第 40 卷 第 6 期 2016 年 3 月 25 日
DOI：10.7500/AEPS20150812003
电力系统自动化
Automation of Electric Power Systems
Vol.40 No.6 Mar.25，2016

文本信息挖掘技术及其在断路器全寿命状态评价中的应用

邱 剑[1]，王慧芳[1]，应高亮[2]，张 波[2]，邹国平[3]，何奔腾[1]
（1. 浙江大学电气工程学院，浙江省杭州市 310027；2. 国网金华供电公司，浙江省金华市 321017；
3. 国网浙江省电力公司电力科学研究院，浙江省杭州市 310014）

摘要：电网企业记录了大量故障与缺陷中文文本，这些文本蕴藏了丰富的设备健康信息。但迄今为止，鲜有电力领域的文本信息挖掘技术研究。以断路器全寿命状态评价为应用研究背景，探索了电网中文文本挖掘方法。首先，根据断路器状态评价的研究现状，提出了构建文本挖掘与全寿命状态评价模型的关键问题。然后，构建了包含文本挖掘信息的全寿命状态评价模型，通过基于隐马尔可夫法（HMM）的文本预处理与向量化、自主区间搜索 k 最近邻（KNN）算法的文本分类和比率型状态信息融合模型完成了断路器全寿命健康状态指数的展示。最后，采用某电网公司实际缺陷文本构建算例。算例表明，文本挖掘技术实现了相似缺陷的相关性学习，比率型信息融合模型能更全面真实地展示健康状态评价的历史流。

关键词：全寿命状态评价；检修消缺；断路器；文本挖掘；隐马尔可夫法（HMM）；k 最近邻（KNN）

0 引言

《中国电力大数据白皮书》指出：多样性是电力大数据最重要的特性之一，包括结构化、半结构化和非结构化数据[1-2]。按照大多数信息化企业的数据管理经验，结构化数据约占数据总量的 20%，它们能被关系型数据库处理，但其余 80%的半结构化和非结构化数据则难以用关系型数据库表达[3]。

在电力系统中，大量非结构化数据分布在设计、安装、运行、检修与退役等整个全寿命周期环节中，主要包括文本、音频和图像等[4]。以检修与维护环节为例，电网企业积累了大量的检修试验记录、巡检消缺记录、故障与缺陷描述报告和事件顺序记录（SOE）等。这些日志和报告主要以夹杂数字、字母符号的中文短文本（以下简称文本）的形式出现，蕴藏着丰富的设备历史运行状态信息、检修效果信息和可靠性信息等，对客观评价设备健康状态发展过程大有裨益。然而，由于文本具备多歧义、难切分、模糊性、多噪声等特点，上述信息还没有得到充分挖掘。

计算机和语言学家专门为此开辟了一门交叉研究领域——自然语言处理（NLP），致力于解决文本和语言处理中的关键性难题。NLP 在语音识别、互联网、搜索引擎、电子医疗报告分析等处均获得了成功的应用[5-6]。但在这些领域中，NLP 通常采用词句长度、精度和词频等作为特征指标，适用于长文档的信息检索和主题分类，因此不能完全照搬至工业界的短文本挖掘。在电力领域，国外已有针对电缆故障工单和操作票的研究[7]，但由于缺乏设备亚健康缺陷信息，仅仅分析了事后故障记录文本，仍处于试探性研究阶段。

迄今为止，尚无有关电网中文文本处理的研究公开发表。一方面是因为专业领域的中文文本处理是 NLP 中最难的问题之一，研究者不仅需要解决中文难切分与模糊性问题[6]，还需掌握扎实的专业背景知识，将电力专业词汇正确地融入文本挖掘模型中；另一方面是因为挖掘这类文本信息，不能仅仅关注于单个主题或单一时间剖面，否则容易造成严重的数据断层，更需要以一种总体的视角进行事物间的综合关联性分析，揭示多因素的相关影响机制，处理好多样化数据的融合[4,8-9]。

鉴于此，本文以断路器全寿命状态评价为应用研究背景，研究故障与缺陷文本挖掘的关键性问题，探索文本挖掘的具体实施流程，不仅为电网其他中文文本信息挖掘提供技术参考，也为其他电力设备实现全寿命状态评价提供范例。

收稿日期：2015-08-12；修回日期：2015-11-28。
上网日期：2015-12-28。
国家电网公司科技项目。

1　断路器状态评价研究现状及存在问题

国家电网公司于2008年颁布了断路器评价导则[10]，许多研究以此为基础发展了一系列状态评价法，如专家系统法、层次分析法、模糊综合评判法、模糊聚类分析法和雷达图法等。上述方法在专家组参与评价的前提下，具有一定可信度，适用于评价没有量化模型支撑的模糊现象[11]。然而，由于不可能把每台设备的每条缺陷都送至专家组进行评价，上述方法可操作性较差。另一类状态评价法是针对实时监测数据进行设备故障诊断分析，主要有神经网络、支持向量机、小波分析和经验模态分解法。这些方法通常将高维监测数据映射为低维特征向量，再利用分类器进行故障识别。上述方法在处理结构化数据时，具有一定优势，适用于濒临故障的紧急状态预警和事后故障原因分析[12]。然而，由于经济成本和技术发展等原因，多数断路器还没有装配状态实时监测装置，因此上述方法应用受限。

目前断路器依据导则定期实施评价，通常情况下评价周期为一年，评价过程包括3步：①根据导则标准对部件状态量进行扣分；②按公式计算部件得分并确定部件定性评价；③确定设备整体评价。导则指出，状态量的获取包括经停电试验、带电检测和在线监测获取的结构化数据，因此可认为基于导则的状态评价已实现结构化数据的利用。

在未来很长一段时间内，人工巡检和带电检测仍将是电力设备最主要的“监测”方式，对于某些模糊性较强的亚健康缺陷，巡检人员很难给出精确的量化参数，只能用文本尽可能将缺陷记录下来。与“稀有”故障事件不同，几乎所有设备都会发生亚健康缺陷。通常情况下，对于无法带电消缺或自行消缺的情形，允许设备带缺陷运行一段时间[8]。但是，在此期间需要加强巡视和跟踪状态评价，通过后续建模分析寻找最优的消缺和检修时机。

为了充分融合不同方式获得的多源状态信息、不同时间序列获得的历史纵向信息以及不同数据颗粒度的异构信息，在含文本信息挖掘的断路器全寿命状态评价研究中，需要解决：①夹杂数字与特殊符号的中文文本的预处理与向量化过程；②设备群中相似性、关联性的学习与挖掘；③历史缺陷信息和状态评价信息在全寿命状态评价中的关系，以及如何实现融合；④包含具体部件缺陷信息的设备健康状态数据流展示。

本文将解决上述问题，实现对中文文本的信息挖掘，并将挖掘结果与基于导则的状态评价结果融合，实现设备全寿命健康状态的展示。

2　包含文本挖掘信息的全寿命状态评价模型

2.1　故障与缺陷记录内容分析

故障记录通常来自于事后故障分析、停电检修试验报告和SOE。缺陷记录通常来自运行及检修人员的巡视记录和带电检测记录[10]。

1)设备基本信息。包括设备型式、电压等级、厂家、出厂时间、名称和所在变电所等。此部分信息通常记录准确。

2)故障/缺陷文本。通常包括故障/缺陷发现或发展时间、故障/缺陷描述、缺陷评判等级。事件发现或发展时间通常有准确记录；故障/缺陷描述的详细程度因案例而异，一般记录完整。缺陷评判等级由现场人员针对具体的描述文本参照国标[10]给出，结果受主观性影响而因人而异。

3)检修/消缺文本。通常包括检修/消缺时间及其效果、技术原因、相关责任人等。上述内容在记录完整性和准确性上差强人意。

由此可见，故障/缺陷文本是相对可靠、信息量最丰富的内容。事件发现或发展时间序列数据结构相对简单，容易向量化处理，已有不少研究者采用时间序列理论(TSM)进行研究[13-14]。本文旨在充分挖掘文本中蕴藏的信息，以期获得比TSM更加丰富的结果。故障/缺陷事件常用单条短文本来描述，一般包括发生部件(主语)及状态量特征(谓语/宾语)，而缺陷程度可能由精准的量词或模糊性形容词/副词(即定语/状语)描述，也可能缺失。通常，现场运维人员会根据相关状态评价标准对缺陷进行评判，给出一般、重要或紧急这3个缺陷等级。

2.2　健康状态指数

为了将历史文本挖掘信息与基于导则[10]的状态评价信息充分融合，并实现设备健康状态的量化表征，本文定义标准化的健康状态指数为H，$0\leqslant H\leqslant 1$，0和1分别代表设备故障和完全健康，且$H=\{H_{\text{text}},H_{\text{score}}\}$，其中，$H_{\text{text}}$为通过故障缺陷文本挖掘得到的量化指标($0\leqslant H_{\text{text}}\leqslant 1$)，3个缺陷等级所对应的具体$H_{\text{text}}$区间，将通过改进KNN算法自主搜索求取；$H_{\text{score}}$为通过状态评价方法获得的部件得分值$S$转化获得，即

$$H_{\text{score}}=\frac{S-S_{\min}}{S_{\max}-S_{\min}} \tag{1}$$

式中：$S_{\max}=100$，$S_{\min}=0$[10]。

2.3　包含文本挖掘信息的全寿命状态评价模型

包含文本挖掘的全寿命状态评价框架如图1所示，中文故障缺陷文本的挖掘主要包括文本预处理

和文本分类算法,得到用 H_{text} 表征的基于缺陷文本的量化健康状态评价指数。随后基于信息融合模型,与基于评价导则的健康状态指数 H_{score} 进行融合,得到全寿命周期 H 的数据流。

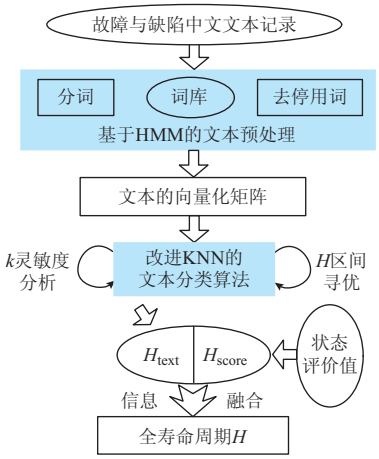

图 1　包含文本挖掘的全寿命状态评价框架
Fig.1　Framework of lifecycle condition assessment including text mining process

3　故障缺陷文本挖掘模型

3.1　基于 HMM 的中文文本预处理流程

中文与英文的 NLP 不同,中文语句字与字之间没有空格所形成的自然分界,因此先要对文本进行分词,形成具有独立意义的合理的词语序列。这也是中文 NLP 领域最大的难点之一[6]。

参考状态评价标准[10]、常用缺陷词语[15]和专家推荐的未登录词等资料后,编纂了"断路器故障与缺陷基本词库"。其中,每一个词由"词/词性"组成(以下简称词库,部分词库见附录 A 表 A1)。未登录词主要指某些部件与特征的缩写词、俗称和数字型合成词。

鉴于断路器缺陷文本为短文本,且句法结构较明确,文本预处理环节仅保留分词、词频统计、去停用词和文本向量化 4 个步骤。

1)分词。以编撰的"断路器故障与缺陷基本词库"为基础,采用 HMM 技术对文本进行初始分词。许多研究者针对不同开发平台公布了开源代码与分词系统,本文采用 Viterbi 算法,详见文献[5-6]。

2)词频统计。对所有分词进行词频统计,并从大到小排序得到词频序列。

3)去停用词。剔除停用词库中的停用词。为了实现数据清理,将所有与附录 A 表 A1 中词库无关的词均视为停用词,部分停用词表见附录 A 表 A2。

4)文本向量化。每个词对应向量空间中的一维,基于词频排序的不重复词序列构成完整的向量空间 $\boldsymbol{W}_{\text{ALL}}=(w_{ij})_{I\times J}$。其中:$w_{ij}=0$ 或 1,当 $w_{ij}=1$ 时,该条文本包含该词向量,反之则为 0;I 和 J 为向量维数。

3.2　基于自主区间搜索 KNN 的文本分类算法

文本预处理完成了文本的向量化过程。为实现无需现场人员或专家给定缺陷等级,将任意缺陷文本输入模型即可自动获取 H_{text} 的目的,根据词库和短文本的特点,对 k 最近邻(KNN)算法进行改进,实现自适应寻优分类。

KNN 算法的核心思想为:如果一个样本在特征空间内的若干个最相邻的样本中的大多数属于某一个类别,则该样本也属于这个类别,并具有这个类别上样本的特性[16]。因此,KNN 化后的 H_{text} 不仅是缺陷等级的归一化,还会涵盖类似设备及其缺陷的共性特征,是一个更加精细化的定量指标。

改进的 KNN 算法包括如下步骤。

1)计算测试集与训练集的相似度。其中,测试集为待归类的已向量化的缺陷文本集合,训练集是已根据专家组经验完成归类好的缺陷文本集合,也称为缺陷文本库。

2)计算测试集缺陷文本对应的 H_{text}。按照测试文本 i 与训练集中文本的相似度进行排序,选出最相似的 k 条文本,然后基于相似缺陷特征,计算测试文本对应的健康指数 H_{text}。

3)计算 H_{text} 的分类区间。经过 KNN 相似度学习,H_{text} 的信息量与数据颗粒度均发生变化。常规 KNN 算法的分类区间是人为设定的,本文以分类准确度最大化为目标,提出 H_{text} 区间搜索算法。

4)k 值寻优。由于不同容量的训练库,分类准确率最高时 k 值可能并不一致。而传统 KNN 算法中 k 值由人为主观选定,本文提出灵敏度算法进行 k 值寻优,得出该训练样本下的最高分类率。

上述步骤的详细计算公式与算法见附录 B。

4　比率型全寿命状态评价信息融合模型

图 2 为采用 H 刻画的事件与时间序列过程。图中:$H_{(i-1)+}$ 为单位健康周期开始时的健康状态指数,一般采用上次状态评价值或新投运或检修完毕后的值,通常接近为 1;H_i 为 i 事件发生(即 $t_{\text{age}i}$)时的健康状态指数,通常为故障与缺陷文本挖掘的结果 H_{text};$H_{(i+1)-}$ 为带缺陷运行直到检修消缺前的健康状态指数,由比率型劣化模型估计获得。

设备于役龄 $t_{\text{age}i}$ 时发生某缺陷事件,文本经 KNN 分类或基于导则状态评价,得到健康指数为 H_i。经带缺陷运行 T_{Delay} 时段后,于 $t_{\text{age}i+1}$ 时刻采取消缺或检修措施,措施前后 H_{i+1} 将发生突变。

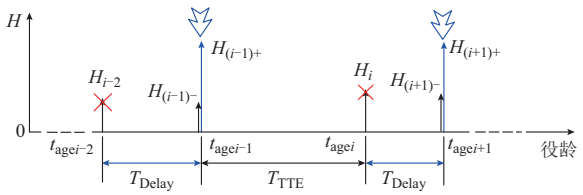

图 2　采用 *H* 定量刻画的事件与时间序列图
Fig.2　*H*-specified event and time sequence process

为此，对缺陷事件定义单位健康周期(SHC)的概念：从设备投入使用开始(包括初始投运、检修/消缺或改造升级完投运，或状态评价为正常后投运)，经 T_{TTE} 时段缺陷事件发生，带缺陷运行 T_{Delay} 时段导致事件程度升级，最后采取措施事件结束的整个过程(措施包括检修、消缺、更换和退役等)。图 2 中，第 i 个 SHC 序列集合为：$I_{SHCi}=\{(t_{agei-1}, H_{(i-1)+}), T_{TTE}, (t_{agei}, H_i), T_{Delay}, (t_{agei+1}, H_{(i+1)-})\}$。

全寿命周期过程中存在着若干个单位健康周期，该复杂序列涉及设备状态劣化升级、延迟消缺/检修时间、带缺陷运行、设备/部件役龄等因素。因此，需要对该复杂过程进行建模。

根据故障或缺陷事件对设备可靠性的影响机制，建模过程参考了含协变量的指数模型、威布尔模型、比率故障率模型等理论[17]。其中，比例故障率模型(PHM)由著名统计学家 Cox 于 1972 年提出，故又称为 Cox-PHM 模型，该模型的基本假设是每个故障事件对个体的故障风险产生比率型的影响，并采用偏似然估计理论估计故障风险概率，被广泛应用在生物医学领域。近年来也有学者将其引入电力系统的研究中[18]，提出如下假设：①每个历史缺陷事件对设备的健康状态产生比率型影响 $\Delta(H)$；②带缺陷运行期间 T_{Delay}，该事件对设备状态产生持续的比率型劣化作用 $\Delta(H)$；③检修/消缺成功，假设措施实施后设备如新，即 t_{agei+1} 时 $H_{(i+1)+}=1$。

基于上述假设，提出了递推关系的比率型状态融合模型(PHFM)用于描述 I_{SHCi}，其状态评价集为 $\{H_{(i-1)+}, \Delta(H_{i-1}), H_i, \Delta(H_i), H_{(i+1)-}\}$，描述为

$$\begin{cases} H_{(i+1)-}=H_i\exp(\delta(t_{agei+1})(H_i-1)) \\ \Delta(H_{i-1})=\dfrac{H_i-H_{(i-1)+}}{t_{agei}-t_{agei-1}} \\ \Delta(H_i)=\dfrac{H_{(i+1)-}-H_i}{t_{agei+1}-t_{agei}} \end{cases} \tag{2}$$

式中：$\Delta(H_{i-1})$ 和 $\Delta(H_i)$ 分别为缺陷事件发生前、后 2 个阶段的健康状态指数劣化率；$\delta(t_{agei})$ 为役龄的示性函数，不带缺陷运行时 $\delta=0$，带缺陷运行时 $\delta=t_{agei}/A$，A 为设备预期服役年限，本文参考国标令 $A=100$ 年。

5　算例分析

5.1　文本预处理与分类结果验证

本文选取了华东某地区电网公司的 454 台有过历史缺陷的断路器作为研究对象。从 2007 年 1 月 1 日至 2015 年 1 月 1 日期间，累计发生缺陷 953 次(即从电网生产管理系统中导出缺陷文本 953 条)，其中，一般/重要/紧急缺陷条数为 271/414/268 条。参考电力公司按年制定检修计划的惯例，选取 2007 年1 月 1 日至 2013 年 12 月 31 日的 744 条文本作为训练样本，2014 年 1 月 1 日至 2015 年 1 月 1 日的 209 条文本作为测试样本。对测试样本按照图 1 的框架进行文本挖掘。

借助基于层次 HMM 的中文开源词法分析系统(ICTCLAS)[6]，导入编纂的“断路器故障与缺陷基本词库”，完成初始分词。部分分词结果见附录 A 表 A3。其余文本预处理步骤通过编写 MATLAB 程序完成。将 209 条测试文本的原来由运维人员评价的等级标签去掉后，采用改进的 KNN 算法进行文本分类和归一化后，最优 H_{text} 分类区间为：

$$H_{text}=\begin{cases}[0.53,0.66] & \text{紧急} \\ (0.66,0.82) & \text{重要} \\ [0.82,0.88] & \text{一般}\end{cases} \tag{3}$$

具体分类结果如图 3 所示。可见，经过改进 KNN 文本分类算法，能够实现 H 的自动化计算，且融合了相似缺陷文本的加权指标，得到的 H_{text} 有更精细的数值，而非 3 个定性的评判等级。此外，分析 H 数值所属区间与由运维人员评价的等级标签，可得到分类准确度。

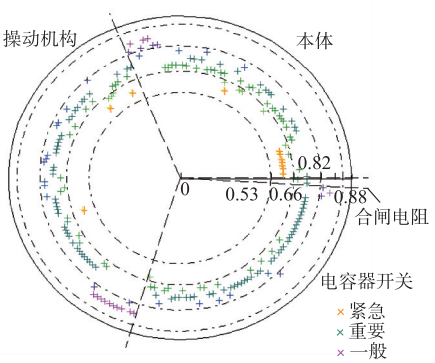

图 3　断路器历史缺陷文本分类
Fig.3　Historical defect text classification of circuit breaker

灵敏度分析如图 4 所示。通过 k 灵敏度算法，得出在 $k=3$ 的情况下，分类准确率最高，即 88.52%。3 个等级的误分类矩阵如表 1 所示。

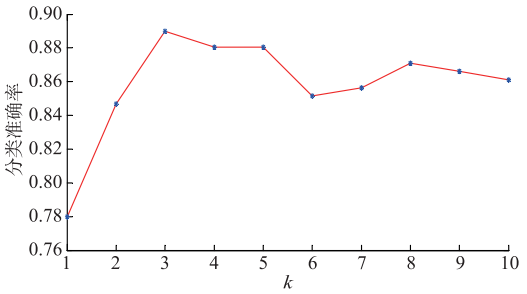

图 4　k 灵敏度分析
Fig.4　k sensitivity analysis

表 1　测试集的误分类矩阵
Table 1　Misclassification matrix of testing sets

原始等级	被分入相应等级的矩阵数		
	紧急	重要	一般
紧急	14	14	0
重要	7	149	2
一般	0	1	22

经分析，H 的误分情况主要发生在新缺陷与多模式失效的情形下。表 1 中，由于训练样本中“紧急”案例较少，而新一年的测试样本中新增了几例从未发生过的“紧急”缺陷，并且还存在并发缺陷事件，进而导致了将“紧急”误分为“重要”的情况。而对于曾发生过类似的缺陷与故障，本模型能够实现近 95％的高准确度分类。

附录 A 表 A4 列举了几个典型缺陷事件的示例，对本文评价结果与导则评价结果进行了对比。其中，完整意义上基于导则的评价分 S 需要针对所有部件进行全方位打分，因此，在仅告知某缺陷文本的情况下，只能给出某部件的扣分值。从表中可以发现，由本模型得到的分类等级无误，但是本模型还无法实现对文本中所有特征、数值进行精确识别并计算精细化的 H，因此目前正在尝试引入语义分析和深度学习技术进行解决。

另外，由于现场情况复杂多变，少数缺陷在国标导则中无法找到扣分依据，本文的相似度学习算法仍能得出 H，为状态评价提供一定的参考，也有助于导则的进一步完善。

5.2　断路器全寿命状态评价展示

选取某台断路器，系统记录了其役龄 16 年之后的检修与消缺记录，其电容器开关发生过 4 次“重要”缺陷，目前正带“重要”缺陷运行中。将该断路器所有文本和状态评价分值输入图 1 框架中，得到该断路器电容器开关的历史健康状态指数流见图 5。

图中绿点为电网企业历年基于导则[10]给出的电容器开关状态分，经式(1)转化后得到 H_{score} 值；蓝点为通过缺陷文本挖掘获得 H_{text}，以及代入融合模型式(2)得到的综合 H，特别地，最后的红点为目前带缺陷运行状态。

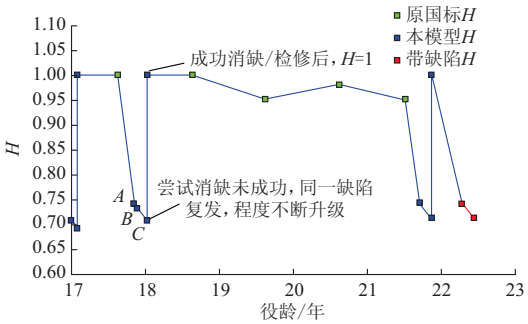

图 5　面向历史数据流展示的设备全寿命状态评价
Fig.5　Asset lifecycle condition assessment for historical data stream demonstration

实线为经过上述两部分融合后展示的该断路器电容器开关的 H 状态流，具有如下特点。

1）经过改进 KNN 文本分类算法，能够实现从文本到 H 的自学习计算，且自动融合了基于导则的状态评价值，实现了自动化展示。

2）不仅能反映每次评价结果，还能反映发现缺陷和维修完毕时刻，以及不同缺陷事件之间的 H 值差异。更重要的是，能体现发现缺陷前亚健康状态期间和带缺陷运行期间 H 的劣化趋势。如该台设备在役龄 16.96 年时发生缺陷，推迟了 0.074 年执行消缺处理，在带缺陷运行期间 H 的劣化率 $\Delta(H_1)=-0.205$。

3）可以清晰表达复发缺陷和消缺不成功情况。如在役龄 17.82 年时，该断路器发现缺陷（A 点），随后在巡检时发现了同一缺陷未消缺成功（B 点），因此 AB 间 H 的劣化速度很快，至役龄 18 年时，才将该缺陷消除（C 点），此时 H 值已劣化较多。

6　结语

本文主要贡献在于以下几个方面。

1）对电网的文本信息处理模型进行了探索，为电网中的相关文本信息挖掘提供了方法。

2）实现了从故障缺陷文本到 H 的自学习映射，改变了运维人员主观评价故障/缺陷等级的方式，并由于挖掘过程中实现了同台设备的历史信息、群体类似信息相似性、关联性的学习，使得 H 的结果更精准，符合电网企业精细化管理的发展要求。

3）通过定义标准健康状态指数实现了基于文本挖掘信息的状态评价结果的融合，通过采用比率型状态融合模型实现了全寿命状态信息流的展示，为电网企业实现全寿命周期管理提供了基础。

在研究过程中，还存在一些难点未得到充分解

决。首先，词库型自然语言处理技术在未登录词识别方面较弱[5-6]，需要从电力设备特性及其故障机制出发，建立电力语义模型，完成新缺陷的识别，提高分类准确率。其次，在文本预处理过程中发现，缺陷文本的规范化录入和数据质量仍有待提高，本文的词库将有助于电网公司对该工作提出进一步的规范化要求。相信随着电力语料库的不断丰富，该模型的学习能力还能够继续提升，届时，可以引入更为先进的机器学习技术挖掘出更有价值的信息。

附录见本刊网络版（http://www.aeps-info.com/aeps/ch/index.aspx）。

参考文献

[1] 中国电机工程学会信息化专委会.中国电力大数据发展白皮书[M].北京：中国电力出版社，2013.

[2] 曲朝阳，陈帅，杨帆.等.基于云计算技术的电力大数据预处理属性约简方法[J].电力系统自动化，2014，38(8)：67-71. DOI：10.7500/AEPS20130601001.
QU Zhaoyang，CHEN Shuai，YANG Fan，et al. An attribute reducing method for electric power big data preprocessing based on cloud computing technology[J]. Automation of Electric Power Systems，2014，38(8)：67-71. DOI：10.7500/AEPS20130601001.

[3] 罗学礼，徐树振，王森，等.电力企业的非结构化数据检索研究[J].计算机与数学工程，2014，42(4)：729-733.
LUO Xueli，XU Shuzhen，WANG Sen，et al. Research on unstructured data retrieval of electric enterprise[J]. Computer & Digital Engineering，2014，42(4)：729-733.

[4] 张东霞，苗新，刘丽平，等.智能电网大数据技术发展研究[J].中国电机工程学报，2015，35(1)：2-12.
ZHANG Dongxia，MIAO Xin，LIU Liping，et al. Research on development strategy for smart grid big data[J]. Proceedings of the CSEE，2015，35(1)：2-12.

[5] JURAFSKY D，MARTIN J H. Speech and language processing：an introduction and natural language processing，computational linguistics and speech recognition[M]. New Jersey，USA：Prentice Hall Press，2005.

[6] 张华平，高凯，黄河燕，等.大数据搜索与挖掘[M].北京：科学出版社，2014.

[7] RUDIN C，WALTZ D，ANDERSON R N，et al. Machine learning for the New York City power grid[J]. IEEE Trans on Pattern Analysis and Machine Intelligence，2012，34(2)：328-345.

[8] 彭小圣，邓迪元，程时杰，等.面向智能电网应用的电力大数据关键技术[J].中国电机工程学报，2015，35(3)：503-511.
PENG Xiaosheng，DENG Diyuan，CHENG Shijie，et al. Research on development strategy for smart grid big data[J]. Proceedings of the CSEE，2015，35(3)：503-511.

[9] 曲广龙，杨洪耕.基于梯形云模型的电能质量数据关联性挖掘方法[J].电力系统自动化，2015，39(7)：145-150. DOI：10.7500/AEPS20140801005.
QU Guanglong，YANG Honggeng. A correlation mining method for power quality data based on trapezoidal cloud model[J]. Automation of Electric Power Systems，2015，39(7)：145-150. DOI：10.7500/AEPS20140801005.

[10] 国家电网公司.SF6高压断路器状态评价导则：Q/GDW 171—2008[S].2008.

[11] LIN P C，GU J C，YANG M T. Intelligent maintenance model for condition assessment of circuit breakers using fuzzy set theory and evidential reasoning[J]. IET Generation，Transmission & Distribution，2014，8(7)：1244-1253.

[12] 孙来军，胡晓光，纪延超.小波包-特征熵在高压断路器故障诊断中的应用[J].电力系统自动化，2006，30(14)：62-65.
SUN Laijun，HU Xiaoguang，JI Yanchao. Fault diagnosis for high voltage circuit breakers with characteristic entropy of wavelet packet[J]. Automation of Electric Power Systems，2006，30(14)：62-65.

[13] 严英杰，盛戈皞，陈玉峰，等.基于时间序列分析的输变电设备状态大数据清洗方法[J].电力系统自动化，2015，39(7)：138-144.DOI：10.7500/AEPS20140111003.
YAN Yingjie，SHENG Gehao，CHEN Yufeng，et al. Cleaning method for big data of power transmission and transformation equipment state based on time sequence analysis[J]. Automation of Electric Power Systems，2015，39(7)：138-144. DOI：10.7500/AEPS20140111003.

[14] 钟锦源，张岩，文福拴，等.基于时间序列相似性匹配的输电系统故障诊断方法[J].电力系统自动化，2015，39(6)：60-67.DOI：10.7500/AEPS20140815001.
ZHONG Jinyuan，ZHANG Yan，WEN Fushuan，et al. A fault diagnosing method in power transmission systems based on time series similarity matching[J]. Automation of Electric Power Systems，2015，39(6)：60-67. DOI：10.7500/AEPS20140815001.

[15] 浙江省电力公司.电网变电一次设备缺陷用语规范：Q/ZDJ 44—2005[S].2005.

[16] HASTIE T，TIBSHIRANI R，FRIEDMAN J. The elements of statistical learning：data mining，inference，and prediction[M]. 2nd ed. Berlin，Germany：Springer，2008.

[17] GORJIAN N，MA L，MITTINTY M，et al. A review on reliability models with covariates[C]// Proceedings of the 4th World Congress on Engineering Asset Management，September 28-30，Athens，Greece：13p.

[18] QIU J，WANG H，LIN D，et al. Nonparametric regression-based failure rate model for electric power equipment using lifecycle data[J]. IEEE Trans on Smart Grid，2015，6(2)：955-963.

邱　剑（1990—），男，博士研究生，主要研究方向：智能电网数据挖掘、故障率预测和状态检修。E-mail：jianqiu@zju.edu.cn

王慧芳（1974—），女，通信作者，博士，副教授，主要研究方向：电网状态检修、继电保护与控制。E-mail：huifangwang@zju.edu.cn

应高亮（1957—），男，高级工程师，主要研究方向：带电检测应用技术。

（编辑　王梦岩）

（下转第 118 页　continued on page 118）

(上接第 112 页　continued from page 112)

Text Mining Technique and Application of Lifecycle Condition Assessment for Circuit Breaker

QIU Jian[1], *WANG Huifang*[1], *YING Gaoliang*[2], *ZHANG Bo*[2], *ZOU Guoping*[3], *HE Benteng*[1]
(1. College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China;
2. State Grid Jinhua Power Supply Company, Jinhua 321017, China;
3. Electric Power Research Institute of State Grid Zhejiang Electric Power Company, Hangzhou 310014, China)

Abstract: In power grids, operating and maintaining engineers have recorded plenty of texts or logs during maintaining and inspecting activities. These textual data contain abundant asset health information. So far, however, few researches, if any, have studied text mining techniques in the power grid. We take the circuit breaker (CB) as a case in point to establish a framework of text mining-based lifecycle condition assessment. Firstly, the key issues of text mining and lifecycle condition assessment models are listed based on reviewing the research of CB condition assessment. Then, use is made of the framework including a hidden Markov model (HMM)-based text preprocessing and vectorizion, self-interval searching k-nearest neighbor (KNN)-based text classification, and a proportional health-index fusion model (PHFM). Finally, we have collected real textual data from a certain power company to demonstrate an example that shows the text mining technique could learn similar defects from other assets by itself, and PHFM shows historical data stream and lifecycle health index much more rigorously.

This work is supported by State Grid Corporation of China.

Key words: lifecycle condition assessment; maintenance and inspection; circuit breaker; text mining; hidden Markov model (HMM); k-nearest neighbor (KNN)