基于 LDA 模型的 95598 热点业务工单挖掘分析

文/王震　代岩岩　陈亮　林晓兰

摘要

95598 热点业务工单的挖掘与分析，对业务详单进行分类，热点问题的及时发现与追踪，起到很重要的作用。目前对于热点业务工单的分类，采用人工查询工单并分类，工作繁琐且效率低。本文提出了一种基于 LDA 的热点业务工单分类模型，对工单中的受理内容进行中文自然语言的处理和数据挖掘，实现对热点业务工单的分类筛选，对准确有效地提高供电服务质量具有十分重要的现实意义。

【关键词】语义分析 文本挖掘 热点工单 LDA

随着电力行业售电侧改革不断加深，对客服管理质量要求越来越高，需要进一步改善客户体验和提升客户满意度。要提升客户满意度，需从客户的热点业务工单入手，分析挖掘热点业务聚焦点，快速有效找出业务短板，提升客户服务质量。

本文依据一般客服问题管理机制和文本挖掘理论，并结合电力企业客服特点，阐述了如何对客服热点工单文本进行挖掘分析以及如何在系统中基于 LDA 算法对其进行分类的应用。业务工单中的投诉工单、客户回访处理不满意的工单能直接反映客户对产品、对服务的感知，是客户满意度的最直接反映。从现状来看，目前的热点工单分类的处理方式，是由调查分析人员通过对 95598 客户诉求数据的分析，对受理的内容进行分析和筛选，然后完成分类。这种方式缺乏有效的辅助分析手段，分析手段单一，影响服务问题的分析和解决效率，因此需利用中文自然语言处理、文本挖掘等技术，结合电力领域的业务特点，对 95598 来电工单进行自动化的智能分析与处理，实现热点业务工单的智能分类与原因挖掘。

1 热点业务工单业务描述

热点业务主要包括停电、乱收费、抄核收、人身伤亡、赔偿、外界关注等的工单，相互之间可以重复统计。通过对工单的挖掘结果，对热点业务工单进行可视化展示，展示维度包括单位、市县公司、以及业务类型。

表 1：各热点业务类型定义

热点业务类型	业务定义
停电	涉及停电内容的工单
乱收费	含有乱收费现象的工单
抄核收	含有抄表、核算、收费业务的工单
人身伤亡	工单内容体现有人身伤亡现象的工单
赔偿	工单中有赔偿现象的工单
外界关注	特殊来电的工单，如 110,12345，媒体来电，政府热线等。

热点业务主要分为以下 6 个大类，分类如表 1。

2 文本挖掘相关理论

文本挖掘（Text Mining，TM）是近几年来数据挖掘领域的一个新兴分支，是以文本数据为特定挖掘对象的知识挖掘。文本挖掘是抽取有效、新颖、有用、可理解的、散布在文本文件中的有价值知识，并且利用这些知识更好地组织信息的过程文本挖掘的要点是分词，根据文本数据中的特征信息进行分词处理，以此构建文本的中间表示。原始的文本数据通常以非结构化或半结构化数据呈现，再利用文本挖掘手段转换为结构化文本，进而发掘新的概念与对应关系。

基于领域特征词表的特征词标注，主要以大量来电工单中反映业务种类、热点问题现象的特征词为基础，设立特征词表，进行基于特征词匹配的子句标注，并依不同维度进行工单分类。

通过构建检测模型和确定模型指标体系、指标阈值等参数，对工单数据进行大数据分析，采取可视化大屏全屏展示的方式进行全方位多角度的展开实时监控、分析、预警和展示，及时发现当前问题、变化趋势，并对问题点改进情况进行跟踪。

2.1 文本自动分类

为了方便对文本进行归类与管理，我们通常会在实际操作中给文本内容指定一个或多个分类类别。传统的人工标注，需要耗费巨大的时间和精力。文档自动分类是文本挖掘领域针对这一业务场景的典型应用。通过相应的分类器，实现文本分类的预测功能。当对一个新文档进行分类时，分类器通常为这个文档指定一个或多个类别标签，并根据算法策略给出分类标签的可信度。

按照机器学习方式的不同，文档自动分类的实现大体上分为两类，监督学习和非监督学习。

监督学习方法是在训练集上建立模型，针对每个训练集，需人工为每个训练集中的文档打上类别标记，接着用训练集训练一个分类器。训练完成后，这个分类器将能够预测任何一个给定文档的类别。非监督学习方式与监督学习方法的不同点，在于他们不需要训练数据集，可以在一批文档中自动发现相似文档并完成分组。

实际应用中，分类器一般由数据集整理，数据预处理，分类算法等三部分组成。数据集，需要整理足够数量的高质量文档，为了将数据集转化为便于进行文本挖掘的格式，同时为提高结果的精度，数据预处理主要包括中文分词、词项的权值修正等步骤。分类算法与策略主要依据相应的文本挖掘模型计算文档的特征，最终实现对文档的分类处理。

2.2 主题模型

主题模型（Topic Model）是在机器学习和自然语言处理等领域是用来在一系列文档中发现抽象主题的一种统计模型。通常来说，若文档有一个中心思想，即主题，那么文档中就会频繁出现与主题关联密切的词项然而，实际上文档会包含多个主题，并且每个主题所占比例也不相同。因此，，如果一篇文档和主题 A 相关的内容占 10%，和主题 B 有关的内容占 90%，那么和主题 B 有关的词项出现的次数大概会是和主题 A 有关的词项出现次数的 9 倍。主题模型试图用数学框架来体现文档分类的这种特点，先对每个文档进行自动分析，再统计文档内词语出现的频率，最后根据统计信息来判断当前文档包括哪些主题，以及每类主题的所占比例。

主体模型的优势如下：还有如下两个优点：

（1）无监督学习完全自动化，在训练过

● 山东省自主创新及成果转化专项项目“电力行业大数据平台的研制及产业化应用”（项目编号：2014ZZCX10105-1）。

表 2：热点工单分类结果

工单编号	上级单位	业务类型	受理内容	热点业务类型
2016070827240728	潍坊	意见	【人身伤亡】【民事赔偿】客户来电反映，在 6 月 23 号客户父亲不慎在田地里触电，之后找过供电公司工作人员处理触电问题，但一直未处理好，客户再次来电继续处理。请供电公司相关部门尽快核实并答复客户。	人身伤亡
2016070526540266	济宁	投诉	【业扩报装】客户来电反映三年前在纸坊供电所营业厅申请三相电，电工让其缴纳 7000 元，表示电表开户费用。之后电工只给客户安装三户照明电，并且未退还费用，也未安装三相电。客户问电工费用问题，电工告知找所长，所长说要问电工，存在推脱。客户表示非常不满，要求供电公司相关部门尽快核实处理并尽快给客户合理解释。	乱收费
2016070125645785	淄博	故障报修	【多户无电】客户报修此处多户居民客户停电，请处理。	停电
2016071328518724	东营	业务咨询	【咨询总户号】查询户号信息	外界关注（12345）
2016070827173699	淄博	服务申请	【抄表数据异常】客户来电反映，0510453424 户号近期抄表数据 7 月份电费过高且 7 月 5 日到 7 月 8 日电费过高。电费用的过快，现申请对抄表数据核实，请相关工作人员核实处理。	抄核收
2016070827273133	青岛	业务咨询	【施工现场恢复】客户来电反映在 2016 年 7 月 7 日，有供电公司工作人员在此处维护线路，换电线和电线杆时，砸坏了客户 2 棵树木，给客户造成损失，现要求赔偿，客户表示已经联系当地电工但一直未得到答复，请尽快核实处理。	赔偿

程不需要引入人工的标注，而是以概率计算为基础，进行分类训练。

（2）满足多种不同的语言形式，都可以经过分词处理后进行主题模型的训练。

3 基于LDA的热点工单分类

在 LDA 主题模型中，一个主题是由一些词项的分布定义的，每个主题由带有分布率的一系列词项构成。一篇文本则是由一些主题构成的。LDA 主题模型的产生过程，主要是按照概率分布，选择部分主题，从主题中再按照概率，选择部分词语，这些词语的无序组合就组成了最终文档。

若上述两个概率分布能被我们计算清楚，则可得到一个模型，根据某偏文档推断出其主题分布，也就是分类。文档生成的过程与由文档推断主题的过程互为逆过程。

3.1 LDA主题模型

LDA 模型的数学原理比较复杂，其 Gibbs Sampling 公式如下：

$$p(z_i = k|\vec{z}_{\neg i}, \vec{w}) \propto \frac{n^{(k)}_{m,\neg i}+\alpha_k}{\sum_{k=1}^{K}(n^{(t)}_{m,\neg i}+\alpha_k)} \cdot \frac{n^{(t)}_{k,\neg i}+\beta_t}{\sum_{t=1}^{V}(n^{(t)}_{k,\neg i}+\beta_t)} \quad (1)$$

公式的右边部分其实就是文档—> 主题—> 词语的路径概率，其物理意义在于 K 条的路径采样，K 为主题的个数。LDA 主题模型的文档分类过程分为两步：训练过程和推理过程。训练过程即根据当前训练文档集建立模型。同时在建模过程中，对各种估计参数进行选取与调优，直至训练过程结束。训练过程结束后，模型建立和参数优化已经完成。而推理过程则是，根据当前模型与参数，对新的文档进行主题分布的计算过程。

训练过程如下：

（1）随机初始化：给语料中每篇文档中的每个词 w，随机的赋一个主题编号 z。

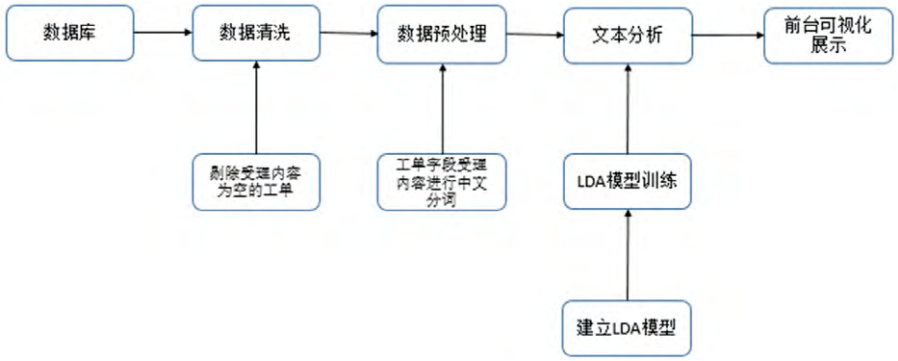

图 1：基于 LDA 模型的热点工单文本挖掘过程

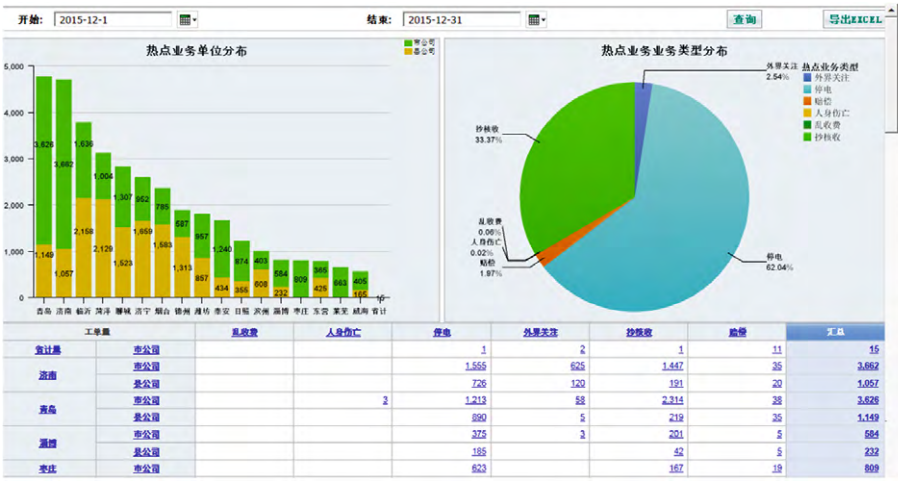

图 2：热点工单分类结果统计分析

（2）更新主题：对语料库进行重新扫描，根据公式（1）重新采样主题并更新。

（3）重复采样，直至 Gibbs Sampling 公式结果收敛。

（4）建立 LDA 模型：统计语料库中主题 - 词语共现频率矩阵。

推理过程如下：

经过训练后，得到参数文档 - 主题分布矩阵 Θ 与主题 - 词语分布矩阵 Φ。其中对文档 - 主题分布矩阵 Θ 一般不进行保存。而在推理过程中需要使用主题 - 词语分布矩阵 Φ。根据 Gibbs Sampling 公式，对新文档中每个词的主题进行抽样，得到此文档的主题分布 θ_new，

>> 下转 192 页

配方批次追踪系统的设计与实现

文/杨晓宾

摘要

卷烟自动化生产物流中配方高架库负责为制丝工序提供烟叶原料，处于卷烟生产源头把控的重要位置。按照工艺要求配方库的出入库作业都是以配方批次为基本单位的，但为了保证效率，在实际生产中都按照烟叶等级类型进行批量入库，两者的错位给配方库造成了较大的管理混乱和质量隐患。针对该问题，本文提出了线外辅助系统的解决思路，通过线外辅助系统完整全面地对整个配方批次全生命周期中的各个过程进行描述，可以实现与实际物流运转情况的双向核对，实现自动化的质量隐患报警。配方批次追踪系统上线后通过前后工作的分析比较，对保障配方批次“原批原用”取得了显著效果。

【关键词】配方高架库 配方批次 线外辅助系统 双向核对 原批原用

昆明卷烟厂配方高架库主要职责是为制丝部门及时、准确、柔性化地提供生产所需批次烟包，整个自动化立体仓库的计算机物流上位管理系统 WMS(Warehouse Management System) 采用了昆船集团的整体集成物流管理软件 TIMMS，通过计算机统筹的管理调度命令控制协调现场大量物流设备的高效运转和相互配合，满足配方入库作业和配方出库作业的正常运转以及灵活多变的业务管理需求。

为保证入库作业效率，配方入库作业在实际操作中都是按照基础烟叶类型——烟叶等级进行运转和管理的，但为了满足制丝工艺要求，配方出库作业又是以基础烟叶类型组合——烟叶配方为单位出库的 (使用最新的配方入库，以递增的方式为每一批入库配方编制批次号，以批次号为标识定位到某一入库的批次配方后进行出库)，频繁的配方等级加剧了“每天配方入库计划分解后是否与实际入库等级相一致？”、“结算意义下配方库存烟包是否能构成一个个完整配方”、“每天配方出库烟包是否与对应入库配方相一致”等方面的疑问，而解决这些问题的核心在于如何高效、准确地进行配方批次追踪。

1 系统设计思路

目前大部分的自动化物流上位系统都是由专业的物流软硬件集成制造商提供的，TIMMS 也一样，软件源代码受版权保护，再加上 TIMMS 配方入库设计理念是以烟叶等级为基础单位进行运作和管理的，造成了现有 TIMMS 框架基础上完善地解决配方批次追踪问题的代价较大。

为此我们提出了线外辅助系统的解决思路，即通过信息化技术研发基于 B/S 架构的配方批次追踪系统，该系统不会对自动化物流上位系统的运转产生任何影响，两个系统相互独立互不干涉。

配方批次追踪系统着重点在于对配方批次入库、出库对应关系的表示，包括配方批次的入库日期、配方牌号、配方的具体明细、出库时间、出库班次等信息，完整全面地对整个配方批次全生命周期中的各个过程进行了描述，以批次号为核心对配方库入库、出库、库存进行配方粒度级的高效管理，这里需要强

<< 上接 191 页

同时在利用公式计算条件概率的时候，公式中的 φ 保持不变。具体过程如下：

（1）随机初始化：给语料中每篇文档中的每个词 w，随机的赋一个主题编号 z;

（2）重复扫描当前文档，按照 Gibbs Sampling 公式，对于每个词 w，重新采样它的主题；

（3）重复以上过程直至 Gibbs Sampling 收敛；

（4）统计文档中的主题分布，该分布即为所求的主题分布 θ_new。

3.2 基于LDA的热点工单内容分类过程

本文在对热点工单受理内容的分类过程中，首先进行数据清洗和预处理，剔除 95598 热点工单受理内容的文本为空或者格式不正确的工单。其次对工单内容进行分词，即基于 IK Analyzer 这个轻量级的中文分词工具包，对热点工单的内容进行分词。再次建立 LDA 模型进行文本语义分析，包括 LDA 模型的训练和 LDA 模型的推理过程，把工单受理内容按照乱收费、人身伤亡、停电、外界关注、抄核收、赔偿等六个主题进行文本分类。最后在 95598 运营分析系统热点业务分析栏进行结果的汇总和展示。

4 业务价值展现

首先从效率上来讲，对热点业务工单分析和分类替代了人工查找、分类和汇总，能提高工作速率。工单的受理内容多，数量多，仅凭人工肉眼去辨别，不仅耗时巨大，可操作性也不高，当类别等因子需求产生变化时，很难对结果进行调整和再利用。而通过该系统，利用大数据挖掘、语义分析技术、文本分类等技术。计算时间短，时效性更强，复用性高，更有助于及时决策。

其次从质量上来讲，利用基于 LDA 的热点工单分类模型对数据进行处理，经实验验证，能达到较高的准确率，数据质量较优。

5 结语

本文利用基于 LDA 的文本挖掘技术，结合山东电力业务需求，热点业务工单专题研究，大大改善目前人工进行热点工单分类效率较低的状况，实现热点业务工单的智能分类与原因挖掘。专题的应用，将会提高客服部门的工作效率，为客服管理人员作出决策提供技术支持，提高了用户的满意度。

参考文献

[1]JiaweiHan. 数据挖掘：概念与技术（原书第三版）[M]. 北京：机械工业出版社，2012.

[2]Ronen FeIdmarl, James Sanger. 文本挖掘 [M]. 北京：人民邮电出版社，2009.

[3]Mitchell T.M, 曾华军 . 机器学习 [M]. 北京：机械工业出版社，2008.

[4] 吕镇超，姬东鸿，吴飞飞 . 基于 LDA 特征扩展的短文本分类 [J]. 计算机工程与应用，2015, 51(04): 123-127.

[5] 姚全姝，宋志理，彭程 . 基于 LDA 模型的文本分类研究 [J]. 计算机工程与应用，2011, 47(13): 150-152.

作者单位

山东鲁能软件技术有限公司　山东省济南市 250001