# 油田环保安全领域标准数据关联性监测技术研究

王凯月[1] 黄 珊[1] 王逸飞[3] 孙红军[2] 苏雪松[3] 延 伟[1]

（1.中国石油化工股份有限公司胜利油田分公司技术检测中心；2.中国标准化研究院；
3.胜利油田检测评价研究有限公司）

**摘 要：**本论文开展油田环保安全标准关联性监测技术研究，针对油田环保安全标准相关国内外动态信息（如：标准动态、政策法规、智库报告、情报产品、热点栏目）进行油田环保安全标准领域自动化关联性监测，遵循“油田环保安全领域标准数据需求识别、油田环保安全领域标准数据源的确定依据、油田环保安全领域标准关联数据自动抓取、油田环保安全领域标准关联监测内容分析”的研究思路，利用大数据分析与知识关联技术，实现对所需监测数据基本内容的自动化统计与分析，动态可视化地展示或分析所需监测数据的内容，及时跟踪与推送油田环保安全标准前沿与热点内容，支持用户便捷了解油田环保安全标准领域最新发展动态，为开展油田环保安全领域标准知识库建设提供多元数据支撑。

**关键词：**油田环保安全，标准数据，关联性监测，机器学习

**DOI编码：**10.3969/j.issn.1674–5698.2024.02.008

# Research on Correlation Monitoring of Standard Data in Oilfield Environmental Safety Field

WANG Kai-yue[1] HUANG Shan[1] WANG Yi-fei[3]
SUN Hong-jun[2] SU Xue-song[3] YAN Wei[1]

（1. Technology Testing Center of Shengli Oilfield Branch, China Petrochemical Co., Ltd.;
2. China National Institute of Standardization; 3. Shengli Oilfield Testing and Evaluation Research Co., Ltd.）

**Abstract:** This paper carries out research on the correlation monitoring technology of oilfield environmental safety standards, and conducts automatic correlation monitoring of domestic and foreign dynamic information related to oilfield environmental safety standards such as standard dynamics, policies and regulations, think tank reports, intelligence products, and hot columns. The research ideas are to identify data requirements of oilfield environmental protection safety standards, determinate standard data sources, automatically capture associated data and make an analysis of associated monitoring content. The paper uses big data analysis and knowledge correlation technology to realize automatic statistics and analysis of the basic content of the required monitoring data, dynamically and visually display or analyze the monitoring data, timely track and publish the forefront and hot content of oilfield environmental safety standards, help users to easily understand the latest development in the field of oilfield environmental safety standards, and provide multivariate data support for the construction of the oilfield environmental safety standard knowledge base.

**Keywords**: oilfield environmental protection safety, standard data, correlation monitoring, machine learning

**作者简介：**王凯月，工程师，学士，主要研究方向为企业标准化建设、企业标准数字化转型。

## 0 引言

以新一代信息技术为代表的新一轮科技革命和产业变革加速演进，经济社会数字化转型成为时代趋势。标准作为经济活动和社会发展的技术支撑，以及国家基础性制度的重要方面，无论在深度还是在广度上都将受到这一趋势的影响。标准数字化转型已成为新时代我国标准化事业发展的重要战略方向，对增强我国科技发展的标准化互动支撑能力、影响全球标准化生态变革具有重要意义。随着人工智能、开源、区块链等技术的持续发展，标准化领域受其影响，出现了多种标准数字化相关概念、标准形式与制定方式。2019年国际标准化组织（ISO）和国际电工委员会（IEC）提出SMART（Standard Machine Applicable，Readable and Transferable）标准的概念，将标准数字化能力划分为5个阶段，该模型在国际层面已经形成共识。2021年10月，中共中央、国务院发布《国家标准化发展纲要》，要求“推动标准工作向数字化、网络化、智能化转型”。随着我国社会不断发展，油田行业也逐渐发展起来。同时，油田行业的经济基础也与日俱增，油田行业的环保安全意识日益增强，而油田环保安全领域标准对于规范和引导油田行业安全生产、绿色发展和效率提升具有重要作用。在数字化时代，油田行业对于安全环保标准智能化应用和服务要求更高。当前，油田环保安全领域标准面临尚未形成标准动态数据源分析与监测方法，具体问题如下。

油田环保安全领域系统化和一体化的标准动态数据源尚未建立。在高质量发展新时代，标准数字化既是经济社会发展、数字技术变革，也是实现国家质量基础设施数字化转型的关键内容。随着标准数字化的发展，标准的普及与使用更加广泛，在标准数字化发展过程中，油田企业高质量发展对标准动态数据质量提出了更高要求，尚未形成高相关、系统化、一体化的标准动态数据源分析与监测方法，无法及时跟踪全球有关油田环保安全领域标准数据、政策法规、战略规划、科技成果等，不能实时跟踪与推送前沿与热点内容，自然无法满足支持用户便捷了解科技标准的最新发展动态。同时，油田环保安全领域各标准管理平台在标准数字内容来源、加工、更新、管理和存储格式等方面都有所区别，形成了“各自为政的局面”，亟待建立统一的标准数据“源”。

因此，为有效支撑国家和国家标准数字化战略有效实施，本论文开展油田环保安全标准关联性监测技术研究，针对油田环保安全标准相关国内外动态信息（如：标准动态、政策法规、智库报告、情报产品、热点栏目）进行油田环保安全标准领域自动化关联性监测，及时跟踪与推送前沿与热点内容，支持用户便捷了解油田环保安全标准领域最新发展动态。

## 1 研究综述

标准关联性监测（Association Monitoring）是指围绕某一领域标准通过对多个相关事件或数据点之间的关联进行实时监测和分析，以发现新的关联模式、趋势或异常情况的过程。这种监测技术在不同领域中都有广泛的应用，如：市场分析、金融风控、社交媒体挖掘等。

数据源关联监测相关技术的发展可以追溯到互联网的兴起和数据爆炸的时代。数据源关联监测是指对数据源进行实时或定期的监测和分析，以识别数据的变化和趋势。在互联网和大数据时代，数据源的规模和多样性迅速增加，对数据源关联监测的需求也日益增加。随着技术的不断发展，相关技术在过去几十年发生了巨大变化。

在数据源关联监测的发展历史中，最早应用的技术之一是网络爬虫[1,2]。网络爬虫技术最早出现在20世纪90年代末，用于搜索引擎的数据收集和索引。当时的搜索引擎如：Altavista和Excite都使用了网络爬虫技术来抓取互联网上的网页。随着互联网规模的迅速扩大，网络爬虫技术也得到了进一步的发展和改进。现在，网络爬虫已广泛应用于各种领域，如：舆情监测、新闻采集和金融数据收集等。另一个重要的技术是文本挖掘[3]，在20世纪90年代末至2000年初开始得到关注和发展。

当时，研究者开始使用自然语言处理和机器学习技术，对大规模文本数据进行分析和挖掘。这为数据源关联监测中的文本分析提供了基础。通过文本挖掘技术[4]，可以从数据源中提取关键词、主题和情感等信息，以便判断数据源的变化和趋势。随着计算能力和数据量的增加，机器学习技术也开始应用于数据源关联监测。机器学习[5,6]是一种通过算法让计算机从数据中学习和提取模式的技术。在数据源关联监测中，机器学习可以用于构建模型并预测数据的变化和趋势。研究者可以使用机器学习算法如：支持向量机、决策树和神经网络等来构建模型，以自动识别异常行为和趋势，并提供预测结果。统计分析[7]也是数据源关联监测中的重要技术之一，是一种用于分析和解释数据的技术。在数据源关联监测中，统计分析技术常用于比较和分析不同时期的数据。通过统计分析，可以检测数据的趋势、方差和相关性等统计指标，帮助用户理解数据源的动态变化。

此外，随着人工智能和大数据技术的快速发展，数据源关联监测也融合了一些新兴的技术。例如：自然语言处理和语义分析技术可以进一步提高文本数据的理解和处理能力。深度学习技术[7]的应用可以帮助处理复杂的模式和结构。同时，云计算和分布式处理技术可以加速数据源关联监测的速度和效率。区块链技术的引入可以保证数据的安全性和可信度。

数据源关联监测相关技术在过去几十年中取得了长足的发展。网络爬虫、文本挖掘、机器学习、统计分析和数据可视化等技术的进步不仅提高了数据源关联监测的效率和准确性，还为决策者和研究人员提供了更好的数据分析和洞察力。随着新兴技术的不断涌现，比如：自然语言处理、深度学习和区块链等，数据源关联监测将进一步发展和创新。这些技术应用于油田环保安全领域标准数据源关联监测，也将促进油田环保安全领域标准数字化的发展。

目前已经开始对标准关联性监测进行探索，中国标准化研究院通过监测国内外相关网站实现实时追踪抓取国内外相关标准化信息情报，在此基础上形成标准舆情化产品。国家科技图书文献中心（NSTL）建成了科技标准重点领域信息门户，该门户聚焦标准化与科技创新互动、资源环境标准化、质量研究、农业食品标准化、高新技术标准化等领域，跟踪全球有关科技标准的政策法规、战略规划、科技成果等，实时跟踪与推送前沿与热点内容，支持用户便捷了解科技标准的最新发展动态。目前门户已经监测了国内外183个相关标准化机构。但是上述尝试均是基于全领域标准开展相关关联性监测研究，鉴于此，本论文也将开展油田环保安全标准关联性监测技术研究。

## 2 油田环保安全领域标准关联性监测技术的主要内容

### 2.1 油田环保安全领域标准数据需求识别

针对公司对油田环保安全领域相关业务标准数据需求模糊、不明确等问题，开展大规模跨部门的实地调研与专家研讨，明确不同部门对标准数据及其来源需求的关键要点，绘制不同部门标准数据需求清单，并对业务相关标准数据需求数据进行聚类组织和处理，并反馈给各个业务部门，通过不断迭代优化，最终精准识别不同业务标准数据需求。上述工作方案的关键在于如何开展大规模的实地调研与专家研讨，本论文的具体方案如下所示。

实地调研确定需求的方案流程如下。

（1）确定调研目的。明确标准数据源范围调研的目的，为了了解不同业务部门对不同标准数据源需求情况。

（2）制定调研计划。设计一个调研问卷或面谈指南，包括一些开放性问题和封闭性问题，以便业务部门可以详细描述他们对标准数据的需求。

（3）选择合适的受访人员。选择每个部门中的关键人员，包括管理层、业务分析师和其他涉及数据使用的员工。

（4）进行调研。采访被选择的受访人员，确保问卷或面谈过程中能够深入探讨他们的需求和期望。

（5）整理和分析数据。将调研数据整理和分

析，找出各部门的共同需求和特定需求。这可能需要使用一些统计方法和数据分析工具。

专家研讨确定需求的方案流程如下。

（1）召集专家组。邀请各个部门的专家，包括业务领域的专家和数据分析专家，参与研讨会议。

（2）明确定位议程。制定会议议程，确保在会议中全面涵盖各个部门的需求，并确定确切的问题，以便专家们能够提供有针对性的意见。

（3）组织研讨会议。进行研讨会议，鼓励专家分享他们的见解、经验和建议。

（4）记录和整理意见。记录专家的意见和建议，包括可能的解决方案和实施策略。

（5）综合分析。将实地调研和专家研讨的结果综合起来，寻找共同点，确定优先级，制定数据需求的详细计划。

（6）反馈和确认。将综合分析的结果反馈给相关部门，确认他们的需求是否被准确理解，如果有误会或遗漏，及时进行修正。

（7）制定实施计划。基于综合分析的结果，制定数据需求的实施计划，包括数据收集、处理、分析和报告的具体步骤和时间表。

### 2.2 油田环保安全领域标准数据源的确定依据

针对油田环保安全领域业务标准数据源范围确定规则或依据缺乏的问题，研究面向不同业务需求的标准数据源范围确定的规则和框架要点，提出集“战略目标、问题导向、业务流程、前沿热点、重点任务、权威可信”等多维度为一体的标准数据源确定依据，并制定参照指标，采用多维评价指标体系等方法综合确定标准数据源。

通过建立符合标准源规则或框架要点的标准数据源筛选依据，并采用多维评价指标体系等方法对标准数据源重要程度进行打分，最终建立具有重要度评价的标准数据源头体系。当前标准源的评价研究大多采用单一或几个指标数据来进行测算，由于标准数据源是一个多元复杂系统，所以采用单一或几个测量指标无法准确表征标准数据源应有内涵。鉴于此，后续本文将采用多维指标体系方法来评估标准源重要水平。在多维指标体系下，其中一个重要问题就是对指标设置权重，根据设置权重方法不同，可将标准源常用测度方法划分为主观权重法、客观权重法、综合计量法。主观权重法包括综合加权法和层次分析法，客观权重法包括主成分分析法和熵值法，综合计量法包括随机前沿分析法（SFA）和数据包络分析法（DEA）。综合计量法更适用于包含投入和产出要素的绩效评估方法，即评估对象如何以较少的资源投入获得较多产出结果的多属性评估，这种方法要求指标体系中指标之间存在明显或严格的投入—产出关系。同时，由于熵值法是根据各项指标数值的变异程度来确定指标权数的，避免了人为因素带来的偏差，但该方法忽略了指标本身重要程度，有时确定的指标权数会与预期的结果相差甚远，同时熵值法不能减少评价指标的维数。鉴于此，本文后续将采用主观和客观相结合的方法从不同维度对标准数据源重要程度进行打分。

### 2.3 油田环保安全领域标准关联数据自动抓取

针对油田环保安全领域标准关联信息自动化抓取水平较低的问题，聚焦上述确定的油田环保安全领域标准数据的国内外相关数据源，采用大规模关联数据自动化抓取技术，自动搜集、挖掘和揭示相关领域或机构发布的标准相关新闻、政策、法规、报告、项目、成果等标准情报资源。其中，大数据关联数据自动化抓取是通过各种技术手段自动从不同数据源中提取数据并将其整合到一个数据存储中，以便进一步分析和处理。本论文制定的自动化抽取技术方案如下。

（1）Web 抓取和爬虫技术

爬虫框架：使用像Scrapy（Python）、Apache Nutch（Java）或者其他开源爬虫框架，能够自动化地从网页上抓取数据。

数据解析：使用HTML解析库（比如：BeautifulSoup、Jsoup）或正则表达式从网页中提取所需数据。

（2）API 调用

API 抓取：很多网站和在线服务提供API接口，可以通过API调用直接获取数据。使用工具如：Requests（Python）来进行API调用。

认证和授权：如果API需要认证，应确保拥有

正确的API密钥或令牌，并且了解API的限制和配额。

（3）数据库连接和查询

数据库连接：使用数据库连接库（例如：JDBC、ODBC）连接到数据库系统。

SQL查询：编写SQL查询语句来选择和提取所需的数据。对于非关系型数据库，可以使用相应的查询语言（例如：MongoDB的查询语言）。

（4）日志文件监控

日志分析：对服务器日志文件进行实时监控，并分析其中的数据。使用工具如：Apache Flume可以用来收集、聚合和移动大量的日志数据。

（5）消息队列

消息队列：使用消息队列系统（例如：Apache Kafka、RabbitMQ）来收集和传输数据。生产者将数据放入消息队列，消费者从中获取数据。

（6）数据仓库抽取（ETL）

ETL工具：使用ETL工具（例如：Apache NiFi、Talend、Apache Airflow）来提取、转换和加载数据。这些工具通常提供可视化界面，方便配置数据流程。

（7）实时数据流处理

流处理框架：使用实时数据流处理框架（例如：Apache Storm、Apache Flink、Apache Kafka Streams）来处理数据流，可以在数据抵达时进行实时处理。

（8）机器学习和自然语言处理

NLP 技术：如果需要从文本中抽取信息，可以使用自然语言处理（NLP）技术。工具如：NLTK（Python自然语言处理库）可以帮助处理文本数据。

机器学习模型：利用机器学习模型（例如：文本分类、命名实体识别）来自动从非结构化数据中抽取结构化信息。

（9）数据爬虫和机器学习结合

自动化学习模型：利用自动化学习模型（例如：AutoML工具）来构建能够适应不同网站结构的数据爬虫，从而实现智能化的数据抓取。

在选择合适的技术时，需要考虑数据源的类型、数据量、抓取频率、数据的格式等因素。综合运用这些技术，可以实现高效、稳定和自动化的大数据关联数据抓取过程。

### 2.4 油田环保安全领域标准关联监测内容分析

针对公司不同业务对标准数据分析和应用能力较差的问题，利用大数据分析与知识关联技术，分别从区域、时间、发布机构、关键词、摘要、单位合作网络、被引用频次等方面开展相关自动化识别，实现对所需监测数据基本内容的自动化统计与分析，实时、动态、可视化地展示或分析所需监测数据的内容。

自动化采集的标准信息可以通过各种机器学习和数据分析技术进行深入分析。本论文将综合采用以下大数据分析技术尝试进行监测数据分析。

（1）文本挖掘（Text Mining）

自然语言处理（NLP）：使用NLP技术，如：分词、命名实体识别、情感分析等，对文本进行处理和理解。

主题建模：使用主题建模算法（如：Latent Dirichlet Allocation）发现文本数据中的主题和关键词。

文本分类：使用文本分类算法（如：朴素贝叶斯、支持向量机）对文本进行分类，例如：垃圾邮件过滤、新闻分类等。

（2）数据挖掘（Data Mining）：

聚类分析：使用聚类算法（如：K均值聚类、层次聚类）将数据分成不同的簇，揭示数据的内在结构。

关联规则挖掘：使用关联规则挖掘算法（如：Apriori算法）找出数据中的关联规律，例如：购物篮分析。

异常检测：使用异常检测算法（如：孤立森林、LOF算法）找出数据中的异常点，用于欺诈检测、设备健康监测等。

（3）可视化分析

使用可视化工具（如：Matplotlib、Seaborn、D3.js）将分析结果以图表、图形的形式呈现，帮助用户直观理解数据。

利用地理信息系统（GIS）技术，将数据可视化在地图上，用于地理空间分析。

（4）实时分析

使用流处理技术（如：Apache Kafka、Apache Storm）进行实时数据分析，对持续产生的数据进行快速处理和响应。

在进行机器分析之前，需要进行数据预处理，包括数据清洗、缺失值处理、特征工程等步骤，以保证分析结果的准确性。选择合适的算法和工具，结合领域知识，可以更好地发现数据中的模式和规律。

## 3　研究结论

本文聚焦油田环保安全领域标准关联性监测技术研究，针对我国油田环保安全领域相关业务标准数据需求模糊、结构化处理需求差异较大、标准数据分析和知识关联能力较弱等问题，采用专家研讨和总结归纳等方法，绘制不同部门标准数据需求清单，精准识别不同业务标准数据需求。研究面向不同业务需求的标准数据源范围确定的框架要点，提出集“战略目标、问题导向、业务流程、前沿热点、重点任务、权威可信”等多维度为一体的标准数据源筛选依据，采用多维评价指标体系等方法综合确定标准数据源。采用大规模关联数据自动化抓取技术，自动搜集、挖掘和揭示相关领域或机构发布的标准相关新闻、政策、法规、报告、项目、成果等标准情报资源。利用大数据分析与知识关联技术，实现对所需监测数据基本内容的自动化统计与分析，实时、动态、可视化地展示或分析所需监测数据的内容。研发标准数据源监测状态提示与预警技术，对标准源状态（如：采集中、待审核、暂停、异常、新增数据量、采集时间等）进行提示或预警，及时优化和调整相应标准源。利用标准重点相关内容的关联技术，将获取的标准信息关键词或重点内容与关联知识进行大数据匹配，从而将与标准信息相关的创新成果、技术、评价等类型信息或数据纳入监测数据源中，实现多类型、高关联的标准数据监测，为开展标准知识库建设提供多元数据支撑。

**参考文献**


[1] S. Brin,L. Page. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN Systems, 1998,30(1-7): 107-117,.

[2] M. Najork, J. L. Wiener. Breadth-first crawling yields high-quality pages[C]. In Proceedings of the 10th International Conference on World Wide Web, 2001.

[3] C. Aggarwal. Data Mining: The Textbook[M]. Springer, 2015.

[4] Y. Huang, S. Zhang, J. Chen. A novel web data monitoring approach based on deep learning[C]. In Proceedings of the 19th International Conference on Big Data Analytics and Knowledge Discovery, 2017.

[5] Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction[M]. Springer, 2009.

[6] 唐亮,段建国,许洪波,等. 基于信息论的文本分类模型[J]. 计算机工程与设计, 2008,29(24):6312-6315.

[7] 尹江,尹治本,黄洪. 网络爬虫效率瓶颈的分析与解决方案[J]. 计算机应用, 2008(05):1114-1116+1119.