

doi: 10.7690/bgzdh.2022.03.017

基于文本挖掘技术和信息熵-TOPSIS 法辨识重点工序

唐彦秋¹, 何中其¹, 郑晓冉²

(1. 南京理工大学化学与化工学院, 南京 210094; 2. 中国烟草总公司山西省公司安全管理处, 太原 030000)

摘要: 为辨识影响火药安全生产的重点工序, 利用文本挖掘技术提取相关文献中具有分析价值的信息, 归纳出设备状态、事故损失、事故影响、参数监控、物料特性、事故可能性以及防护设计等 7 个指标, 对某发射药连续化制球工艺中重点工序进行评估。引入信息熵法优化了 TOPSIS 法中指标权重的计算, 综合求得各工序的重要度排序, 由此辨识出重点工序。实例分析结果表明: 挤出、切粒和塑化工序是该制球工艺中最重要工序, 对发射药的生产安全有重大影响; 所得结果与实际相符, 能为相关风险评估工作的有效进行提供思路。

关键词: 文本挖掘技术; 重点工序; 信息熵; TOPSIS 法; 风险评估

中图分类号: TJ55; TQ560.8 **文献标志码:** A

Identification of Key Processes Based on Text Mining and Entropy-TOPSIS Method

Tang Yanqiu¹, He Zhongqi¹, Zheng Xiaoran²

(1. School of Chemistry and Chemical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China;
2. Safety Management Office, Shanxi Branch of China National Tobacco Corporation, Taiyuan 030000, China)

Abstract: In order to identify the key processes affecting the safety of propellant production, the text mining technology was used to extract the valuable information from the relevant literature, and 7 indicators were summarized, including equipment status, accident loss, accident impact, parameter monitoring, material characteristics, accident possibility and protection design, and the key processes of a propellant continuous pelletizing process were evaluated. The information entropy method is introduced to optimize the calculation of index weight in TOPSIS method, and the importance ranking of each process is comprehensively obtained, thus the key process is identified. The results show that the extrusion, pelletizing and plasticizing processes are the most important processes in the pelletizing process, which have a significant impact on the production safety of propellant. The results are consistent with the actual situation, which can provide ideas for the effective risk assessment.

Keywords: text mining technology; key processes; information entropy; TOPSIS method; risk assessment

0 引言

国内外对火炸药生产相关的风险评估已展开广泛研究, 常用安全检查表法、事故树法、HAZOP 法、BZA-1 法等风险评估方法, 随着火炸药的高能化发展, 风险评估方法也在不断改进。曹楷等^[1]基于致因理论提出了改进 HAZOP-LEC 法, 用于某乳化炸药生产线 4 个工艺的风险分析, 辨识出工艺偏差及危险等级, 并提出相应的控制措施。冯睿智等^[2]建立了某单基发射药厂的生产过程安全评价指标体系, 运用层次分析法求得指标权重并提出了多层次模糊综合评价模型, 得出该生产系统的安全等级为中等并提出建议措施。钱辉等^[3]构建模糊事故树, 分析出导致固体推进剂生产过程中捏合工序发生燃爆事故的基本事件。刘保顺等^[4]运用模糊数学改进了传统的 BZA-1 法, 研究了单基药生产的塑化

与成型车间的现实危险度。通过风险评估方法的改进或多种方法联用以提高准确性, 或基于方法本身的原理引入模糊数学相关理论。

风险评估工作中除了方法的应用和改进相当重要外, 重点工序的辨识也不容忽视。重点工序即对火药安全生产影响最大的工序。余成早等^[5]剖析了单质炸药生产线的典型事故案例可知, 梯恩梯的硝化、太安的蒸馏等重点工序一旦发生事故, 常造成惨重的人员伤亡和财产损失。

在风险评估中常将重点工序作为主要研究对象。刘保顺等^[4]根据事故统计得知塑化成型工序易发生燃爆事故, 故从单基发射药生产系统的 3 个子系统中选择塑化成型工序进行定量评估研究。文献[6]在研究影响双基药生产工艺实现自动化的主要因素时, 从原料及中间产物的危险性出发, 因硝化

收稿日期: 2021-11-27; 修回日期: 2021-12-28

作者简介: 唐彦秋(1995—), 女, 贵州人, 硕士, 从事火炸药生产风险评估研究。E-mail: tyqdeyouxiang@163.com。

甘油的危险性最大，故将 NG 自动加料及输送工序作为其研究对象。Zhong T T 等^[7]在研究推进剂压缩成型工艺时凭个人经验认为压延是最危险工序，而后对压延工序开展数值模拟研究。柴涛等^[8]对硝化甘油生产的 5 个工序做事故统计分析，将爆炸次数最多的硝化工序确定为研究对象。

塑化成型、压缩成型、压延等都是上述学者的重点研究工序，但他们对重点工序的辨识却缺乏令人信服的依据，依靠历史事故资料或凭借个人经验等作出判断的方式虽较为简便，但事故资料不齐全、仅凭经验所得的结果主观性大等都是不容忽略的弊端，甚至可能因重点工序的判断失误致使后续评估工作无效。鉴于此，利用文本挖掘技术获取重点工序的判断指标，将信息熵和 TOPSIS 法(逼近理想解的排序法)联用以求出影响火药安全生产工序的重要性排序，由此得出重要性最大的即为重点工序，为火药的安全生产风险评估的研究提供思路。

1 方法

1.1 文本挖掘技术

文本挖掘技术又称文本数据挖掘或文本知识发现，Feldman 于 1995 年首次将数据挖掘与文本分类结合，并提出文本挖掘的概念^[9]，即利用分析工具从大量模糊的、随机的、有噪声影响的文本资料库中，提取出具有预测性或潜在分析价值的信息的过程^[10]。处理结构不规则的、没有预定义数据模型的非结构化或半结构化数据(如文档、图片、视频等)，一般处理过程为：文本预处理(去停用词、分词)、建模(算法)、结果分析及可视化。文本挖掘技术是一门综合性很强的学科，结合了语言学、统计学、计算机科学、数据库技术等学科^[11-13]。文本挖掘技术在人文社科研究中应用广泛，如意见挖掘、情感分析、用户需求及新闻编辑等领域。国内外对该技术的研究趋向不同，国外偏向于热点发现和医疗健康领域的运用，国内趋于知识管理、专利分析、文献计量学等领域的运用^[14]。目前，文本挖掘技术也逐渐应用于安全科学领域，如隧道施工的评估、铁路信号故障分析、安全生产事故数据分析等^[15-17]，可为安全生产提供可靠的决策辅助。

非结构化文本分析软件是能将非结构化文本中词频、词性、词间关系等特征以结构化数据或可视化方式呈现给用户的计算机软件。KH Coder 是一种较为常用的开源的非结构化文本挖掘软件，无需编程基础，且以 ChaSen、MySQL 作为后端，具备词

频统计、词性分析、自动聚类 and 可视化等功能^[18]。

1.2 信息熵-TOPSIS 法

1949 年，Shannon 提出信息熵作为描述信息性质及度量的方法，也可描述系统的不确定性^[19]。信息熵法又称熵权法，是一种客观赋权评价方法，主要根据各指标变异程度，利用信息熵计算出各指标的熵权，通过修正可得到较为客观的指标权重^[20]。信息熵越大说明指标变异程度越低，提供的信息量越少，故在综合评价中该指标的权重就越小^[21]。逼近理想解的排序法即 TOPSIS 法，是一种常用的决策方法，在模一化矩阵中通过计算评价对象与正理想解、负理想解的距离进而求得评价对象的优劣排序^[22]。信息熵-TOPSIS 法将信息熵客观求取的指标权重引入 TOPSIS 法中，确保计算结果的相对客观。

1) 构建决策矩阵。设多决策问题中有 m 个方案 A_1, A_2, \dots, A_m ， n 种属性 X_1, X_2, \dots, X_n ， d_{ij} 为方案 A_i 对属性 X_j 的取值，一般由专家打分得到，故构建初始决策矩阵 D 为：

$$D = (d_{ij})_{m \times n} = \begin{bmatrix} d_{11} & \dots & d_{1n} \\ \vdots & & \vdots \\ d_{m1} & \dots & d_{mn} \end{bmatrix} \quad (1)$$

其中： $d_{ij} \geq 0, i=1, 2, \dots, m; j=1, 2, \dots, n$ 。

2) 标准化决策矩阵。属性可分为效益型属性和费用型属性 2 类，效益型属性值越大越好，而费用型属性值越小越好。对费用型属性值 d_{ij} 作倒数变换，从而将所有属性统一为效益型。然后对 d_{ij} 使用归一法作比例尺度变换：

$$r'_{ij} = d_{ij} / \sum_{i=1}^m d_{ij} \quad (2)$$

其中 $0 \leq r'_{ij} \leq 1$ ， r'_{ij} 为归一化后的值。故标准化的决策矩阵 $R = (r'_{ij})_{m \times n}$ 。

3) 方案关于属性的熵 E_j 。在信息论中，熵衡量的是一条信息的不确定程度，熵理论广泛应用于计算机科学、经济学、天体物理学等领域^[23-24]。在多属性决策问题中利用各方案的固有信息可得到各属性的信息熵：

$$E_j = -k \sum_{i=1}^m r_{ij} \ln r_{ij}, k = 1 / \ln m \quad (3)$$

当 $E_j=1$ 时， X_j 对于辨别方案优劣不起作用；而当 $E_j=0$ 时， X_j 最能辨别方案优劣。

4) 计算属性权重 ω_j 。设属性对于方案的区分度为 F_j ，则 $F_j=1-E_j$ ，将其归一化后可得属性权重：

性进行排序，由此辨识出制球工艺中的重点工序。

2.2 分析归纳

首先，笔者检索从 2010—2021 年与发射药、推进剂生产工艺安全相关的文献 115 篇作为数据来源，将其处理为 Text 格式后批量导入文本挖掘软件 KH Coder 中，经去停用词、分词完成预处理，该软件可实现词频统计、集群分析、绘制词汇网络图等功能。图 2 为前 25 个高频词汇汇总，图 3 为词汇网络。

由图 2、3 可知，“事故”词频最高，达 182 次。在推进剂、发射药的生产中，事故的突发性、灾难性决定了对事故进行分析研究的必然性。“物料”“工艺”“螺杆”体现出这百余篇文献的研究思路：物料的危险性在外界激励条件(热刺激、机械搅拌等)下导致工艺的危险性，而常规的设备由于未在工艺安全阈值内而失效，最终导致事故。为准确辨识出重点工序，必须充分考虑到物料、工艺和设备及三者之间的关系，其中，工艺以“温度”“压力”

等工艺参数的偏离为研究重点。设备以生产中的工作状态正常与否和防护设计的可靠性为重点研究方向；因此可将物料特性、参数监控、设备状态和防护设计作为辨识重点工序的 4 个指标。

#	词汇	词类/活用	频率
1	事故	Noun	182
2	工艺	Noun	155
3	过程	Noun	145
4	药	Noun	144
5	推进剂	Noun	127
6	生产	Noun	121
7	螺杆	Noun	119
8	损失	Noun	110
9	影响	Noun	108
10	温度	Noun	77
11	可能性	Noun	75
12	发射	Verb	74
13	固体	Noun	72
14	技术	Noun	69
15	费用	Noun	68
16	物料	Noun	65
17	研究	Noun	65
18	挤出	Verb	57
19	界面	Noun	53
20	压力	Noun	52
21	设计	Verb	50
22	运行	Noun	50
23	分析	Verb	47
24	防护	Noun	44
25	生产	Verb	42

图 2 词频

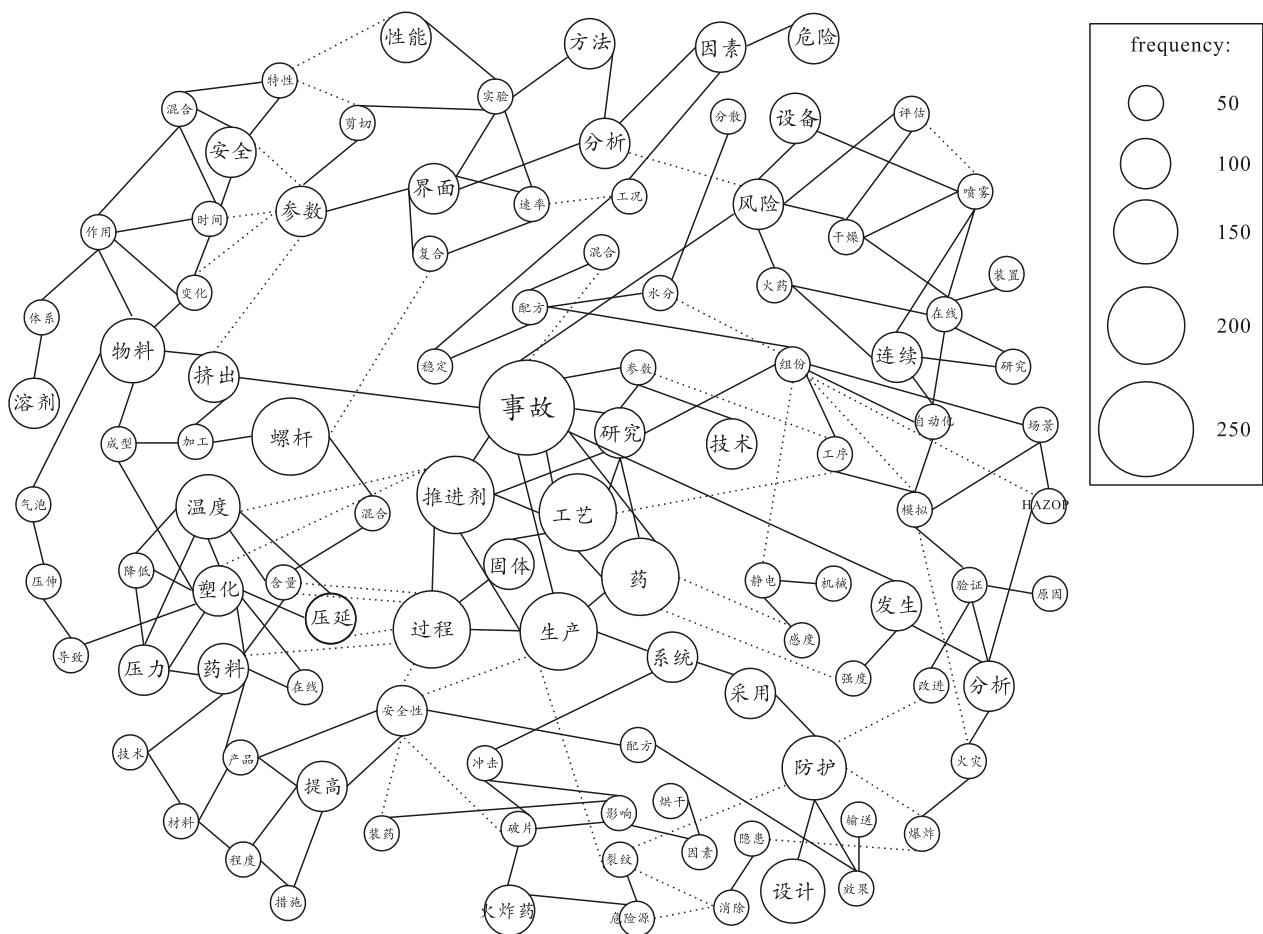


图 3 词汇网络

安全生产风险评估工作中以确定某一特定危险情况发生的可能性和后果为研究内容。经文本挖掘

技术所得高频词“可能性”“影响”“损失”“费用”等不仅与风险评估工作的研究内容相符，还将事故后果的内涵拓展，其严重程度不仅包含经济损失、人身伤亡，还应注意事故发生后所造成的不良社会影响。于是，将事故可能性、事故损失、事故影响作为辨识重点工序的 3 个指标。

综上所述，笔者将设备状态、事故损失、事故影响、参数监控、物料特性、事故可能性以及防护设计这 7 个指标作为判断影响火药安全生产的重点

工序的分析依据。

然后，运用信息熵法确定 7 个指标的权重。因重点工序的确定问题属于多属性决策问题中的一种，故将 7 个工序视为方案，经文本挖掘技术得到的指标视为属性。查阅历史同类型事故数据并邀请 4 名火药生产专家、6 名风险评估专家、3 名原制球工艺线上且有 10 年以上工龄的老员工及 2 名连续化生产线的设计工程师共同为该生产线的工序打分，取其中位数形成决策表格如表 1 所示。

表 1 决策表格

工序名称	设备状态	可能损失/万元	事故影响	参数监控	物料特性	事故可能性	防护设计
驱水	8	20	9	7	8	8	8
塑化	8	25	9	7	7	8	9
挤出	6	33	8	5	6	9	7
切粒	5	30	8	5	5	6	7
整形	9	13	7	9	8	4	8
驱溶	7	32	7	6	7	5	7
辅助	9	6	2	9	9	3	9

因设备状态、参数监控及防护设计是费用型属性，应对其取倒数使之统一为效益型属性，故可得到如下决策矩阵：

$$D = \begin{bmatrix} \frac{1}{8} & 20 & 9 & \frac{1}{7} & 8 & 8 & \frac{1}{8} \\ \frac{1}{8} & 25 & 9 & \frac{1}{7} & 7 & 8 & \frac{1}{9} \\ \frac{1}{6} & 33 & 8 & \frac{1}{5} & 6 & 9 & \frac{1}{7} \\ \frac{1}{5} & 30 & 8 & \frac{1}{5} & 5 & 6 & \frac{1}{7} \\ \frac{1}{9} & 13 & 7 & \frac{1}{9} & 8 & 4 & \frac{1}{8} \\ \frac{1}{7} & 32 & 7 & \frac{1}{6} & 7 & 5 & \frac{1}{7} \\ \frac{1}{9} & 6 & 2 & \frac{1}{9} & 9 & 3 & \frac{1}{9} \end{bmatrix}。$$

对该矩阵归一化处理得：

$$R = \begin{bmatrix} 0.127 & 0.126 & 0.180 & 0.133 & 0.160 & 0.186 & 0.139 \\ 0.127 & 0.157 & 0.180 & 0.133 & 0.140 & 0.186 & 0.123 \\ 0.170 & 0.208 & 0.160 & 0.186 & 0.120 & 0.209 & 0.159 \\ 0.204 & 0.189 & 0.140 & 0.186 & 0.100 & 0.140 & 0.159 \\ 0.113 & 0.082 & 0.140 & 0.103 & 0.160 & 0.093 & 0.139 \\ 0.146 & 0.201 & 0.140 & 0.155 & 0.140 & 0.116 & 0.159 \\ 0.113 & 0.038 & 0.040 & 0.103 & 0.180 & 0.070 & 0.123 \end{bmatrix}。$$

由式(3)、(4)可得：

$$E_j = (0.989 \ 0.947 \ 0.968 \ 0.987 \ 0.992 \ 0.969 \ 0.997)；$$

$\omega_j = (0.075 \ 0.347 \ 0.213 \ 0.086 \ 0.053 \ 0.207 \ 0.018)$ 。故由信息熵法可得各判断指标的权重依次为 0.075, 0.347, 0.213, 0.086, 0.053, 0.207 和 0.018。

运用 TOPSIS 法综合得出工序重要性排序。先是对原决策矩阵模一化并加权处理得规范化加权矩阵，由式(5)得：

$$V = \begin{bmatrix} 0.025 & 0.106 & 0.097 & 0.029 & 0.022 & 0.097 & 0.007 \\ 0.025 & 0.133 & 0.097 & 0.029 & 0.019 & 0.097 & 0.006 \\ 0.033 & 0.176 & 0.086 & 0.041 & 0.017 & 0.109 & 0.008 \\ 0.040 & 0.160 & 0.086 & 0.041 & 0.014 & 0.072 & 0.008 \\ 0.022 & 0.069 & 0.075 & 0.023 & 0.022 & 0.048 & 0.007 \\ 0.028 & 0.170 & 0.075 & 0.034 & 0.019 & 0.060 & 0.008 \\ 0.022 & 0.032 & 0.022 & 0.023 & 0.025 & 0.036 & 0.006 \end{bmatrix}。$$

由式(6)–(8)可依次求得：

$$S^+ = (0.073 \ 0.049 \ 0.015 \ 0.042 \ 0.127 \ 0.055 \ 0.180)；$$

$$S^- = (0.123 \ 0.140 \ 0.175 \ 0.150 \ 0.067 \ 0.151 \ 0.011)。$$

并将相对接近度归一化：

$$C = (0.149 \ 0.177 \ 0.219 \ 0.185 \ 0.082 \ 0.174 \ 0.014)。$$

由此可得制球工艺中 7 个工序重要性排序为：挤出 > 切粒 > 塑化 > 驱溶 > 驱水 > 整形 > 辅助。于是该制球工艺中的重点工序为挤出、切粒和塑化工序，该结果与实际相符。在后续对其做风险评估时需重点分析挤出、切粒和塑化工序的风险。

3 结论

1) 运用文本挖掘技术对近十年火药生产工序

安全相关文献的分析归纳, 得出辨识重点工序的 7 个指标是设备状态、事故损失、事故影响、参数监控、物料特性、事故可能性以及防护设计, 其权重分别为 0.075, 0.347, 0.213, 0.086, 0.053, 0.207 和 0.018。

2) 发射药连续化制球工艺中, 挤出、切粒和塑化工序是其最为重要的工序, 后续的评估工作中需重点分析这 3 个重点工序的风险。

参考文献:

- [1] 曹楷, 叶义成, 骆晶晶, 等. 改进的 HAZOP-LEC 法在乳化炸药生产工艺中的应用[J]. 工业安全与环保, 2018, 44(9): 42-45, 85.
- [2] 冯睿智, 张树海, 苟瑞君, 等. 模糊综合评价在单基发射药生产安全评价中的应用[J]. 中国安全生产科学技术, 2012, 8(9): 128-132.
- [3] 钱辉, 杨森, 刘小琴, 等. 固体推进剂捏合工序燃爆事故的模糊事故树分析[J]. 安全与环境工程, 2018, 25(6): 127-131, 138.
- [4] 刘保顺, 马忠亮. 基于模糊数学改进的 BZA-1 法在单基药生产过程安全评价中的应用[J]. 山西化工, 2016, 36(1): 86-90.
- [5] 余成早, 朱克美, 周兵, 等. 火炸药安全事故分析与预防措施对策探讨: 全国危险物质与安全应急技术研讨会论文集(下)[C]. 重庆: 重庆市人民政府、中国工程物理研究院, 2011: 667-670.
- [6] 何那仁朝格图, 邓国栋, 鲁磊明, 等. 改性双基推进剂制备中硝化甘油的连续加料与计量技术[J]. 爆破器材, 2017, 46(4): 59-64.
- [7] ZHONG T T, RAO G N, PENG J H. Numerical simulation of three dimensional flow fields for extrusion process of GR-35 double-base propellant: 2014 International Symposium on Safety Science and Technology[C]//Beijing, China, 2015: 920-926.
- [8] 柴涛, 刘玉存, 吕林川. 硝化甘油生产过程中硝化工序的安全分析[J]. 兵工安全技术, 1998(1): 17-19.
- [9] FELDMAN R, DAGAN I. Knowledge Discovery in Textual Databases (KDT)[C]. National Conferences on Artificial Intelligence 1998-1999, 2009: 1857-1862.
- [10] 谭章禄, 彭胜男, 王兆刚. 基于聚类分析的国内文本挖掘热点与趋势研究[J]. 情报学报, 2019, 38(6): 578-585.
- [11] 梅馨, 邢桂芬. 文本挖掘技术综述[J]. 江苏大学学报(自然科学版), 2003, 24(5): 72-76.
- [12] PATRICIA C, ROSA P, LUISA M. Research lines on the impact of the COVID-19 pandemic on business. A text mining analysis[J]. Journal of Business Research, 2020, 132(c): 586-593.
- [13] SWAGATO C, DIVESH G, ATUL P, et al. Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application[J]. Journal of Business Research, 2020, 131(c): 815-825.
- [14] 钟美华. 基于非结构化数据管理平台研究与建设[J]. 中国新通信, 2020, 22(23): 57-58.
- [15] 高永涛, 赵宇松, 崔柔杰, 等. 基于文本挖掘和变权 AHP 法的隧道施工工期安全状态评估[J]. 昆明理工大学学报(自然科学版), 2020, 45(6): 128-139.
- [16] 吴建军. 基于文本挖掘的铁路信号故障分析[J]. 集成电路应用, 2020, 37(12): 34-36.
- [17] 徐日华. 基于数据挖掘技术的安全生产事故数据分析研究[D]. 北京: 北京化工大学, 2017.
- [18] 程慧荣, 黄国彬, 郑琳. 非结构化文本分析软件比较研究—以 KH Coder 和 Wordstat 为例[J]. 图书与情报, 2015(4): 110-117, 122.
- [19] ZHANG X C, SHAN W C, ZHANG Z W, et al. AE monitoring of reinforced concrete squat wall subjected to cyclic loading with information entropy-based analysis[J]. Engineering Structures, 2018, 165: 359-367.
- [20] 张完定, 王广三. 基于 TOPSIS-熵权法的陕西创新驱动发展评价指标研究[J]. 长安大学学报(社会科学版), 2019, 21(1): 32-41.
- [21] 孟超, 李晓萌, 孙玉岭, 等. 基于 ISM 和信息熵的铁路隧道塌方风险分析[J]. 中国安全科学学报, 2018(Z1): 64-70.
- [22] 刘倩, 谷晓燕, 贾子睿. 基于熵权-TOPSIS 法的重大错报风险评估研究[J]. 北京信息科技大学学报(自然科学版), 2019, 34(5): 73-77.
- [23] COIFMAN R R, WICKERHAUSER M V. Entropy-based algorithms for best basis selection[J]. IEEE Trans Information Theory, 1992, 38(2): 713-718.
- [24] SKILLING J, BRYAN R. Maximum entropy image reconstruction: general algorithm[J]. MNRAS, 1984, 211(1): 111-124.
- [25] 王飞俊, 杨斐霏, 王江宁, 等. NGEC 基改性双基推进剂的制备及性能[J]. 火炸药学报, 2006, 29(6): 51-53.