

基于自然语言处理技术的电力客户投诉工单文本挖掘分析

吴刚勇, 张千斌, 吴恒超, 顾冰

(国网浙江省湖州供电公司, 浙江 湖州 313000)

摘要:本文主要结合浙江湖州电力业务需求,旨在打破客户对用电诉求存在的盲区,从而提高对用户用电需求的管理程度,实现热点投诉业务工单的原因挖掘。为了更好的深入挖掘投诉工单背后所蕴含的信息,研究基于自然语言处理技术出发,对电力客户投诉工单进行深入文本挖掘,利用隐马尔可夫模型等分词技术分析投诉工单中的受理内容,进行词频统计,通过TF-IDF算法计算关键词重要性权重值,提取权重值大的关键词频作为客户投诉文本挖掘的最终结果,并运用词云分析技术进行分析结果可视化展示;通过文本分类分析,构建文本分类器模型,实现对“热点词频”在不同业务中的分布情况的研究,并根据结果开展相应改进措施。把控住当下电力客户投诉的主要问题,针对性的为不同类型的电力客户提供差异化的服务策略,从而提高客户满意度和忠诚度。专题的推广应用,能够很好的提升客服部门的工作效率,落在实处的为客户解决难题。

关键词:投诉工单;自然语言处理;文本挖掘;词云分析

文章编号:2096-4633(2018)10-0068-06 **中图分类号:**TP3 **文献标志码:**B

随着配售电市场的不断开放,新增配网将允许外部资本投资。注册新的配售电公司已经成为社会资本关注的热点领域,各地售电公司如雨后春笋般纷纷成立,他们在不久的将来极有可能成为电网公司在售电领域的有力竞争者,客户资源流失将是必然,如何减少客户资源流失将是亟待解决的问题。而保障客户稳固的首要措施是提高客户满意度,意味着减少客户投诉,因此,利用自然语言处理技术对客户投诉工单进行文本挖掘分析,了解客户投诉的主要问题,并针对性的提高差异化的服务策略是当下提高客户满意度,增加客户粘性的重要举措。

1 理论模型

1.1 隐马尔可夫模型^[1-5]

隐马尔可夫模型(HMM)是用来描述包含隐含未知参数的马尔可夫过程,该模型是关于时序的概率模型。隐马尔可夫模型的状态不能直接观察到,但是,它能够以观测向量序列观察到,每个观测向量的各种表现状态都是通过概率密度呈现的,每一个观测向量是基于相应概率密度分布的状态序列产生。

隐马尔可夫模型是一个五元组 (S, O, A, B, π) :

S为状态集合:由四种状态构成:词头(标记为

F),词中(标记为M)、词尾(标记为E)、单字成词(标记为W)。

O为观察序列:即实际存在的一个状态的有向序列,如状态 $o_1, o_2, \dots, o_n$,注意状态是存在顺序的。

A为状态转移分布,即S中各元素中,两两之间转移的概率值。比如当前是 s_2 ,下一个状态是 s_9 的转移概率为 $s_{2,9}$ (小于1)。

B为每种状态出现的概率分布。

π 为初始的状态分布

按照机器学习方式的不同,求取参数A、B、 π 的方法大体上分为两类,监督学习和非监督学习。

1.1.1 监督学习方法

如果训练数据集已经给出观测序列及相应的路径序列:

设置初始参数值 $\pi: \pi = \pi(m) = P(m1 = qi)$, $m = 1, 2, \dots, N$,为 $t = 1$ 时刻,观测值处于 q_m 的状态概率。其中: $q_m = F, M, E, W$, $m = 1, 2, 3, 4$ 。

基于统计分析,对每个句子开头第一个字出现频率进行统计,以其统计数除以句子总数,即可计算得到该字的初始状态F、W的概率情况。

假设学习状态转移矩阵A的子元素为 $a(i \rightarrow j)$,那么,子元素 $a(i \rightarrow j) = (\text{由 } qi \text{ 状态变到 } qj \text{ 状态的次数}) / (\text{状态变化总次数})$ 。本文只考虑元素

的状态变化,而不考虑观测值变化。

假设观测概率分布 B 的子元素为 $b_{j(k)}$,那么,
 $b_{j(k)} = (j \text{ 状态下观测为 } k \text{ 的次数}) / (\text{所有状态的总次数})$ 。

总而言之,监督学习方法主要是基于统计频数除以总数,得到相应的的概率,以此构成模型参数。

1.1.2 非监督学习方法

由于监督学习方法需要进行人工标注,这样往往会付出很大的代价,因此,可采用非监督学习的算法来实现。

最后基于维特比算法:基于动态规划算法挖掘出最优路径,即:从 $t = 1$ 开始递归计算,得出在 t 时刻状态为 i 的各条路径的最大概率,到 $t = T$ 时终止,从而实现最终分词。

1.2 TF-IDF 算法^[6-7]

TF-IDF 是一种统计方法,用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降。TF-IDF 实际上是:TF 词频 (term frequency), IDF 逆向文件频率 (inverse document frequency)。TF 表示分词后的词汇 $T (T = t_1, t_2, \dots, t_i)$ 在文本中出现的次数。DF 表示 t_i 的文本频率,即文本集合中含有 t_i 的文本频率。IDF 表示 t_i 的逆文本频率,公式如下:

$$IDF_i = \ln (n / \eta_{DFi})$$

式中 n 为文本总数。

对字词的重要性进行权重计算,计算公式如下:

$$\omega_{i,j} = TF_{i,j} \times IDF_i = TF_{i,j} \times \ln (n / \eta_{DFi})$$

在实际应用中,需要对 $TF_{i,j}$ 进行归一化处理,记 $\lambda_{TF_{i,j}} = \ln (1 + TF_{i,j})$,此时:

$$\omega_{i,j} = \ln (1 + TF_{i,j}) \times \ln (n / \eta_{DFi})$$

1.3 决策树算法

在机器学习中,决策树是一类非常常见的算法,它是一种分类与回归并存的算法,但主要是以分类为主。决策树是一个预测模型;代表的主要是对象值与对象属性二者之间的一种映射关系。并且决策树中每一个节点表示的是某个对象,而每个分叉的路径代表的是某个可能的属性值,每个叶结点又对应从根节点到该叶节点所经历的路径所表示的对象的值。在决策树算法中,仅有单一输出,如果有复数的输出,则可以通过建立独立的决策树来处理不同

的输出情况^[8-11]。

决策树的主要优点是其模型具有很强的可读性,并且分类的速度很快,对噪声数据具有很好的健壮性。在进行模型学习时,利用训练数据集,根据损失函数最小化的原则建立相应决策树模型。在模型预测时,对测试集数据,利用构建的决策树模型进行分类。决策树算法的建模过程通常包括 3 个步骤:特征选择、决策树生成和决策树修剪^[12]。

决策树算法的学习的目标主要是根据设定的训练数据集构建一个决策树模型,使其可以进行准确的分类。其本质主要是通过训练数据集归纳一种分类的规则。在选择决策树的时候,应该选择一个与训练数据矛盾相对较小的决策树,而且选择的模型应该不仅对训练数据集具有很好的拟合效果,而且对未知的数据也要有很好的预测效果^[13]。

机器学习,决策树的典型算法主要有 ID_3 , $C_{4.5}$, CART、 C_5 等。 C_5 算法是基于 $C_{4.5}$ 开发的新版本,它能适用于很多类型的问题^[14-15]。

2 实证研究

基于自然语言处理技术出发,对电力客户投诉工单进行深入文本挖掘,利用分词技术分析投诉工单中的受理内容,对分词结果开展特征选取与降维处理,并进行词频统计,运用词云分析技术进行分析结果可视化展示,把控住当下电力客户投诉的主要问题,针对性的为不同类型的电力客户提供差异化的服务策略,从而提高客户满意度和忠诚度。

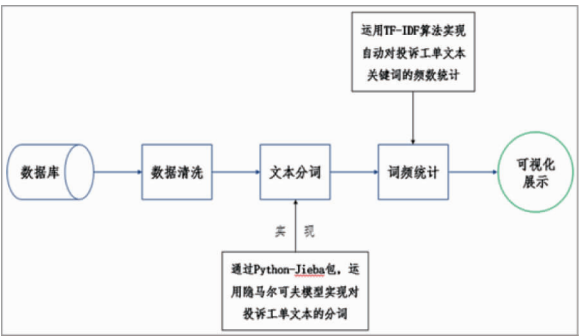


图 1 文本挖掘过程

Fig. 1 Text mining process

2.1 文本分词实现

文本分词是指使用计算机自动对文本进行词语的切分。通过大数据软件 Python 中的 Jieba 包,运用隐马尔可夫模型,实现对客户投诉受理内容的分词。分词结果如图 2 所示。

句子	分词结果
客户从去岁客户来电电压长时间低影响居民生活生产解决客户供电公司核实处理停电信息1701	客户 从去岁 客户 来电 电压 长时间 低 影响 居民 生活 生产 解决 客户 供电公司 核实 处理 停电 信息 1701
户号为3311户号3310004893客户来电投诉供电公司工作人员多早上催费时态度差情况工作	户号 为 3311 户号 3310004893 客户 来电 投诉 供电公司 工作人员 多 早上 催费 时 态度 差 情况 工作
客户反映客户来电供电公司电网建设施工客户经济损失工作人员答应客户年前给予赔偿未	客户 反映 客户 来电 供电公司 电网 建设 施工 客户 经济损失 工作人员 答应 客户 年前 给予 赔偿 未
客户来电投诉客户来电投诉拨打95598报修抢修人员电话13587930156告知客户客户问其回复客	客户 来电 投诉 客户 来电 投诉 拨打 95598 报修 抢修 人员 电话 13587930156 告知 客户 客户 问 其 回复 客
客户投诉客户来电供电公司男性多无故障用户电停客户生活生产带来很大不便客户供	客户 投诉 客户 来电 供电公司 男性 多 无故障 用户 电停 客户 生活 生产 带来 很大 不便 客户 供
客户反映客户来电营业厅申请办理过户业务超出承诺时限客户供电公司相关部门核实处理	客户 反映 客户 来电 营业厅 申请办理 过户 业务 超出 承诺 时限 客户 供电公司 相关 部门 核实 处理
客户来电投诉客户来电供电公司工作人员说今天下午停电安排000安装变压器系统查询不到停	客户 来电 投诉 客户 来电 供电公司 工作人员 说 今天 下午 停电 安排 000 安装 变压器 系统 查询 不到 停
客户反映客户地点日至日内次停电影响居民生活生产解决客户供电公司相关部门彻底解决	客户 反映 客户 地点 日至 日内 次 停电 影响 居民 生活 生产 解决 客户 供电公司 相关 部门 彻底 解决
客户投诉客户投诉供电公司台区负责人李晚言电话号码13511229432告知客户	客户 投诉 客户 投诉 供电公司 台区 负责人 李晚言 电话 号码 13511229432 告知 客户
客户来电投诉客户来电地点计划停电公告时间停电收到停电信息短信通知客户拨打电话现场未	客户 来电 投诉 客户 来电 地点 计划 停电 公告 时间 停电 收到 停电 信息 短信 通知 客户 拨打电话 现场 未
客户投诉客户投诉浙江省湖州市吴兴区环渚街道供电公司工作人员承诺客户客户房屋建全帮	客户 投诉 客户 投诉 浙江省 湖州市 吴兴区 环渚 街道 供电公司 工作人员 承诺 客户 客户 房屋 建全 帮
客户来电投诉客户来电日到营业厅申请检查系统内能查询流程检查承诺时限工作日内出具检测	客户 来电 投诉 客户 来电 日到 营业厅 申请 检查 系统 内能 查询 流程 检查 承诺 时限 工作 日内 出具 检测
客户反映客户来电浙江省湖州市南浔区南浔镇三长行政村施家兜自然村组号前期线路改造施工人	客户 反映 客户 来电 浙江省 湖州市 南浔区 南浔镇 三长 行政村 施家兜 自然村 组号 前期 线路 改造 施工 人
客户来电投诉客户来电京邻国线路接错邻国户号3318073086电工确认表计线路接错工作人员查	客户 来电 投诉 客户 来电 京 邻国 线路 接错 邻国 户号 3318073086 电工 确认 表计 线路 接错 工作人员 查
客户来电投诉客户来电办理计划结算电费协议系统显示预交电费缴费方式扣除元费用客户不解	客户 来电 投诉 客户 来电 办理 计划 结算 电费 协议 系统 显示 预交 电费 缴费 方式 扣除 元 费用 客户 不解
客户来电投诉客户来电投诉浙江省湖州市吴兴区龙溪街道七里亭行政村凌西路西侧西南分区G1	客户 来电 投诉 客户 来电 投诉 浙江省 湖州市 吴兴区 龙溪 街道 七里 亭 行政村 凌西 路 西侧 西南 分区 G1
客户来电投诉客户来电2011营业厅申请三相电缴纳3000元押金合同条款写退还客户近期营业厅	客户 来电 投诉 客户 来电 2011 营业厅 申请 三相 电 缴纳 3000 元 押金 合同 条款 写 退还 客户 近期 营业厅
客户来电投诉客户来电户号3330147750电能表接线客户停电推上邻国电表家中电视	客户 来电 投诉 客户 来电 户号 3330147750 电能表 接线 客户 停电 推上 邻国 电表 家中 电视

图2 分词结果

Fig.2 Result of word segmentation

2.2 特征选取与降维

2.2.1 特征选取

对 255 条文本数据进行分词后,每个单词均可作为标识文本的特征,各特征在整个文本集合中出现的次数如图 3 所示。

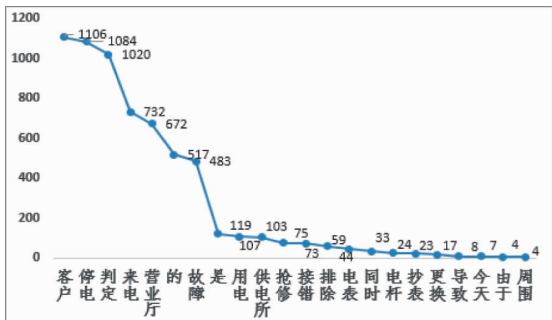


图3 分词结果折线图

Fig.3 Line chart of participle result

2.2.2 特征降维

特征降维主要是为了对特征进行识别剔除,剔除掉对文本区分程度很少的特征,如:与电力业务关系不大的特征,以降低后续文本聚类的算法复杂度,主要包括以下情形:

- (1) 剔除掉几乎每条文本都出现词,如:“客户”、“来电”、“判定”等在 255 多条文本数据中出现 200 次以上的高频无用词汇。
- (2) 剔除掉常用特殊词,主要包括常见的称谓词、结构词、语气助词,如“我”、“你”、“是”、“啊”等与电力业务无关词汇。
- (3) 去除一些词频很小的特征,如“导致”、“今天”、“由于”、“周围”等在 255 多条文本数据中出现次数少于 10 的低频词汇。

通过对出现频率设定相应的阈值(上限,下限)来自动实现特征的降维。

2.3 关键词频提取

通过上述对分词结果进行特征选取与降维,实

现对无关词汇的过滤,留下与电力业务相关的关键词。结合实际电力业务,对现有关键词进一步筛选,通过 TF-IDF(词频-逆文档频率)算法计算关键词重要性权重值,提取权重值大的关键词频作为客户投诉文本挖掘的最终结果。

2.4 可视化展示

通过 Python 软件,运用词云分析实现投诉工单文本挖掘结果展示:

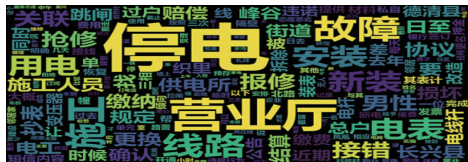


图4 投诉文本词云

Fig.4 Word cloud of complaint text

由图 4 可知在客户投诉中,词语“营业厅”、“停电”、“故障”等出现频数较多,表明客户主要对营业厅、停电、故障等意见较大,可从这几个方面入手,如:提高营业厅服务水平、减少停电或停电信息通知到位、加强故障检修减少故障发生等等措施,从而提高客户满意度,改善客户投诉问题。

2.5 分类器模型构建

通过相关大数据算法,根据选定的训练集数据生成分类器模型,利用该模型实现对工单数据的自动分类,了解热点事件在不同业务类型的分布情况,并根据结果提出相应措施,实现供电企业的主动式服务,提高客户满意度,降低投诉度,从而增加客户粘性。

2.5.1 分类器模型确定

本文主要选取“停电”为例进行分析,从投诉工单中随机选取部分疑似“停电”工单和一定比例非“停电”工单作为训练集。

构建 C_s 树、SVM 算法、贝叶斯网络、C&R 树等算法的文本分类模型,通过“准确率”“遗漏率”等指标对比分析不同分类器模型的在训练集上的分类效果,最终选取决策树 C_s 树模型作为热点事件“停电”的最终分类器模型。

2.5.2 分类评估及分类结果

将训练集中的待分类工单输入至 C5 决策树模型中进行分类,并进行效果评估,评估结果如图 5 所示。

从图 5 可以看到,模型的查全率为 93.7%,遗漏率为 9.1%,对疑似“停电”的工单基本能够全部识别,但从查准率与准确率这两个指标来看,查准率为 34.1%,准确率为 79.8%,表明虽能基本

识别出疑似“停电”的工单,但也将其他的热点工单判定成该热点工单,因此需要对该分类模型进行调优与改善。

表 1 分类结果验证

Tab1 Verification of classification result

评估指标	查全率 R/%	查准率 P/%	准确率 R’/%	遗漏率 L/%
指标值	93.7	34.1	79.8	8.1

对于模型的调优改进主要通过三个方面:①丰富文本挖掘的同义词库与专业术语词库,从而使分词的结果更加准确;②对于特征词的选择需要更加准确;③对 C5 决策树算法进一步调参,如对模型的深度与叶子节点进行调优。通过对分类器模型进行调整后,再将训练集工单数据输入模型中进行分类,得到的结果如图 6 所示:

图 2 模型调优后分类结果验证

Tab.2 Verification of classification results after model tuning

评估指标	查全率 R/%	查准率 P/%	准确率 R’/%	遗漏率 L/%
指标值	89.2	94.3	93.6	12.2

从表 6 可以看到,经过一番的模型调参改进后,查准率与准确率两个指标分别提高到了 94.3% 与 93.6%,在遗漏率指标上虽然略有增高,但仍在可接受

表 3 疑似“停电”工单在不同业务类型中的分布情况

Tab.3 Distribution of suspected “string household” work orders in different businesses types

营业投诉	14	停送电诉	74	服务投诉	2	供电质量	21	电网建设	7
电能计量	3	抢修服务	27	服务行为	1	电压质量	18	电力施工	7
抄表催费	5	停电问题	42	服务投诉	1	供电可靠性	3		
用电变更	2	停送电信息公告	5						
当眼价电费	1	业扩报装	2						
业务收费	1								

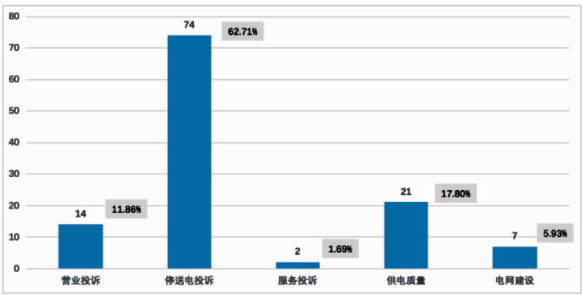


图 5 疑似“停电”工单在不同业务类型中的比例情况

Fig.5 Proportion of suspected “blackout” work orders in different business types

范围内。通过对训练集数据的验证,可以明确调参后的分类模型能够较为准确的识别出热点事件的工单。

2.5.3 分类结果及应用

将全部投诉工单输入至调整后的分类模型中,共识别出疑似“停电”工单 118 张,其中的工单业务类型情况如表 3、图 5 所示。

从图 7 和图 8 中的结果可以看到,建立的 C5 决策树分类模型在进行投诉工单文本分类时,可以忽略座席人员在记录投诉工单过程中的多级业务类型的层级,并直接通过工单的手里内容进行深入挖掘,识别出深藏在不同业务类型中的热点事件“停电”的投诉工单。其中,停送电投诉类工单共 74 张,占疑似“停电”工单的 62.71%;供电质量类工单工 21 张,占疑似“停电”工单的 17.80%;营业投诉类工单共 14 张,占疑似“停电”工单的 11.86%;电网建设类工单工 7 张,占疑似“停电”工单的 5.93%;服务投诉类工单数量最少,为 2 张,仅占疑似“停电”工单的 1.69%。

通过分类后的热点事件工单与营销部门的业务进行相关联,可以明确不同类型的业务对客户满意度和投诉度的影响程度,并以此提出有效的改进方法与解决措施,为供电企业提高主动式服务,在提高满意度和降低投诉度的基础上增加客户的忠诚,增大电力局在配电市场的有效竞争力^[16-18]。

3 应用价值

95598 投诉工单的深入分析与研究是基于“客户诉求”出发,深入客户投诉工单受理内容,挖掘客户的真实需求与投诉原因。应用大数据分析技术,采取隐马尔可夫模型、分词等分析方法对投诉工单开展文本挖掘,打破原有对客户投诉需求模糊不清的壁垒,把控住当下电力客户投诉的主要问题,针对性的为不同类型的电力客户提供差异化的服务策略,提高客户粘性和满意度。

4 结语

本文利用基于自然语言处理的文本挖掘技术,结合浙江湖州电力业务需求,热点业务工单专题研究,打破了客户对用电诉求存在的盲区,提高对用户用电需求的管理程度,实现热点投诉业务工单的原因挖掘。专题的应用,将会提高客服部门的工作效率,为实现主动、精准的客户服务提供决策支持,以提升客户服务能力。

参考文献:

- [1] 厉建宾,朱雅魁,付立衡.基于大数据技术的客户诉求分析与应用[J].电力大数据,2017,20(10):14-17.
LI Jianbin,ZHU Yakui,FU Liheng. The customer demands analysis and application based on big data [J]. Power Systems and Big Data,2017,20(10):14-17.
- [2] 林亚平,刘云中,周顺先,等.基于最大熵的隐马尔可夫模型文本信息抽取[J].电子学报,2005,33(02):236-240.
LIN Yaping,LIU Yunzhong,ZHOU Shunxian, et al. Using hidden markov model for text information extraction based on maximum entropy[J]. Acta Electronica Sinica,2005,33(02):236-240.
- [3] 周顺先,林亚平,王耀南,等.基于二阶隐马尔可夫模型的文本信息抽取[J].电子学报,2007,35(11):2226-2231.
ZHOU Shunxian,LIN Yaping,WANG Yaonan, et al. Text information extraction based on second-order hidden Markov model [J]. Acta Electronica Sinica,2007,35(11):2226-2231.
- [4] 陈友仁,赵正校.基于隐马尔可夫模型的车牌自动识别技术[J].红外与激光工程.2001,30(02):102-107.
CHEN Youren,ZHAO Zhengxiao. Automatic recognition of vehicle license based on hidden Markov model [J]. Infrared and Laser Engineering,2001,30(02):102-107.
- [5] 于江德,肖新峰,樊孝忠.基于隐马尔可夫模型的中文文本事件信息抽取[J].微电子学与计算机.2007,24(10):92-94.
YU Jiangde,XIAO Xinfeng,FAN Xiaozhong. Event information extraction from chinese text based on hidden Markov models [J]. Microelectronics & Computer,2007,24(10):92-94.
- [6] 黄承慧,印鉴,侯昉.一种结合词项语义信息和TF-IDF方法的文本相似度度量方法[J].计算机学报,2011,34(05):856-864.
HUANG Chenghui,YIN Jian,HOU Fang. A text similarity measurement combining word semantic information with TF-IDF method [J]. Chinese Journal of Computers,2011,34(05):856-864.
- [7] 周源,刘怀兰,杜朋朋,等.基于改进TF-IDF特征提取的文本分类模型研究[J].情报科学.2017,35(05):111-118.
ZHOU Yuan,LIU Huailan,DU Pengpeng, et al. Research on text classification model based on improved TF-IDF feature extraction [J]. Information Science,2017,35(05):111-118.
- [8] YOUWEI WANG,LIZHOU FENG. A new feature selection method for handling redundant information in text classification [J]. Frontiers of Information Technology & Electronic Engineering, 2018,19(02):221-234.
- [9] Ximing LI, Jihong OUYANG, You LU. Topic modeling for large-scale text data [J]. Frontiers of Information Technology & Electronic Engineering,2015,16(06):457-465.
- [10] 彭敏,黄佳佳,朱佳晖,等.基于频繁项集的海量短文本聚类与主题抽取[J].计算机研究与发展,2015,52(09):1941-1953
PENG Min,HUANG Jiajia,ZHU Jiahui, et al. Mass of short texts clustering and topic extraction based on frequent itemsets [J]. Journal of Computer Research and Development,2015,52(09):1941-1953.
- [11] LONGXIANG WANG,XIAOSHE DONG,XINGJUN ZHANG, et al. TextGen: a realistic text data content generation method for modern storage system benchmarks [J]. Frontiers of Information Technology & Electronic Engineering, 2016, 17 (10):982-994.
彭京,杨冬青,唐世渭,等.一种基于语义内积空间模型的文本聚类算法[J].计算机学报.2007,30(08):1354-1363.
- [12] PENG Jing,YANG Dong-qing,TANG Shiwei, et al. A novel text clustering algorithm based on inner product space model of semantic[J]. Chinese Journal of Computers,2007,30(08):1354-1363.
- [13] 宋锦萍,侯玉华,杨晓艺,等.基于小波域多状态隐马尔科夫树模型的自适应文本图像分割算法[J].电子学报,2007,35(01):118-122
SONG Jinping,HOU Yuhua,YANG Xiaoyi, et al. Context-adapted document segmentation based on multi-state hidden Markov tree models in the wavelet domain[J]. Acta Electronica Sinica,2007,35(01):118-122.
- [14] 黄晓斌,赵超.文本挖掘在网络舆情信息分析中的应用[J].情报科学,2009,27(01):94-99
HUANG Xiaobin,ZHAO Chao. Application of text mining technology in analysis of net-mediated public sentiment [J]. Information Science, 2009,27(01):94-99.
- [15] 阮光册.基于文本挖掘的网络新闻报道差异分析[J].情报科学,2012,30(01):105-109.
RUAN Guangce. Analysis on web media report differences based on text mining [J]. Information Science, 2012,30(01):105-109.
- [16] 罗欣,张爽.深度学习在电力潜在投诉识别分类中的应用[J]浙江电力.2017,36(10):83-86.
LUO Xin,ZHANG Shuang. Application of deep learning in identification and classification of potential complaints of electric power [J]. Zhejiang Electric Power, 2017, 36 (10):83-86.
- [17] 梁浩波.基于文本挖掘的用电客户诉求智能聚类研究[J].广东电力,2016,29(08):45-50+66

LIANG Haobo. Research on intelligent clustering for electricity customers' demands based on text mining [J]. Guangdong Electric Power,2016,29(08):45-50+66.

[18] 赵晓明,张学强,曹岚. 基于关键词的电力系统“大数据”与“云计算”专题文献分析[J]. 浙江电力,2016,35(02):27-30.

ZHAO Xiaoming,ZHANG Xueqiang,CAO Lan. Thematic analysis of "big data" and "cloud computing" in power system based on key words[J]. Zhejiang Electric Power,2016,35(02):27-30.

收稿日期:2018-08-25

作者简介:



吴刚勇(1972),男,本科,高级工程师,主要从事综合能源服务工作。

(本文责任编辑:范 斌)

Text mining analysis of power customer complaint worksheet based on natural language processing technology

WU Gangyong,ZHANG Qianbin,WU Hengchao,GU Bing

(State Grid Zhejiang Huzhou Electric Power Supply Company,Huzhou 313000 Zhejiang,China)

Abstract: This paper mainly combines the demand of Zhejiang Huzhou power business, aiming at breaking the blind spot where customers have complaints about power consumption,thereby improving the management level of users' electricity demand and realizing the reasons mining for hot complaints business work orders. In order to better dig deeper into the information contained in the complaints worksheet,based on natural language processing technology,this paper makes a in-depth text mining of power customer complaints work orders,uses hidden Markov model and other word segmentation techniques to analyze the acceptance of complaints in the work order content,and carries out the word frequency statistics,the keyword importance weight value is calculated by TF-IDF algorithm,and the keyword frequency with large weight value is extracted as the final result of customer complaint text mining. Finally,the word cloud analysis technology is used to visualize the analysis results,Through the text classification analysis,the text classifier model is constructed to realize the research on the distribution of "hot word frequency" in different business,and the corresponding improvement measures are carried out according to the results.The main problems of current power customer complaints are controlled,and differentiated service strategies are provided for different types of power customers,thereby improving customer satisfaction and loyalty. The promotion and application of the topic can improve the work efficiency of the customer service department,and solve the problem for the customer.

Key words: complaint work orders;natural language processing;text mining;word cloud analysis