词的个数为：$n_d^{(k)}$ ，则对应的多项分布的计数可以表示为

$$\vec{n}_d = (n_d^{(1)}, n_d^{(2)}, \dots, n_d^{(k)}),$$

利用 Dirichlet-multi 共轭，得到θ_d的后验分布为：

$$Dirichlet(\theta_d|\vec{\alpha} + \overrightarrow{n_d}),$$

如果在第k个主题中，第v个词的个数为： $n_k^{(v)}$，则对应的多项分布的计数可以表示为

$$\overrightarrow{n_k} = (n_k^{(1)}, n_k^{(2)}, \dots, n_k^{(V)}),$$

利用 Dirichlet-multi 共轭，得到β_k的后验分布为：

$$Dirichlet(\beta_k|\vec{\eta} + \vec{n}_k),$$

但隐狄利克雷不太合适短文本，所以在针对“留言主题”的学习中表现效果不是很好。

3.2 热度评判具体步骤

针对此问题，选择使用“留言主题”作为挖掘热点信息的样本；使用训练数据集对作为衡量文本相似度指标的余弦相似度的阈值进行选择，此过程中用了全部的训练数据。

通常来说短文本具有以下几个特点：

1、 稀疏性：因为内容较短通常包含的内容也有限，有意义的词语较少，故抽取的词向量会较为稀疏；

2、 不规范性：短文本表述简洁，简称、网络用语等不规范用语使用较为广泛，文本收到噪声的影响较大；

3、 实时性：易于传播；

4、 海量性：数据量大。

但针对政务网站的留言主题这类短文本还有以下几个特殊性：首先，它较为精炼，关键词较为集中，能够集中体现"留言内容"的摘要和信息；其次，它的适用于较为规范，虽然有较多的带有地理特征的词汇，但重复较多，经典的文本表示方法效果不会太差。

对于“留言主题”文本数据进行分词后，使用 TF-IDF 算法转化为词向量。在训练阶段，我们尝试了设置 max_df 和 min_df 参数，但效果不是很好，最终我们选择使用词表不设置参数。基于留言主题这类文本的特性，得到基于词典的词频矩阵后，为了不损失信息以及简化模型，也没有使用常见的 TruncatedSVD 进行降维。

经过模型的预训练，最终将用于判定聚合文本的指标余弦相似度的阈值设定为 0.4，既保证了能够充分聚合相似的文本减少损失，也

保证了相似度不会太低。

针对每一个留言（即每一个词向量），我们选取其前后 90 天内的留言计算相似度。即对于每一条留言，在特定的时间（即 90 天内）是否有多条相似留言，以此挖掘热点。

经过聚合及去重，最终得到 10 个留言数约 180 天（前后各 90 天）内超过 10 条热点话题。我们以留言条数+点赞数+反对数+留言用户数（通过用户 id 区分）的均值作为得分，从高到低进行排序。

3.3 结论

经过以上步骤及规则，最终得出的前五个热点话题分别为：

问题 ID	热度	时间范围	地点/人群	问题
1	40.75	2019/8/29 至 2020/1/26	A 市万家丽南路丽发新城居民	附近搅拌站扰民
2	29.5	2019/7/7 至 2020/9/1	伊景园滨河苑业主/广铁集团职工	捆绑销售车位
3	14.75	2019/1/18 至 2019/5/15	热心市民	关于加快 A 市建设的相关建议
4	9.75	2019/6/5 至 2019/9/2	A7 县星沙居民	旧城改造项目
5	9	2019/7/28 至 2019/9/25	魅力之城小区居民	临街门面油烟直排扰民

（表 4 热点问题表）

相应热点问题对应的留言信息详见“热点问题留言明细表.xlsx”文件。

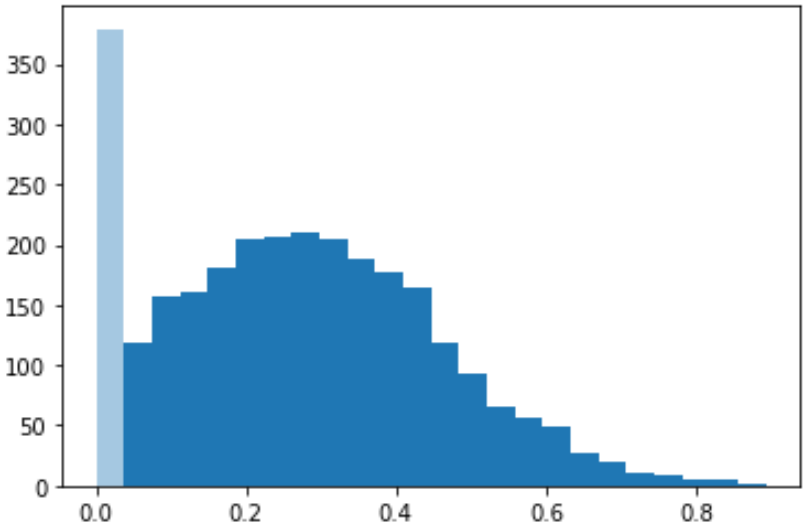
4. 答复意见的评价

针对留言答复意见的文本特点与现实意义，我们从相关性（50%）、时效性（30%）、易读性（10%）、完成度（10%）4 个方面，制定了评价标准，对答复进行评分并且实现。

4.1 相关性分析

相关性即答复内容与留言内容的文本相似度，也就是判断答复的是不是留言群众所关心的内容。这是评价答复意见最重要、最关键的标准，该指标占总分的 50%。

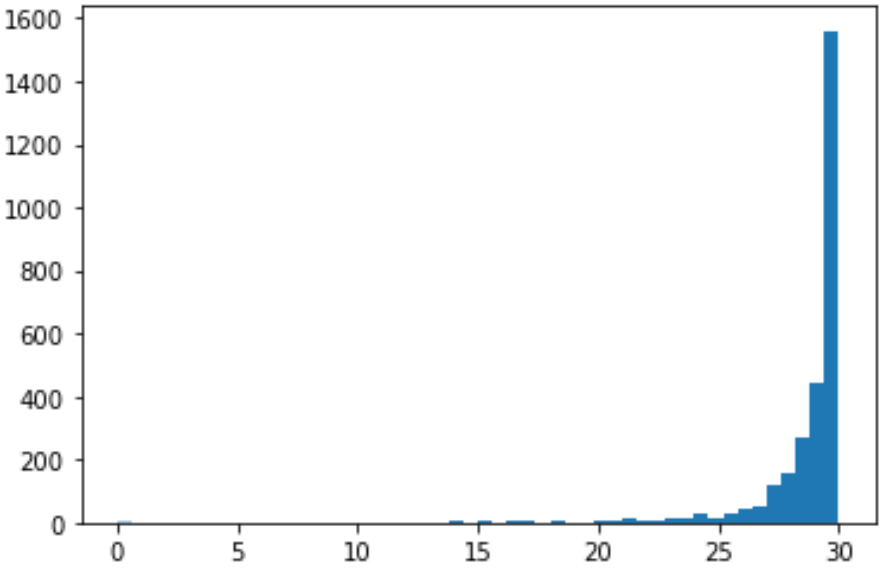
基于 sklearn 库，考虑到向量方向对文本内容的重要性，我们将“留言主题”与“留言详情”列合并，与“答复意见”一起进行分词、TF-IDF 处理后，转为词向量，计算其余弦相似度。cosine_similarity 函数计算结果为[0,1]范围内的浮点数，具体分布如图 2：



（图 2 相似度指标分布图）

4.2 时效性分析

考虑到答复的及时性对于群众来说是非常重要的，因此我们将时效性作为评价标准的 30%。计算方法采用了最简单的，将答复时间与留言时间相减，按一年 365 天 12 个月换算，并将该指标控制在[0,30]范围内。样本的时效性计算结果分布如图 3：



（图 3 时效性指标分布图）

4.3 易读性分析

第三个评价指标为文本易读性，即答复内容是否便于群众阅读与理解，是否对群众有意义。本指标占评价标准的 10%。关于答复内容的易读性，我们借鉴了英文文本中的 Flesch 可读性公式，并结合中文的特性，构造了一个简单的易读性计算公式。

Flesch 可读性公式由作家 Rudolph Flesch 于 1948 年应用心理学杂志的文章《新的可读性尺度》中提出，被认为是最古老、最准确的可读性公式之一，最适用于学校文本。目前，它已成为许多美国政府

机构使用的标准可读性公式。

Flesch 可读性公式的数学表达式为：

$$\mathrm{RE} = 206.835 - (1.015 \times \mathrm{ASL}) - (84.6 \times \mathrm{ASW}),$$

其中，

ASL 为平均句子长度（=文本总单词数/总句子数），

ASW 为单词的平均音节数（=文本总音节数/单词字数），

RE 是[0,100]之间的数，RE 值越大，文本越易读。

根据 Flesch 长期的调查与研究，RE 值在 60 以上的文本普遍被认为是可接受的，RE 值在 50 以下的文本往往被认为是晦涩难懂的。

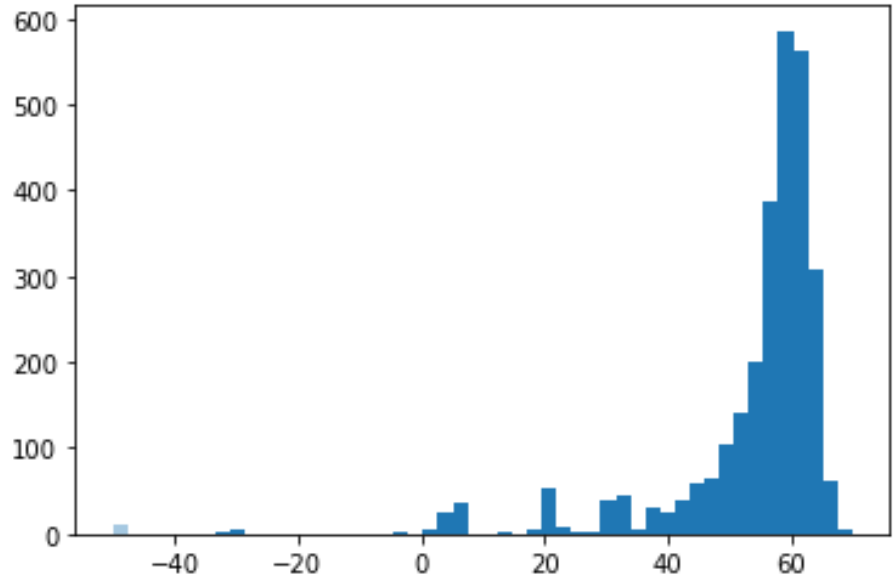
由于汉字是象形文字，与英文的拼音文字大有不同，不能简单地通过音节数甚至笔画数来判断是否易读。于是我们将公式中的 ASW 指标改为文本中的分句数，即用逗号、顿号、引号、书名号、分号等分句标点符号隔开的句子数。直观上易知，文本中使用上述分句标点越少，连续不间断的字数越多，文本越不易读。

此外，考虑到留言答复普遍需要较长较详尽的特点，我们对公式中的系数进行了一些调整，最终得到的易读性计算公式为：

$$\mathrm{RE} = 76.71 - (0.275 \times \mathrm{ASL}) - (\frac{211.5}{\mathrm{ASSN}}),$$

其中，ASSN 为文本中的分句数。

最终结果在$[-50,100]$之间，且普遍集中在[30,70]范围内。过于简短空泛（例如只有一句话或一个日期）的答复虽然不难懂，但由于没有任何意义，RE 值结果会为负值，答复评分会倒扣。样本的易读性计算结果分布如图 4：



（图 4 易读性指标分布图）

4.4 完成度分析

完成度指的是答复意见中体现出的问题的解决程度。我们根据前500 条答复的文本内容，构建了完成度正向词包，即词包中的词被视为是对群众所反映的问题有意义的，例如“已完成”、“核实”、“修缮”、“落实”等词。每出现词包中的 1 个词，完成度可加 1 分，最高为 5 分。

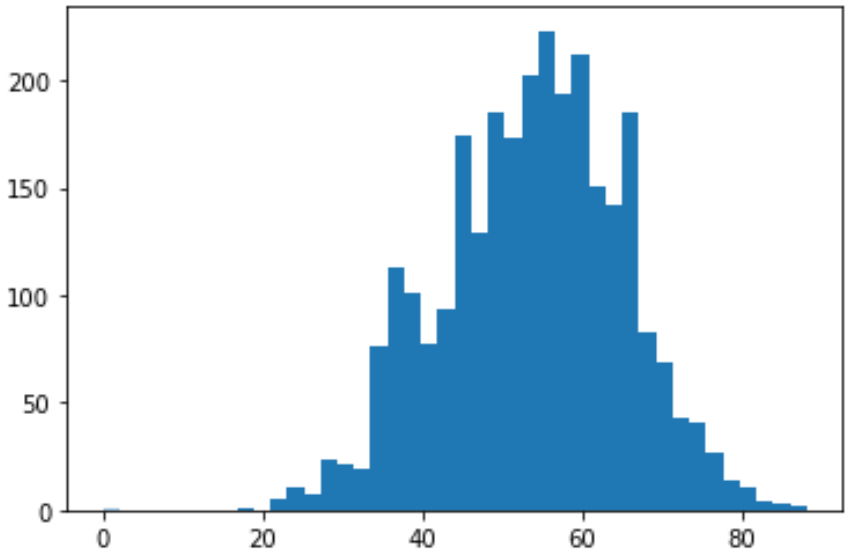
同时，考虑到中文文本表达方式多样，易重复，易受其他助词与否定词的影响，本指标仅占评分标准的 10%。样本的完成度计算结果分布如下：

完成度分值	数据量
5	1003
4	320
3	402
2	407
1	416
0	268

（表 5 完成度指标分布表）

4.5 评价标准

最终评分标准为百分制，总分 = 50 × 相关性 + 时效性 + 0.1 × 易读性 + 2 × 完成度。最终评价分值计算结果分布如图 5：



（图 5 答复意见评分分布图）

根据分值，建议将答复意见的质量分为以下 5 个档次：

0~20 分：极差；

20~40 分：较差；

40~60 分：合格；

60~80 分：良好；

80~100 分：优秀。