基于知识分层提取模型的服务台知识库建设①

曹逸峰, 陈晓伟

(中国农业银行股份有限公司 数据中心, 上海 200131)

摘 要: 针对传统运维知识库建设的不足, 提出了一种面向服务台的生产运维知识自动分层提取模型. 通过建立生产运维特征词库, 对事件工单的短文本进行向量化解析, 并利用改进的 KNN 算法实现事件短文本分类, 最终通过领域主题规则完成知识的发现. 将此方法应用到企业级服务台知识库建设中, 完成了由事件工单到知识的自动转化, 弥补了手工创建知识的缺陷, 促进了整个运维流程的自动化.

关键词: 运维知识库; 服务台; 事件短文本; 特征向量; KNN 算法

Construction of Service Desk Knowledge Base Based on Hierarchical Knowledge Extraction Model

CAO Yi-Feng, CHEN Xiao-Wei

(Agricultural Bank of China Data Center, Shanghai 200131, China)

Abstract: In view of the shortcomings of traditional operational knowledge base construction, proposed an automatic hierarchical knowledge extraction model to build the knowledge base on service desk. Firstly, we made quantitative analysis on short text of events worksheet by the establishment of production and operation feature lexicon. Then, we used the improved KNN text classification algorithm to classify the short text of events. Finally, we completed knowledge discovery by the field topic rules. We applied this method to the construction of the knowledge base of enterprise service desk. It can automatically complete the transformation of events worksheet to knowledge, which not only makes up the defect of the hand to create knowledge, but also promotes the automation of the whole operational process.

Key words: operational knowledge base; service desk; short text of events; feature vector; KNN algorithm

随着信息化由“技术驱动”向“业务驱动”转变, IT 部门的角色已经由技术支持变为信息服务的提供者. 近几年国外先进运维理念的涌入, 越来越多的企业开始意识到运维服务的重要性, 截止 2013 年 10 月中国大陆金融业已有 37 家公司通过 ISO20000 认证[1], 各个公司都建立起了面向流程管理的 IT 服务管理平台(服务台). 其中知识库的建设是服务台功能的重要组成部分, 在 IT 运维支持方面发挥了显著的作用. 目前大多数服务台都具备知识流程管理功能, 通过人工创建、提交、审核、更新、归档等流程实现企业 IT 部门知识的交流、共享和学习[2]. 然而知识库建设是一项繁琐艰巨的工作, 这种传统的知识管理流程存在很多缺陷, 比如日常运维中积累了大量事件工单, 其中蕴藏着丰富的运维知识经验, 由于数据量大, 若要人工去提取显然不现实, 因而导致了大量知识的流失. 另外, 繁琐的流程耗费人力, 常导致知识库形同虚设, 长时间得不到更新和维护.

知识库建设的重点是要突出实用, 尤其对一线人员, 当接到故障报警时能够通过知识库快速地查找, 方便地使用知识极为重要. 因此, 建立一套能处理大量数据信息并自动生成知识, 不需要过多人工维护的服务台知识库管理体系, 将具有重要的意义和研究价值.

服务台中流转的事件工单, 其本质是文本信息的载体, 在机器学习和数据挖掘领域有很多自动处理文本数据的方法[3], 因此, 实现知识自动生成的关键就是

① 收稿时间:2014-05-27;收到修改稿时间:2014-06-27

实现工单数据的文本化并选择合适的文本处理算法. 本文基于以上研究, 提出了一种基于向量化解析和短文本自动分类[4]的知识分层提取模型, 实现知识库知识的自动更新与维护.

1　知识分层提取模型

本文在传统知识库建设的基础上, 提出了一种面向非结构化事件文本信息的知识分层自动提取模型. 通过分层提取模型实现对生产运行异常事件库中的生产运维相关知识的自动提取, 整个模型结构如图 1 所示.

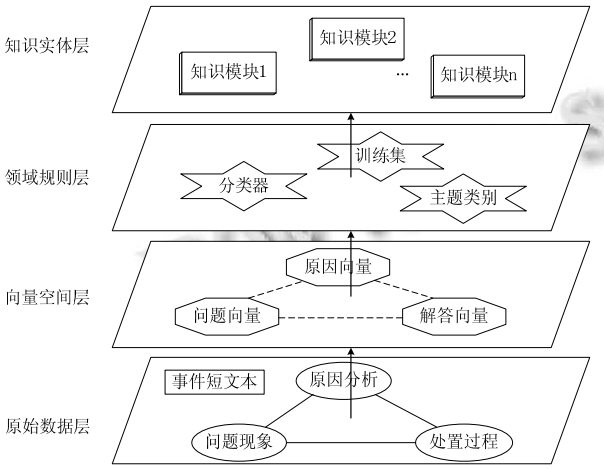

图 1　知识分层提取模型

知识分层提取模型由四层构成, 从下往上分别为原始数据层、向量空间层、领域规则层和知识实体层. 其中, 最底层是原始数据层, 主要指承载知识的生产运维相关的异常事件信息, 包括服务台事件工单和异常事件库问题库等; 第二层是向量空间层, 是将原始数据层经过一系列向量化处理得到的向量空间集, 以便于机器识别和处理; 第三层是领域规则层, 主要功能是将向量空间集进行分类处理并按主题规则保存处理后的数据, 包括分类器、主题类别、训练集等; 最顶层则是知识实体层, 它是按照领域规则提取得到的知识模块并对各个模块打上属性标签便于知识的快速检索. 四个层相互关联递进, 形成统一整体. 当新的生产运行事件发生时, 事件信息经过知识分层模型一系列处理, 便可自动提取得到蕴含其中的知识.

2　事件向量化解析

生产运行异常事件库中的信息都是以短文本形式保存的非结构化数据, 为了后期能进行机器自动分类处理, 必须对其进行预处理以便计算机能识别. 本文采用了基于向量空间模型的文本表示方法, 对生产运行异常事件进行分词、去禁用词、特征向量表示和特征扩展处理, 最终形成异常事件的向量化空间集.

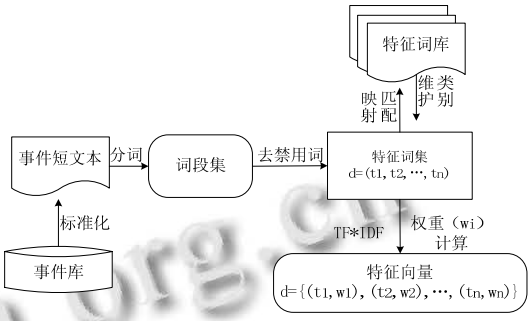

图 2　事件短文本向量化解析过程示意图

2.1 事件短文本预处理

对服务台中的生产运行的异常事件按问题现象、原因分析和处置过程三个部分进行标准化提取和保存, 形成标准化的事件库. 这样每个事件都由问题现象、原因分析和处置过程这三个短文本构成. 事件短文本预处理过程包括分词和去禁用词.

分词过程一般采用分词软件将短文本分成若干词段, 作为后续特征提取的基础. 目前中文分词方面的开源软件种类很多, 最常用的是中科院ICTCLAS分词系统[5], 但这些分词软件并不能很好地识别 IT 运维领域的专业词汇, 因此需要对分词软件进行改进, 将生产运维相关词条加入词库. 另外还可以对匹配模式进行调整, 如果一个长度为 a 的词条被一个长度为 b 的词条包含($a<b$), 且 b 长度词条涵盖所有 a 词条, 则只将 b 词条加入候选集合中. 例如“电子银行系统”这个词, 最初的分词系统会分成“电子”、“银行”和“系统”这三个词, 而改进的分词系统则可将其作为一个整体词汇, 这样可以更精准的保证语义.

去禁用词则是在分词后将词条碎片中大量的高频无意义的词语过滤掉, 保留能够代表文本特征的核心的名词、动词等. 这里对现有的中文停用词表稍作修整就能很好满足去噪音词汇的需求.

2.2 特征向量表示

事件短文本经过预处理后, 提取得到的特征值就组成特征词集, 表示成 $d=(t1,t2,\ldots,tn)$. 其中 d 代表一个事件短文本, t 为每个短文本中能代表文本主题属性的特征词汇.

特征词集中每个特征词与文本的关联程度不同, 即每个词代表文本主题的权重不同, 因此需要对这种关联关系进行量化处理. 如果一个词在某一文本中出现的频率越高, 则它越能代表词文本的主题含义, 对应的权重也越高. 基于这种统计规律的权重计算可以利用 TF-IDF 方法[6]实现, 其中 TF 代表特征项频率指特征词 *ti* 在文本 *d* 中出现的频率, IDF 表示逆向文本频率. 权重 *Wi* 计算公式如下:

$$Wi = tfi \times \log(\frac{N}{dfi}) \quad (1)$$

其中, *tfi* 表示特征词 *ti* 在文档 *d* 中出现的次数, *dfi* 表示包含特征词 *ti* 的文本数, *N* 总文本数.

计算特征词权重后, 每个事件短文本都可以表示成一个二维特征向量, 从而实现事件的向量化解析. 特征向量 $d=\{(t1,w1),(t2,w2),\ldots,(tn,wn)\}$.

2.3 特征词库建设与维护

特征词库包含了生产运维的各类主题词汇比如故障类别、系统信息等, 是长期运维经验的提炼. 最初的特征词库需要人工维护, 根据现有的运维体系, 归纳总结各类与生产运维相关的词汇, 形成运维专业领域词库, 并建立主题类别, 每一主题类别都由若干相关的特征词组成.

原始的特征词库建立后, 进行事件的向量化解析, 得到特征词集与特征词库进行映射匹配, 如果特征词集中的特征项在特征词库中无对应, 则将其纳入特征词库. 后期随着新事件不断增加, 从事件短文本提取得到的特征词集就可以不断充实丰富现有的特征词库.

另外, 事件的短文本在处理过程中与传统文本最大的差别就是短文本信息量少, 易导致生成的向量空间语义缺失和高维稀疏[7]. 利用特征词库已有的主题类别对特征词集进行特征项扩展能很好解决此类问题. 将短文本特征词集中的特征项与特征词库中主题类别进行匹配, 如果特征项与主题类别中某一类别相对应, 则可将这一类别中特征项扩展到短文本特征词集中.

3 事件文本分类与知识发现

文本分类的目的是为了挖掘有价值的文本知识, 实现文本知识的自动提取, 其基本思想是按照预先定义的主题类别, 为文档集合中每个文档确定一个类别, 方法有朴素贝叶斯、K 近邻、支持向量机、决策树、神经网络等. 事件短文本的分类与传统的文本分类类似, 也包括训练和分类两个过程. 本文采用 KNN 算法[8]构造分类器并对已有事件短文本进行分类训练形成若干训练集类别, 若有新事件文本加入, 分类器按照算法规则自动将其归类. 分类后的信息按照预先设定的规则进行筛选从而达到知识发现的目的, 同时根据知识特点给短文本信息打上不同的属性标签, 方便知识的检索.

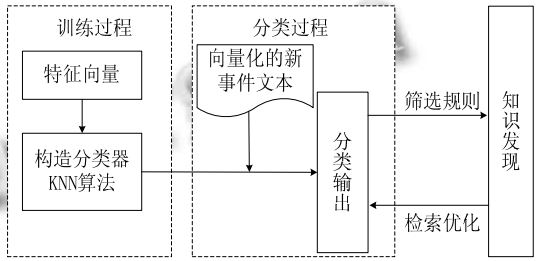

图 3 知识分类提取过程示意图

3.1 事件短文本自动分类

文本自动分类的关键是构造分类函数即分类器, 利用分类函数将待分类文本划分到相应类别空间中. 由于 KNN 算法具有简单、高效且适用样本容量大、类域交叉分类等优点, 本文采用KNN算法对事件短文本进行类别学习. 该算法思路是: 给定新事件短文本后, 在训练文本集中找到与该新文本距离最近的 *k* 篇文本, 根据这 *k* 篇文本的类别判断新文本所属的类别. 本文在现有KNN算法的基础上对其稍作改进, 具体实现步骤如下:

① 初始化训练集, 即根据特征词库重新描述训练文本向量;

② 向量化解析, 新的事件文本到达后, 预处理并进行特征项扩展, 最后向量化表示;

③ 相似度计算, 在训练文本集中选出与新文本最相似的 *k* 个文本, 计算公式:

$$Sim(di,dj) = \frac{\sum_{k=1}^{M} Wik \times Wjk}{\sqrt{(\sum_{k=1}^{M} W_{ik}^2)(\sum_{k=1}^{M} W_{jk}^2)}} \quad (2)$$

其中 *k* 是一个经验值, 需要不断调整才能使分类结果最优;

④ 类别筛选, 在新文本的 *k* 个邻居中, 计算每类的权重, 计算公式:

$$p(x,Cj) = \sum_{di \in knn} Sim(x,di)y(di,Cj) \quad (3)$$

其中 x 为新文本的特征向量, $Sim(x,di)$为相似度公式同公式 2, $y(di,Cj)$为类别属性函数[9], 若 di 属于类 Cj 则函数值为 1, 否则为 0;

⑤ 比较类的权重, 将新文本分到权重最大的那个类别中.

3.2 分类结果筛选

事件短文本分类成功后, 在类别内部及类别间根据设定的筛选规则即可实现知识信息的提取. 筛选规则的设定可以根据生产运维的需要进行灵活设置, 例如事件短文本中针对异常现象相近的“问题现象”短文本, 对其对应的“原因分析”短文本进行分类, 并按次数多少进行排序, 排在前面的就是最可能的故障原因. 同理对原因相近的“处置过程”短文本, 分类后按处置时间长短排序即可得到效率最高的处置方案.

在实现知识的筛选提取的同时, 还需要快速方便的实现知识的检索, 传统的知识检索都是按照树状结构一级一级地实现查询的, 这种逐级打开类目的检索方式大大降低了运维人员的效率. 本文中引入了知识的属性标签 Tag 的概念[10], 同一个知识可以有多个属性标签. 这些属性标签可以用相应的特征词代替, 与类目相比更加离散、灵活, 也缩减了类目的深度, 运维人员只要检索相应的关键词就能快速定位相关知识, 同时也解决了类目交叉的问题.

4 服务台知识库构建

传统的服务台中, 事件管理流程与知识管理流程基本都是相互独立、隔离的. 事件解决关闭后, 往往未能继续利用, 仅仅限于后期的查阅, 而知识建设也仅仅限于人工的提交和审核, 缺少其它自动处理渠道. 本文在知识库的构建上采用了一种闭环流程结构, 如图 4 所示. 在事件、问题管理流程中, 建立与知识库流程的接口, 实现事件与知识的互通. 通过事件工单可以提取有价值的知识, 同时通过知识库也能快速查找到最佳故障解决方案和对应的事件工单信息.

当运维人员接到故障报警后, 服务台会自动分析提交事件的标题, 在知识库中查找统计, 并通过相似度计算分析, 将最可能贴近的处置方案推荐给支持人员. 这将大大节省支持人员的时间, 即使需要支持人员来手动收索, 内置的知识库也要比跳转到另一个页面查找知识库方便的多. 当一个事件解决后, 相关的事件单信息就会通过知识分层提取模型, 经过向量化解析和分类处理, 并按一定的领域主题规则抽取有价值的知识, 自动充实知识库.

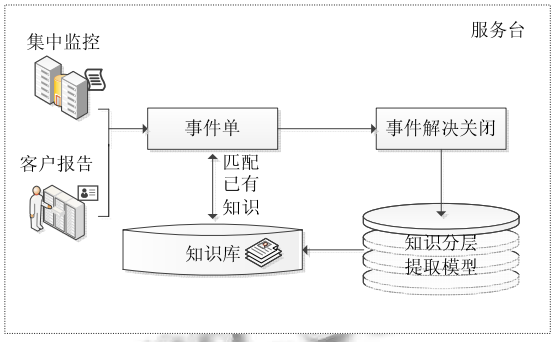

图 4 服务台知识库工作示意图

5 结束语

本文创造性地将机器学习和数据挖掘相关技术运用到运维自动化领域, 通过对传统文本分类技术进行改进, 提出了一种针对事件短文本的知识自动分层提取模型, 并阐述了具体的实现过程. 其中特征库的建立和分类算法的调整与改进是工作的重点和难点. 特征词库的好坏直接关系到分类的结果, 但是特征库建设的工作量非常大, 在模型建立的初期需要人工收集特征词, 整理主题类别. 如果企业前期已有与生产运维操作方面的经验总结和文档则可以大大降低工作量. 另外在分类算法的调整与改进方面, 需要选择合适的分类算法. 但并不是所有的分类算法从一开始就能得到符合要求的结果, 需要对其进行改进以适用于事件短文本的分类, 同时针对分类结果不断进行校正调整, 逐渐提高分类的准确率.

智能化的知识库能为运维工作提供有力的支撑, 不仅能实现经验分享, 还能为故障的解决提供实时参考. 随着运维自动化水平的不断提高, 这种自动化的知识提取模型将有广阔的应用前景, 尤其是在突发事件的事前预警和事中控制方面, 能提供很好的决策支持, 提高事件处置效率.

参考文献:

1 Bin-Abbas H, Bakry SH. Assessment of IT governance in organizations: A simple integrated approach. Computers in Human Behavior, 2014, 32: 261–267.

2 王毅.基于 ITIL 的 IT 服务运维管理体系研究.硅谷,2014, (3):149–151.

3 李小庆.基于数据挖掘与知识发现的银行决策分析.金融科

技时代,2014,22(1):56–59.
4 范云杰,刘怀亮.基于维基百科的中文短文本分类研究.现代图书情报技术,2012,(3):47–52.
5 张华平,刘群.汉语词法分析系统 ICTCLAS,2010.
6 Wang XL, Yang L, Wang D, Zhen LH. Improved TF-IDF keyword extraction algorithm. Computer Science & Application, 2013, 3(1): 64–68.
7 林伟,孟凡荣,王志晓.基于概念特征的语义文本分类.计算机工程与应用,2011,47(28):139–142.
8 孙荣宗,苗夺谦,卫志华,李文.基于粗糙集的快速 KNN 文本分类算法.计算机工程,2010,36(24).
9 江涛,陈小莉,张玉芳,熊忠阳.基于聚类算法的 KNN 文本分类算法研究.计算机工程与应用,2009,45(7):153–155.
10 林晓燕,高良才,汤帜.中文电子文档的数学公式定位研究.北京大学学报(自然科学版),2014,(1):17–24.