科学通报 2018 年 第 63 卷 第 8 期: 777 ~ 783
论 文
《中国科学》杂志社
SCIENCE CHINA PRESS

一种基于管道大数据的腐蚀因素相关性分析方法

张河苇[1], 金剑[2], 董绍华[1*], 李宁[2], 张来斌[1]

1. 中国石油大学(北京)机械与储运工程学院, 北京 102249;
2. 中石油管道有限责任公司西部分公司, 乌鲁木齐 830000
* 联系人, E-mail: dongshaohua72@sina.com

2017-09-27 收稿, 2018-01-29 修回, 2018-01-31 接受, 2018-03-19 网络版发表
国家重点基础研究发展计划(2017YFC0805800)和中石油管道有限责任公司西部分公司科研项目(XG-JCGL-CX-KJXX-01-JL-03/201604)资助

摘要 我国天然气管道建设在过去的10年取得了巨大的发展, 目前中国的长输管道总数量已达到12.5×10^4 km. 随着管道系统的运行时间增长和材料性能的老化, 由管道腐蚀、失效泄漏引起的事故频发, 并且由于管径的增大和运行压力的提高, 事故后果严重性也随之提高, 特别是对高后果区的影响更为严重. 油气管道的安全问题越来越受到重视. 因此, 提取管道腐蚀性影响因素、发现风险演变规律对于预测预防管道泄漏尤为重要. 由于数字化管道已经逐步建立, 设计、施工、运行过程中产生大量的数据得以保存, 内检测数据信号总量也达到TB级别, 这些均为腐蚀分析提供了信息基础. 本文基于大数据的方法, 融合管道系统在设计、施工、运行过程中产生的结构化和非结构化数据, 基于互信息理论方法, 建立腐蚀等级与多因素之间的关联模型. 通过现场数据验证模型的有效性, 确定影响管道腐蚀的关键因素, 为管道寿命预测奠定基础. 结论表明, 利用管道"大数据"挖掘管道各数据集之间的关联关系, 可用于获取管道腐蚀信息, 为事故预防提供依据. 除此之外, 该方法还可以在确定主要因素之后缩小管段范围, 研究其中某一种因素, 如进一步确定腐蚀与哪一种土壤的相关性更强.

关键词 腐蚀, 大数据, 互信息理论, 相关性分析, 管道

管道是运输石油、天然气和成品油最经济、安全的方式之一, 但因为外部环境及管材、施工等的影响, 导致腐蚀、断裂等问题, 进而存在火灾、爆炸的危险[1]. 我国管道地域广阔、环境复杂, 管道总长度近80×10^4 km(含长输管道12×10^4 km、油田集输管道33×10^4 km、城市燃气管道35×10^4 km), 由于管道材质一般为钢管, 在复杂的土壤环境中, 腐蚀问题始终是贯穿其安全生产的重要因素[2,3]. 在美国政府运输部管道及危险物品安全管理局对1992~2012年美国管道事故的统计分析中(http://primis.phmsa.dot.gov/comm/reports/safety/PSI.html?nocache=8699), 研究者发现引起集气管道事故的首要因素是腐蚀, 其次才是材料、焊接、装备失效[4]. 因此, 分析管道腐蚀的相关因素, 对掌握腐蚀规律及提出防护措施具有十分重要的意义.

目前国内外研究人员对管道腐蚀因素进行了大量的研究工作. 例如, 范开峰等人[5]从管道的内外壁腐蚀两个方面论述了导致天然气管道腐蚀的原因和机理; 牛雁坡[6]从化学原理的角度分析了可能会导致腐蚀速率增加的因素, 并提出了相应的防护措施. 上述方法所提出的因素均建立在理论分析层面, 未采用实际的现场数据进行验证. 骆正山和郜阳[7]采用灰色支持向量机的方法进行湿天然气集输管道腐蚀研究, 分析预测了腐蚀速率的相关因素; 骆正山等人[8]提出了基于PCA-SVM的预测模型. 但是, 腐蚀是内外环境、运行条件共同作用的结果, 而目前所使用的方法多从内腐蚀或外腐蚀中的某一方面入手,

引用格式: 张河苇, 金剑, 董绍华, 等. 一种基于管道大数据的腐蚀因素相关性分析方法. 科学通报, 2018, 63: 777–783
Zhang H W, Jin J, Dong S H, et al. A corrosion correlation analysis method based on pipeline big data (in Chinese). Chin Sci Bull, 2018, 63: 777–783, doi: 10.1360/N972017-01038

© 2018《中国科学》杂志社

分析结果存在一定偏差. 近年来, 数字化管道已在国内外管道企业建立起来, 集成了管道系统在设计、施工、运行过程中产生的大量数据[9]. 随着物联网、云计算、大数据等技术的发展, 数据的深度挖掘应用极为紧迫, 决策支持方法和模型亟待建立. 本文采用互信息法建立基于管道大数据的相关性分析模型对管道腐蚀因素进行全面分析, 融合管道系统在设计、施工、运行过程中产生的结构化和非结构化数据, 包括4个数据集, 从中提取36个腐蚀数据要素, 建立腐蚀等级与多因素之间的关联模型, 找出相关关系. 通过现场数据验证模型的有效性, 确定影响管道腐蚀的关键因素, 为管道寿命预测奠定基础, 为建立预测模型提供参考.

1 管道大数据

自2011年5月, 麦肯锡公司发布了关于大数据的调研报告*The next frontier of innovation*(《大数据: 下一个前沿, 竞争力、创新力和生产力》)[10]之后, 大数据技术在互联网、金融、物流领域发展迅速, 体现出极高的社会价值[11]. 例如, 淘宝、沃尔玛等能够从人们的购物习惯中发现购买规律并向客户进行商品推荐; 百度、谷歌等搜索引擎也会搜索网页背后的数据链并据此帮助用户最快找到有用信息. 大数据为发现事物背后的规律提供了可能.

中国目前长输管道总里程已经达到12.5×10^4 km, 每天都产生大量数据, 基本实现电子化记录和网络化管理, 这标志着管道大数据时代的来临[12]. 管道大数据包括管道建设期数据、管道地理数据、检测监测数据、管道资产设备数据、管道腐蚀数据、运营数据、市场数据. 然而在现阶段, 由于管道企业间各数据相对独立, 即使在同一个企业中, 不同类型的数据也依赖于不同的存储系统, 这就导致了数据资源的浪费, 隐含着管道安全信息的"大数据"绝大部分被忽视和丢弃. 如果能够以管道内检测数据为基础, 将历史资料、日常管理数据等进行对齐、统一整合就可以真正实现管道大数据的价值.

管道大数据的引入也会带来分析方式上的改变. 冯庆善[13]提出当具备了管道大数据的基础, 就可以从事故的内因出发, 分析导致某一事故发生的因素在总因素所占的比重进而判断该事故发生的概率. 这种分析方法是与传统的从表象出发的分析方式截然不同的. 同样地, 在小数据的模式下, 是用样本数据来进行预测分析, 这需要辅助主观筛选, 加入了人为因素的干扰, 而大数据则是从全体数据中逐步筛选判别其中的相关因素, 这个过程大多依靠计算机来完成.

其实, 在整个能源行业中, 大数据技术还处于起步阶段, 仅在电力行业中有一定的应用案例[14~17], 在管道系统中也仅限于在管道风险分析、内检测等方面有初步的探索[18~20], 尚未有实质性应用. 鉴于此, 本文从相关性分析的角度出发, 以内检测数据为核心, 将不同系统中的数据分类整理, 采用互信息法建立管道腐蚀与其他因素的相关分析模型, 发掘影响管道腐蚀的关键因素, 为腐蚀预测奠定基础.

2 互信息

信息理论是Shannon和Weaver[21]为定量描述通信过程中数据传输和储存开销而建立的一系列理论. 信息量的大小可以用所消除的不确定性的大小来表示[22].

从信源得到的平均信息量可由系统的熵来定义. 假定是一个离散随机变量X, 则变量X的信息熵$H(X)$为

$$H(X)=-\sum_{x\in X}p(x)\log_b p(x), \tag{1}$$

其中, X为信源, $p(x)$表示变量X取值为x的概率; b为底数, 当b取不同的值表示信息熵有不同的量纲, 当b=2时量纲为bit(比特), 当b=e(自然对数底)时量纲为nat(纳特).

假如随机变量X已知, 那么随机变量Y关于X的条件熵$H(Y|X)$定义为

$$\begin{aligned}H\left(Y\middle|X\right)&=\sum_{x\in X}p\left(x\right)H\left(Y\middle|X=x\right)\\&=-\sum_{x\in X}p\left(x\right)\sum_{y\in Y}p\left(y\middle|x\right)\log_b p\left(y\middle|x\right)\\&=-\sum_{x\in X}\sum_{y\in Y}p\left(x,y\right)\log_b p\left(y\middle|x\right).\end{aligned} \tag{2}$$

条件熵反映了随机变量Y在得到X提供的信息之后, 剩余的不确定性. 熵与条件熵存在以下的不等式关系:

$$H\left(X\middle|Y\right)\leqslant H\left(X\right). \tag{3}$$

当$H(X|Y)$=0, 表示Y包含了X的全部信息; 当$H(X|Y)$=$H(X)$, 表示两个变量是相互独立, 即Y不能给X提供信息.

互信息是信息理论中用来表示变量间相关程度的一个基本概念[22]. 两个随机变量X和Y的互信息

$I(X,Y)$定义为

$$\begin{aligned} I(X,Y) &= H(X)-H(X|Y) \\ &= -\sum_{i=1}^{n} p(x_i)\log_b p(x_i) \\ &\quad +\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_i)\log_b p(x_i/y_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\log_b \frac{p(x_i,y_j)}{p(x_i)p(y_j)}, \end{aligned} \tag{4}$$

其中, x_i代表变量x中的第i个值; y_j代表变量y中的第j个值; $p(x_i, y_j)$表示变量x为第i个值且变量y为第j个值的联合概率密度, $p(x_i)$和$p(y_j)$为独立密度. $I(X,Y)$越大, 变量X包含关于Y的信息就越多, 因此可用互信息来度量变量间的相关性[23].

3 相关性分析模型

管道腐蚀的产生与运行条件、外界环境密不可分. 根据地域和管道本体属性的不同, 管道发生腐蚀的可能性也不同. 本文首先对管道数据进行对齐整合, 然后采用互信息法对管道腐蚀与其他影响因素进行分析和排序, 确定导致管道腐蚀产生的关键因素, 为管道的安全管理提供依据.

具体建模步骤如下: 首先进行数据采集. 收集某一管段的所有信息, 然后将数据划分为管道基本属性数据、工程建设数据、内外检测数据、运行维护数据4个数据集. 收集范围包括管道建设期数据、管道地理数据、检测监测数据、管道资产设备数据、管道腐蚀数据、运营数据; 然后进行以内检测数据为基准, 将其他数据与内检测数据进行对齐; 第三步为属性域离散化. 将腐蚀等级作为决策属性, 数据集中的其他因素作为条件属性. 对决策属性和条件属性的属性域离散化处理, 也就是把属性域中的最大值与最小值之间的距离划分为距离相等的子区间, 再将每个属性的数值分配到对应区间中, 最终得到每个子区间中的数值个数. 根据以下公式求解互信息值:

$$I(X,y) = -\sum_{i=1}^{M_i}\frac{N_i}{N}\cdot\log_2\frac{N_i}{N} + \sum_{j=1}^{M_j}P(y_j)\left[\sum_{i=1}^{M_i}\frac{N_{ij}}{N}\cdot\log_2\frac{N_{ij}}{N}\right], \tag{5}$$

其中, M_i为决策属性X的属性值个数; N_i为决策属性X第i个属性的数值个数; N为条件属性Y和决策属性X所有数值个数之和; M_j为输入的条件属性$y(y\in Y)$的属性值个数; $P(y_j)$为输入的条件属性y的第j个区间的数值个数除以y的所有数值个数; N_{ij}为当输入的条件属性y在第j个区间时, 决策属性恰好在第i个区间的数值个数.

假设腐蚀等级为X, 其他数据信息构成条件属性集$Y=(y_1, y_2, \cdots, y_p)$, 则管道缺陷与其他条件属性之间的互信息可表示为

$$I(X,Y) = \left[I(X,y_1)\cdots I(X,y_j)\cdots I(X,y_p)\right], \tag{1}$$

对得到的互信息值进行排序, 互信息越大表示二者之间的相关性越强.

构造相关分析模型的流程图如图1所示.

4 案例分析

以某长输管道中长度为100 km的管段为例进行分析. 收集该管段的管道建设期数据、管道地理数据、检测监测数据、管道资产设备数据、管道腐蚀数据、运营数据等数据信息.

将采集的数据重新划分为管道基本属性数据、工程建设数据、内外检测数据、运行维护数据等4个数据集, 然后以管道桩号和参考点为基准将数据集中的数据进行对齐. 最终得到8772条数据信息, 每条数据信息中均包括缺陷桩号、钢管类型、腐蚀等级、焊缝类型等36个因素. 部分数据表格如表1所示. 将腐

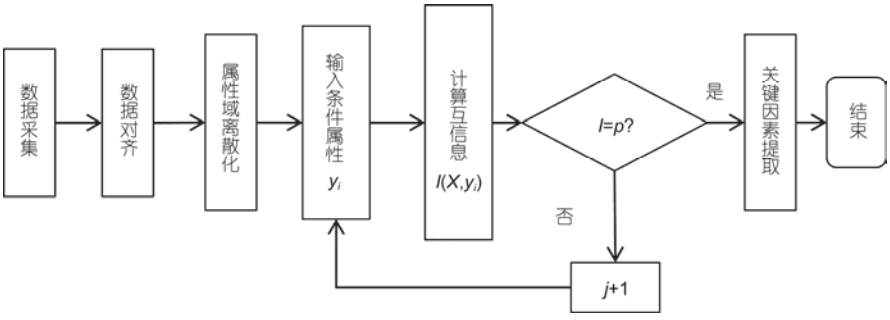


图 1 相关分析模型流程图

Figure 1 The flow chart of the correlation analysis model

表 1 数据总表节选

Table 1 Excerpt from the data matrix

桩号	缺陷长度(mm)	埋深(m)	钢管类型	管径(mm)	管道壁厚(mm)	焊缝类型	土壤类型	钢管材质	施工单位
1333	154	1.4	直管	660	7.14	螺旋焊缝	黄土	L415(X60)	中油一建
1334	67	1.4	直管	660	7.14	螺旋焊缝	黄土	L415(X60)	中油一建
1335	67	1.4	直管	660	7.14	螺旋焊缝	黄土	L415(X60)	中油一建
1336	48	1.2	直管	660	7.14	螺旋焊缝	黄土	L415(X60)	中油一建
1337	38	0.7	直管	660	7.14	螺旋焊缝	黄土	L415(X60)	中油一建
1338	24	0.9	直管	660	7.14	螺旋焊缝	黄土	L415(X60)	中油一建
1339	67	0.9	直管	660	7.14	螺旋焊缝	黄土	L415(X60)	长庆油建
1340	86	1	直管	660	8.74	螺旋焊缝	黄土	L415(X60)	长庆油建
1346	38	1.6	直管	660	10.3	螺旋焊缝	黄土	L415(X60)	长庆油建
1347	34	1.7	直管	660	10.3	螺旋焊缝	黄土	L415(X60)	长庆油建
1348	67	1.3	直管	660	10.3	螺旋焊缝	黄土	L415(X60)	长庆油建
1352	58	–	直管	660	–	螺旋焊缝	黄土	L415(X60)	长庆油建
1353	96	2.5	直管	660	7.14	螺旋焊缝	黄土	L415(X60)	长庆油建
1354	48	1.3	直管	660	7.14	螺旋焊缝	黄土	L415(X60)	长庆油建
1355	24	0.6	直管	660	7.14	螺旋焊缝	黄土	L415(X60)	长庆油建
1358	58	1.4	直管	660	7.14	螺旋焊缝	黄土	L415(X60)	长庆油建

蚀等级作为决策属性, 其余35个因素作为条件属性, 然后对36个属性值进行离散化.

根据式(4)求得每个条件属性与决策属性之间的互信息数值, 如图2(属性对应关系表见表2)所示, 将其中较其他数值有波动的点单独取出来如图3所示.

由图3中可以看出, 这10个因素依次为缺陷宽度、深度、长度、埋深、管道类型、管径、壁厚、焊接类型、土壤类型、管道材质. 其中数值最大的因素为管道材质、壁厚、管径, 较大的因素有宽度、深度、长度、埋深. 从数据集的角度来说, 本段管道腐蚀等级与内检测数据(深度、宽度、长度)以及管道基本属性数据(管道材质、壁厚、管径埋深)的相关性较强. 因此可以将这7类数据作为预测管道腐蚀等级的重要参考因素.

同时可以采用此方法进一步分析, 如数据中的土壤还包括黄土、砂石土、黏土和湖泊等, 进一步分析后得到图4的互信息值, 由图中可以看出黄土和黏土等地区管道腐蚀更为严重.

5 结论

从现场收集管道资料和历史数据, 将不同系统中的多源数据按照一定规则进行分类整理, 基于互信息理论建立腐蚀数据的相关性分析模型, 用于分析条件属性与决策属性之间的相关关系, 其结果可为管体腐蚀预测奠定数据基础. 研究得到以下结论.

(1) 以管道内检测数据为基准, 可以校正其他数据的位置对应性, 融合所有数据集形成管道大数据集, 通过建立多源性管道大数据分析表结构, 将腐蚀因子离散化, 使多源数据具有大数据分析的功能性特征.

(2) 运用互信息理论对管道缺陷的关联性因素进行判别, 挖掘出各种潜在关联因素和管道缺陷之

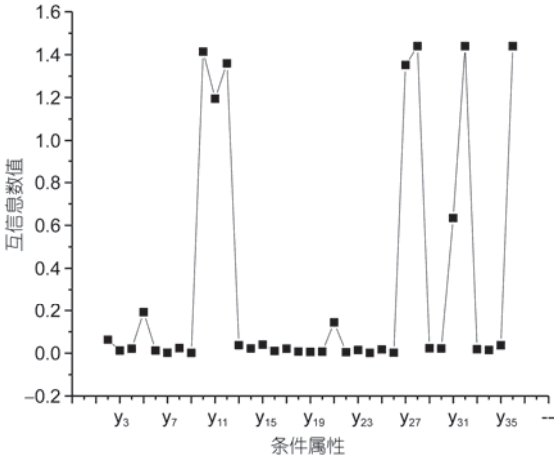


图 2 互信息分析结果

Figure 2 Mutual information analysis results

表 2　属性对应关系表

Table 2　Property correspondence table

编号	属性	编号	属性	编号	属性
y_1	腐蚀等级	y_{13}	无损检测单位	y_{25}	高后果区类型
y_2	防腐层类型	y_{14}	监理单位	y_{26}	风险评价结果
y_3	运行压力	y_{15}	施工单位	y_{27}	埋深
y_4	设计压力	y_{16}	高程值	y_{28}	管径
y_5	焊缝类型	y_{17}	内检测单位	y_{29}	行政区域
y_6	焊缝数量	y_{18}	检测监理单位	y_{30}	人口
y_7	降雨量	y_{19}	检测器类型	y_{31}	土壤类型
y_8	气候	y_{20}	内检测方法	y_{32}	管材
y_9	地貌	y_{21}	管道类型	y_{33}	穿跨越方式
y_{10}	深度	y_{22}	检测时间	y_{34}	环向位置
y_{11}	长度	y_{23}	地灾风险值	y_{35}	补口防腐类型
y_{12}	宽度	y_{24}	评价方法	y_{36}	壁厚

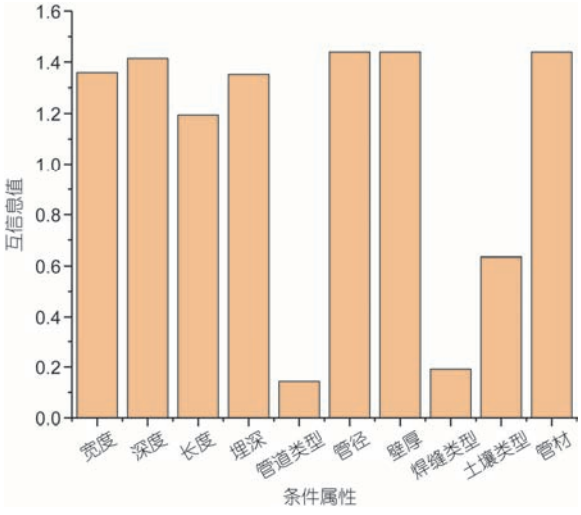


图 3　(网络版彩色)波动点

Figure 3　(Color online) Fluctuation points

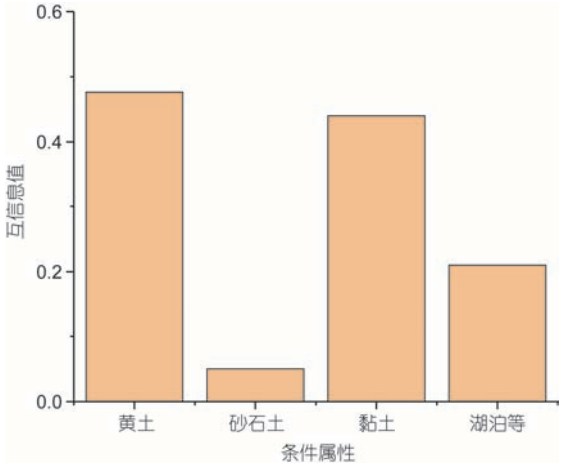


图 4　(网络版彩色)土壤类型分析结果

Figure 4　(Color online) Soil analysis results

间的关联关系, 这里的关联关系是一个相关关系, 而不是因果关系, 无需考虑数据的具体关系(线性关系或非线性关系), 该相关性分析方法不局限于结构化的数据, 也适用于非结构化数据的分析, 非结构化数据如焊接底片图像缺陷识别、第三方位置大数据等, 同样适用于管道大数据.

(3) 本案例中腐蚀相关因素的提取和划分可能较为笼统, 但分析方法针对数据提取的越多, 数据表越细致, 精度会更高. 海量数据的大数据分析遵循由面到线再到点的原则, 可以在确定主要因素之后缩小管段范围, 研究其中某一种因素, 如进一步确定腐蚀与哪一种土壤的相关性更强, 可以提高后期预测的精准性.

参考文献

1　Feng Q S, Chen J F, Ai M Y, et al. Application of pipeline integrity management in earthquake disaster response (in Chinese). J Pet, 2010, 31: 139–143 [冯庆善, 陈健峰, 艾慕阳, 等. 管道完整性管理在应对地震灾害中的应用. 石油学报, 2010, 31: 139–143]

2　Miao C W, Lu Q M. The planning of natural gas pipeline industry and corrosion protection technique in China. In: Proceedings of the

2nd China International Corrosion Control Conference. Beijing, 2002. 18–26
3 Liu J. Corrosion causes and prevention measures of natural gas pipelines (in Chinese). NeiJiang Tech, 2012, 33: 104–105 [刘佳. 天然气管道的腐蚀原因及防治措施. 内江科技, 2012, 33: 104–105]
4 Di Y, Shuai J, Wang X L, et al. Cause analysis and classification method of oil and gas pipeline accidents (in Chinese). Chin J Saf Sci, 2013, 23: 109–115 [狄彦, 帅健, 王晓霖, 等. 油气管道事故原因分析及分类方法研究. 中国安全科学学报, 2013, 23: 109–115]
5 Fan K F, Wang W Q, Sun R, et al. Corrosion and anticorrosion analysis of natural gas pipelines (in Chinese). Contemp Chem, 2013, 42: 653–656 [范开峰, 王卫强, 孙瑞, 等. 天然气管道腐蚀与防腐分析. 当代化工, 2013, 42: 653–656]
6 Niu Y P. Study on corrosion cause and corrosion rate prediction of oil transportation pipelines (in Chinese). Piping Tech Equip, 2011, 5: 45–47 [牛雁坡. 石油运输管道腐蚀原因和腐蚀速率预测研究. 管道技术与设备, 2011, 5: 45–47]
7 Luo Z S, Gao Y. Study on corrosion of wet natural gas gathering pipeline based on grey support vector machine (in Chinese). Chin J Saf Sci, 2013, 23: 62–67 [骆正山, 郜阳. 基于灰色支持向量机的湿天然气集输管道腐蚀研究. 中国安全科学学报, 2013, 23: 62–67]
8 Luo Z S, Bi A R, Wang X W. Corrosion prediction of high-sour oil-gas mixed pipelines based on PCA-SVM (in Chinese). Chin J Saf Sci, 2016, 26: 85–90 [骆正山, 毕傲睿, 王小完. 基于 PCA-SVM 的高含硫油气混输管道腐蚀预测. 中国安全科学学报, 2016, 26: 85–90]
9 Dong S H, Zhang H W. A full lifecycle intelligent pipe-network solution based on large data (in Chinese). Oil Gas Storage Transp, 2017, 36: 28–36 [董绍华, 张河苇. 基于大数据的全生命周期智能管网解决方案. 油气储运, 2017, 36: 28–36]
10 Brown J S, Hagel J. The next frontier of innovation. McKinsey Q, 2005, 3: 82–91
11 Li G J, Cheng X Q. Research status and scientific thinking of big data—Strategy Policy Decision Research (in Chinese). J Chin Acad Sci, 2012, 27: 647–656 [李国杰, 程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考. 中国科学院院刊, 2012, 27: 647–656]
12 Dong S H. Theory and Application of Pipeline Integrity Evaluation (in Chinese). Beijing: Petroleum Industry Press, 2014 [董绍华. 管道完整性评估理论与应用. 北京: 石油工业出版社, 2014]
13 Feng Q S. The theoretical model of the cause of "tree-born" accident (in Chinese). Oil Gas Storage Transp, 2014, 33: 115–120,128 [冯庆善. "树生"事故致因理论模型. 油气储运, 2014, 33: 115–120,128]
14 Fisch D, Gruber T, Sick B. SwiftRule: Mining comprehensible classification rules for time series analysis. IEEE Trans Knowl Data Eng, 2011, 23: 774–787
15 Gill S, Stephen B, Galloway S. Wind turbine condition assessment through power curve copula modeling. IEEE Trans Sustain Energy, 2012, 3: 94–101
16 Torres R S, Falcão A X, Gonçalves M A, et al. A genetic programming framework for content-based image retrieval. Pattern Recogn, 2009, 42: 283–292
17 Zhu Y L, Yin J L. Study on application of multi-kernel learning relevance vector machines in fault diagnosis of power transformers (in Chinese). Proc CSEE, 2013, 33: 68–74 [朱永利, 尹金良. 组合核相关向量机在电力变压器故障诊断中的应用研究. 中国电机工程学报, 2013, 33: 68–74]
18 Lin X X, Li Y X, Zhou X, et al. Data management of pipeline inspection in large data environment (in Chinese). Oil Gas Storage Transp, 2015, 34: 349–353 [林现喜, 李银喜, 周信, 等. 大数据环境下管道内检测数据管理. 油气储运, 2015, 34: 349–353]
19 Feng Q S. Consideration of pipeline risk assessment method based on large data condition (in Chinese). Oil Gas Storage Transp, 2014, 33: 457–461 [冯庆善. 基于大数据条件下的管道风险评估方法思考. 油气储运, 2014, 33: 457–461]
20 Wang W B. Large data management structure and application of long-distance oil and gas pipeline (in Chinese). Oil Gas Storage Transp, 2015, 34: 229–232 [王维斌. 长输油气管道大数据管理架构及应用. 油气储运, 2015, 34: 229–232]
21 Shannon C E, Weaver W. The Mathematical Theory of Communication. Urbana: University of Illinois press, 2015
22 Cao X H, Zhang Z C. Information Theory and Coding (in Chinese). Beijing: Tsinghua University Press, 2004 [曹雪虹, 张宗橙. 信息论与编码. 北京: 清华大学出版社, 2004]
23 Han M, Liu X X. Multi-step input variable selection multiple sequence prediction based on mutual information (in Chinese). J Autom, 2012, 38: 999–1006 [韩敏, 刘晓欣. 基于互信息的分步式输入变量选择多元序列预测研究. 自动化学报, 2012, 38: 999–1006]

Summary for “一种基于管道大数据的腐蚀因素相关性分析方法”

A corrosion correlation analysis method based on pipeline big data

Hewei Zhang[1], Jian Jin[2], Shaohua Dong[1*], Ning Li[2] & Laibin Zhang[1]

[1] *College of Mechanical and Transportation Engineering, China University of Petroleum (Beijing), Beijing 102249, China;*
[2] *China Petroleum Pipeline Co., Ltd. West Branch, Urumqi 830000, China*
* Corresponding author, E-mail: dongshaohua72@sina.com

At present, the total number of long-distance pipeline in China has reached 125000 km, achieved great development in the past decade. With the increase of the running time of the pipeline system and the aging of material properties, the frequent accidents are caused by pipeline corrosion and failure leakage. What’s more, the severity of the accident consequences has been increased since the pipe diameter and operation pressure has increased, especially for the impact on the high consequence area. Thus, the safety problems of oil and gas pipelines are paid more and more attention. Extracting the corrosive factors of pipeline and the law of risk evolution is important to predict the prevention of pipeline leakage. With the development of digital pipeline, the data produced during the process of design, construction and operation with in pipeline system were saved and the amount of detection data signal has also reached the level of TB. All of this has provided the fundamental for information extraction. However, because of lack of relationship between these data sets, most part of the data in which the safety information of pipeline hidden was ignored and discarded. In order to fulfill the potential of the “big data” and analyze the causes of pipeline corrosion in term of data, structural and unstructured data produced in the design, construction and operation of the pipeline system are constructed in this paper, and based on the mutual Information theory method, the correlation model between the corrosion grade and the multi-factor is established, which aim is to extract all the relevant factors of pipeline corrosion defects from the “big data” and determine the key indexes of the defects. This method is not limited to structured data, and it is also applicable to the analysis of unstructured data. The basic modeling steps include: collect all the information of a pipe segment first, then divide the data into four datasets; the third step is to attribute domain discretization; finally, a correlation model based on mutual information theory is established to determine the relationship between corrosion and other factors and to rank all factors according to the size of influence values. It is noteworthy that the relationship here is a correlation, not a causal relationship, that is, no need to consider the linear or non-linear relationship of the data. By validating the validity of the model, the key factors affecting the pipeline corrosion are determined, which lays a foundation for the prediction of pipeline life. The conclusion shows that the correlation between the data sets of pipeline “big data” can be used to obtain the information of pipeline corrosion and to provide the basis for accident prevention. In addition, the method can be used to study one of the factors, such as further determine which soil is more relevant with the corrosion, after the main factors are extracted to narrow the scope.

corrosion, big data, mutual information theory, correlation analysis, pipeline

doi: 10.1360/N972017-01038