文章编号:1009-1831(2016)S1-0055-03

DOI: 10.3969/j.issn.1009-1831.2016.S1.022

基于文本数据挖掘技术的95598业务工单主题分析应用

丁　麒[1],庄志画[2],刘东丹[3]

(1. 国网浙江省电力公司 电力科学研究院,杭州　310000; 2. 国网宁波供电公司,浙江 宁波　315000; 3. 朗新科技股份有限公司,杭州　310000)

The matic analysis and application of 95598 business order based on the techniques of text data mining

DING Qi[1], *ZHUANG Zhi-hua*[2], *LIU Dong-dan*[3]

(1. Electric Power Research Institute, State Grid Zhejiang Power Company, Hangzhou 310000, China;
2. State Grid Ningbo Power Supply Company, Ningbo 315000, China;
3. Lang-shine Science and Technology Co., Ltd., Hangzhou 310000, China)

摘要:运用LDA文档主题生成模型对海量95598业务工单进行文本挖掘,建立工单标签知识库,通过95598业务工单与标签知识库的识别匹配,形成样本工单。对样本工单开展主题分析,通过多维度数据对比,并结合专家经验和业务分析,提出专题策略建议,为客户提供精准化供电服务。

关键词:95598业务工单;文本挖掘;主题分析

Abstract: The article initially carries out a text mining on quantitative 95598 business orders by LDA text themes model and establishes knowledge base of order tags. The article further studies on the data match between 95598 business order and knowledge base of order tags so as to develop the exemplary order and thematic analysis. By comparing multi-dimension data and adopting expertise comments and business analysis, this article proposes specific strategies and suggestions so as to provide more precise supply services for customers.

Key words: 95598 business order; text mining; thematic analysis

中图分类号:F407.61　**文献标志码**:C

1　运用技术和模型

1.1　文本挖掘技术

文本挖掘TM(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[1-2]。文本挖掘的要点首先是进行分词,根据分词结果,从文本数据中抽取的特征信息形成文本的中间表示[3]。把原来的文本数据以结构化的数据呈现,再利用分类、聚类等技术转化为结构化文本,并根据该结构发现新的概念和获取相应的关系[4]。

1.2　LDA文档主题生成模型

LDA(latent dirichlet allocation, LDA)是一种文档主题生成模型,也称为三层贝叶斯概率模型,包含词、主题和文档三层结构。其采用了词袋(bag of words)的方法[5],这种方法将每一篇文档视为一个词频向量,每一篇文档代表了一些主题所构成的一个概率分布,而每一个主题又代表了很多单词所构成的一个概率分布[6]。

由于LDA模型是全概率生成模型,因此具有清晰的内在结构,并且可以利用高效的概率推断算法进行计算;同时LDA模型参数空间的规模与文档数量无关,因此更适合处理大规模语料库。LDA模型已经在机器文本学习的诸多领域以及信息检索中得到应用[7-8]。本文运用该模型与专家经验一起构建知识库。

2　主题分析模型

2.1　总体模型架构

95598业务工单主题分析的总体模型架构主要包括:知识库构建、样本选取(主题选取)、数据处理和主题分析等。

2.2　知识库构建

2.2.1　通过专家经验构建

知识库由问题敏感词库、原因敏感词库以及对应的标签库等构成。通过对95598业务工单的研读

收稿日期:2016-07-14

和分析，根据专家的丰富经验，以行业标准用语为基础，将各类型日常工单高频词汇的语义与问题点和原因点匹配，并剔除常见的干扰词汇，梳理出与问题点和原因点对应的敏感词，建立问题敏感词库和原因敏感词库，再对大量的敏感词进行分类归因，定义为不同的标签，为每个敏感词“贴上”标签，从而形成问题标签库和原因标签库。

2.2.2 通过LDA模型挖掘文本

以专家经验的方式进行知识库的构建，无法实现常态化词汇补充。引入LDA文档主题生成模型，运用数据挖掘技术，对文本进行挖掘分析，通过机器自动文本学习，同样以标签形式对问题敏感词和原因敏感词进行分类归因，从而达到常态化补充知识库的目的，减少人工工作量，有效提升工作效率。例如：在被“贴上”天气原因这一标签的95598业务工单中，词汇N频繁地出现，那么通过机器文本学习，判断词汇N和天气原因的关联概率，并将词汇N添加到天气原因这一标签下的敏感词中。

2.2.3 知识库甄别规则

知识库甄别规则主要分为标签匹配、信息识别和业务分类3类。

（1）标签匹配规则

在95598业务工单文本搜索问题敏感词库和原因敏感词库中的词汇，搜索结果与对应标签进行匹配，再将标签“贴在”相应工单上，如果工单文本搜索结果为对应多个标签的不同敏感词，则将多个标签同时“贴在”相应工单上。

（2）信息识别规则

95598业务工单基础信息中，根据信息要素或是要素中的关键字段来识别该工单客户的相关属性。例如：通过识别基本信息中客户身份证号码第7—10位，以此计算该工单对应户主年龄。

（3）业务分类规则

通过95598业务工单文本词汇的语义匹配将工单业务类型重新归类，并与工单基础信息记录的业务类型对比，起到矫正或重新归类的作用。

2.3 样本选取（主题选取）

2.3.1 样本工单筛选和基础信息检查

首先将知识库导入95598业务工单，通过问题敏感词和原因敏感词的识别和匹配，为业务工单“贴上”问题标签和原因标签，以此对干扰、重复和空白的无效工单进行第一次剔除。其次，检查筛选95598业务工单的基础信息，包括身份信息、联系信息、地址信息等，剔除信息大量缺失、数据分析价值不高的工单，从而选取出内容完整、信息量丰富的样本工单。

2.3.2 主题选取

主题选取采用专家提出和指标测算相结合的方式进行。首先根据专家经验选出若干个需引起重视的分析主题，再通过测算各个主题工单的地区占比、增量幅度、引起投诉占比和引起故障占比等指标，排除问题不显著、个发性、临时性的主题，从而选取出时段性或地域性强、典型突出的主题进行数据分析，并有针对性地开展对策研究与制定。

2.3.3 主题和标签关联

选取确定分析主题后，判断分析主题与标签的关联关系，在上文的样本工单中搜索该主题包含的标签，由此形成与主题相关的样本工单集合。

2.4 数据处理

2.4.1 样本字段重新定义

将样本工单内容的字段重新进行定义，从而形成有数据分析价值并利于分析的样本。例如：某样本工单中时间字段为“2016/06/16 10:30:00”，将该时间字段重新定义为“第二季度、6月、工作日、非假期、工作时间”。同样地，针对样本工单中的地址字段根据台区、线路等进行重新定义，并可对地址进行重构，使得地址字段信息具有层次化结构关系。

2.4.2 数据处理过程

数据处理从样本工单选取开始，按前文所述方式选取样本后，将工单数据的“基础数据、文本数据和关联系统数据”3个部分分别进行字段重新定义，实现工单所有内容标签化，使样本工单数据最终以标签化形式存在。当分析主题确定后，自动对样本工单数据进行标签关联匹配，搜索出符合该主题分析需求并具备数据分析价值的的样本工单集合。

2.5 主题分析

在选取样本工单并对样本工单内容的字段重新进行定义后，根据专家经验和指标测算相结合的方式选定分析主题，开展对样本工单的主题分析。主题分析主要包括6个方面。

（1）总体显著性分析：主要分析各地市该主题工单数量占全省该主题工单数量的比例，展现各地市该主题工单的总体发生情况。

（2）分维度分析：从不同的维度对主题工单进行分析，主要包括的维度有：地市、区县、性别、年龄段、时间段等。通过多维度的数据比较，分析该主题工单在不同维度下所展现的特性。

（3）精准定位：发现该主题工单发生比例较高的区域，一般精准到区县，由此可对该主题工单在该区域的发生情况作进一步的对比分析。

（4）专家经验分析：对某主题工单在某区域的发生情况，从历史文化、经济发展、生活传统等区域特性对该主题工单反映的问题进行分析，是数据分析的重要辅助和归因说明。

（5）业务分析：通过与95598业务工单业务类型相结合进行关联分析，为制定解决问题的专业措

施提供数据支撑和分析依据，实现主题分析的最终目的。

（6）总结和建议：根据上述的多方位分析结果，总结提炼存在问题和隐患，并针对性地提出切实可行的措施建议。

3 主题分析示例

以方言问题为例对2015年95598业务工单进行挖掘分析。2015年，浙江省方言类工单占全部工单的比例为0.63%，方言主题的全省总体显著性较低，但区域差异较大，部分地区方言类工单相对较为集中。

宁波、绍兴和杭州3地的方言类工单占全省方言类工单的比例分别为25.88%、22.16%、14.09%，方言类工单发生相对较为集中。按照区县细分，排名最靠前的20个区县中，宁波占8个，绍兴占5个，杭州占4个，其中排名第一的为宁波奉化市。

从专家经验角度分析，奉化市出现方言类工单相对最为集中的情况有其历史文化原因。奉化市唐代建县，因以“民皆乐于奉承土化”而得名，语言形成历史悠久。同时，该地为蒋介石故居，在不少影视剧中，蒋介石的口音即为奉化口音，给当地中老年百姓留下了深刻印象。另外，奉化方言发音较为特殊，多句话时首字发音多数较重，这些特点使奉化方言显得“硬”、“拗”，难改也难懂。

从业务角度分析，奉化市发生的方言类工单，绝大多数为业务咨询类工单，而投诉工单所占的比例极低，仅占所有奉化市方言类工单总量的1.32%，同时显著低于宁波市方言类工单总量中投诉工单占比。

通过对方言类主题工单的分析研究，方言问题在全省范围内并不严重，仅在部分地区较为突出，但其中的投诉占比极低。因此，针对方言问题，无需开展区域专项治理，但为避免因语言误解而引发投诉，建议应对措施如下：当遇到方言客户时，客服人员应保持耐心，礼貌引导其说普通话或建议客户由身边会说普通话的人士代为表述；下发工单至地市，并做好相关备注，由懂当地方言的地市工作人员核实处理；梳理完善浙江省各地市、区县方言词库对照表，特别是方言类工单发生比例较高的地区。

4 结束语

本文详细阐述了95598业务工单主题分析方法，对运用的常态化文本挖掘技术、主题分析模型作了深入介绍，研究如何从不同类型的海量客户服务工单数据中准确筛选与识别客户感知诉求，挖掘感知诉求背后的问题。针对海量工单数据，利用文本挖掘、敏感词索引匹配等技术，实现工单自动化筛选分类并识别问题事件。同时结合“方言”主题示例的展示，呈现了主题分析方法在实际业务工作中应用的成效。

参考文献：

[1] 王丽坤，王宏，陆玉昌. 文本挖掘及其关键技术与方法[J]. 计算机科学，2002(12)：12-19.

[2] 潘钢. 上海移动公司客户投诉管理研究及应用[D]. 上海：上海交通大学，2013.

[3] 谌志群，张国煊. 文本挖掘研究进展[J]. 智能识别与人工智能，2005(1)：65-74.

[4] 陈阳，凌俊民，蒙圣光. 投诉数据智能挖掘分类管理系统[J]. 数字技术与应用，2011(6)：146-149.

[5] 李文波，孙乐，张大鲲. 基于Labeled LDA模型的文本分类新算法[J]. 计算机学报，2008(4)：620-621.

[6] 刘兴平，章晓明，沈然，等. 电力企业投诉工单文本挖掘模型[J]. 电力需求侧管理，2016(2)：57-60.

[7] Blei D, Ng A, Jordan M. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3)：993-1 022.