Mine Engineering 矿山工程, 2024, 12(3), 388-394
Published Online July 2024 in Hans. https://www.hanspub.org/journal/me
https://doi.org/10.12677/me.2024.123047

基于N-gram剪枝技术的隐患文本自动评估模型

叶洪胜[1]，刘　洪[2]，周宝山[3]，兰　莉[2]，邹巧兰[2]，周啟梦[2]，王海宇[2]

[1]重庆科技大学智能技术与工程学院，重庆
[2]重庆科技大学安全工程学院，重庆
[3]中国石油冀东油田分公司，河北 唐山

收稿日期：2024年5月9日；录用日期：2024年6月28日；发布日期：2024年7月10日

摘　要

为了自动分析海上钻井平台隐患文本中蕴含的隐患响应程度信息，量化隐患严重程度，提出一种基于N-gram词袋向量的隐患响应等级量化评估模型。首先针对1565条钻井平台的现场隐患记录进行分词与过滤处理；其次再以N-gram作为特征单元重塑词袋维度；然后提出使用逆TF-IDF值来强化特征值；最后，使用朴素贝叶斯构建隐患量化模型。结果表明：使用该方法的隐患量化评估模型具有较高的精确率、召回率及F1值。

关键词

语义分析，钻井平台，N-gram，词袋向量，隐患量化

An Automatic Assessment Model Based on N-gram Pruning Technique for Hidden Danger Text

Hongsheng Ye[1], Hong Liu[2], Baoshan Zhou[3], Li Lan[2], Qiaolan Zou[2], Qimeng Zhou[2], Haiyu Wang[2]

[1]School of Intelligent Technology and Engineering, Chongqing University Science and Technology, Chongqing
[2]School of Safety Engineering, Chongqing University Science and Technology, Chongqing
[3]PetroChina Jidong Oilfield Branch, Tangshan Hebei

Received: May 9th, 2024; accepted: Jun. 28th, 2024; published: Jul. 10th, 2024

Abstract

To automatically analyze the response level information of hidden dangers contained in hidden danger texts and quantify the severity, a quantitative evaluation model based on N-gram word bag

文章引用: 叶洪胜, 刘洪, 周宝山, 兰莉, 邹巧兰, 周啟梦, 王海宇. 基于 N-gram 剪枝技术的隐患文本自动评估模型[J]. 矿山工程, 2024, 12(3): 388-394. DOI: 10.12677/me.2024.123047

vectors is proposed for the response level of hidden dangers. Firstly, segment and filter the on-site hazard records of 1565 drilling platforms; Secondly, using N-gram as feature units to reshape the bag of words dimension; Then, it is proposed to use the inverse TF-IDF value to enhance the feature values; Finally, use naive Bayes to construct a hazard quantification model. The results show that the hazard quantification evaluation model using this method has high accuracy, recall, and F1 value.

Keywords

Semantic Analysis, Drilling Platforms, N-gram, Word Bag Vector, Hazard Quantification

1. 引言

近年来，随着陆上油气资源进入瓶颈期，我国日渐突出的能源供需矛盾迫使油田开发转向海上，海上油气开采作业越来越多。钻井平台在油气田的开采中起着决定性的作用，海上油气资源开发具有高投入、高技术、高风险等特点，平台的定期检查与隐患评估关系着人员的安危和财产安全造成的损失。由于人员设备高度集中且位于海上，一旦发生事故应急救援困难，对海洋环境也会带来严重影响[1]。例如，我国蓬莱钻井平台井涌事故，B 平台轻质油溢出 13 天未引起重视，C 平台又发生油基泥浆渗漏[2]，油井紧急封闭并造成渤海湾水体污染，生态损害无法估量。因此，利用 nlp 技术挖掘隐患记录背后蕴含的信息，量化隐患严重程度，让生产管理单位更直观的掌握平台状态，为安全管理人员提供有效决策依据，对配置人员高效响应隐患问题具有重要意义。

词袋模型[3]-[5]是自然语言处理和信息检索中的一种典型方法，能将文档转化成特征单元并用于分类或微妙意图分析，其中典型的代表如 one-hot，TF-IDF 等表示方法，在安全隐患文本分析领域有较为广泛的应用。Yildirim 等[6]提出基于单词的句子情感预测，但是仅仅将词袋维持在单词的维度，模型分析文本含义精度不高，而将维度拓展到 n 元组短句又面临维度灾难的问题。谭章禄等[7]基于文本聚类对煤矿安全隐患类型进行了挖掘和研究。陈孝慈等[8]提出利用 Bigram 二字串作为特征单元，结合 SVM 对安全隐患文本分类，为海上钻井平台隐患分析引入了运用高维短句分析的思想。胡瑾秋等[9]使用 TF-IDF 计算词在文档中重要性特征值结合关联规则算法挖掘企业生产事故中的隐患，为企业安全管理提供了预警和可视化方案。黄春梅等[10]使用词袋模型结合 TF-IDF 提取文本特征，对短文本进行分类研究，得到了较高的归类准确率，为文本特征提取提供了参考方向，为 nlp 技术引入隐患短文本分析的应用奠定了基础。孟涛等[11]使用短文本拓展后的特征向量，得出了拓展维度有助于处理短文本的结论。韩天园等[12]基于文本挖掘算法剖析重大交通事故案例，构建了重大交通事故的层级模型。目前改进的词袋模型及文档特征值表示方法仅仅只是扩充和保留更多的词项特征维度[13]或者替换特征矩阵的值[14]来优化文档分类和信息挖掘效果，在对于更深层次和更隐晦的含义挖掘上效果不尽人意[15]。这些研究主要利用词袋模型重塑文档或者使用 TF-IDF 算法来计算文档中的词分布及重要性，但对于出现在同领域且词差异性小的隐患分析评估任务上，准确率不高，且传统方法随文档量增加词袋维度会迅速扩张，为了保证隐患文本的自动分析和量化的可靠性，需要一种改进方法来优化传统模型。

因此，本文对收集到的 1565 条海上钻井平台的现场隐患记录，使用 N-gram 字串和支持度置信度算

法，在将词袋模型从单词拓展到短语的情况下，利用统计学中的关联规则挖掘算法对维度进行剪枝，使词袋维度在大量收缩的同时仍保留有效的隐患信息，避免了模型的过拟合问题。并且提出了一种基于逆TF-IDF 值的特征值增强方法，进一步提高了隐患评估模型的精度，为海上平台隐患文本的无人分析及评估提供了方法与思路。

2. 平台安全隐患文本词袋构建

2.1. 海上平台隐患文本特征

以某石油公司 2017~2021 年的 1565 条隐患记录为数据源，其中原始数据包括设备分类、隐患类别、描述、是否有直接 HSE 风险、发现时间等(表 1)。在对文本反应的安全隐患进行评估时，遵照公司隐患响应等级标准(表 2)，主要分析表 1 中的隐患描述。

Table 1. Raw data information of offshore platform hazards
表 1. 海上平台隐患原始数据信息

设备分类	隐患类别	描述	是否有直接 HSE 风险	发现时间
设备设施类	设备缺陷	F33H 井井下漏液压油情况较严重，导致井口盘油位下降较快	否	2017-01-19
电气设备	其它	下层甲板设备接地线断开	否	2018-05-31
平台结构	设计缺陷	油田 CEPJ 平台生活楼四楼吊货平台应加踢脚板	是	2021-03-23

Table 2. Hazard response level standard for offshore platforms
表 2. 海上平台隐患响应等级标准

隐患等级	响应等级	影响范围	示例	影响程度
特别紧急	1	设备本身	断裂/破裂/漏油等	实质性损坏，已经影响使用
紧急	2	设备本身	轻微磨损/松动等	实质性损伤，即将影响使用
较紧急	3	非本身/其他设备	安全标识/设计隐患/灯具等	非实质性损害
一般紧急	4	记录	相关记录	纸质档记录

2.2. 特征单元分析

在挖掘文本蕴藏的隐患信息前，中文文本需要先进行分词和停用词过滤。本文采用 Jieba 中文分词工具包和通用停用词库对隐患文本进行预处理，将文档转化成词列表。有别于微博等社交型文本，每条安全隐患文本的词汇数在 10 到 30 之间，语言精炼且包含庞大的信息量。如果仅使用词作为特征，过少的特征单元难以保证隐患评估的最终效果。因此将词进行连接，扩展成 N-gram 短语用以拓展词袋维度是较为有效的方法。

由于用作构建 N-gram 的词数越多，其在整个领域的文档中出现的次数就越少，超过 Four-gram 的短语基本上已经失去了作为最小特征单元的意义。使用 Bi-gram，Tri-gram，Four-gram 来拓展词袋维度具有较好的拟合能力和效果。以 Tri-gram 字串为例，即是以相邻三个词构成的短语作为特征单元。词袋模型是将文档视作最小词单元的线性组合，通过构建以词为列的矩阵来表达原文档，在用于垃圾邮件分类问题或多领域文本的分类上具有较好效果。但由于其忽略了词序及组合的信息，在词的出现基本相同的同领域文档的语义分析问题上收效甚微。而将最小单元拓展到 N-gram 短语则在一定程度上克服了这个问题，以“单点带缆甲板”为例，分词工具将其切割成由“单点、带缆、甲板”三个词汇组成的列表，失去了词序信息的排列组合很难真正等同于原始短语。在隐患文本这个领域中，“单点－带缆－甲板”这

个 Tri-gram 字串才是真正指向某个具体设备的最小单位信息。

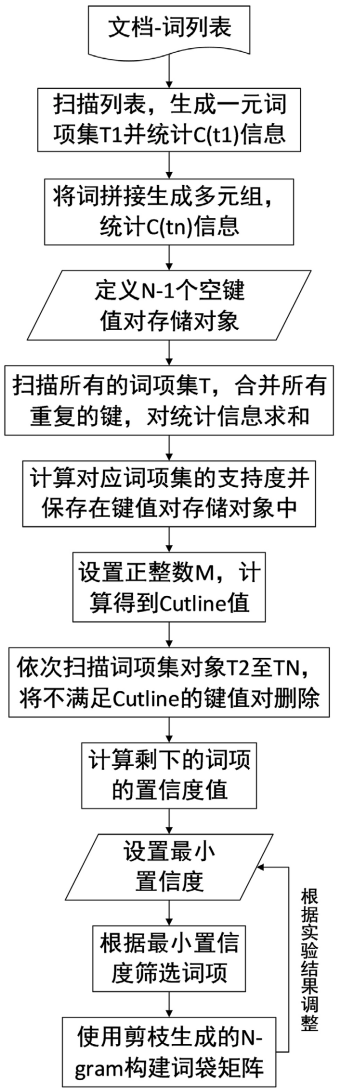

Figure 1. N-gram flow chart of pruning generation algorithm
图 1. N-gram 剪枝生成算法流程图

如果仅仅只是依照文档的词序生成不重复的 N-gram 字串，词袋的维度会扩张到一个极为庞大的状态，这不仅会生成过多的噪音特征来隐藏真正有用的信息，造成判断精度的下降，由于维度过大，计算资源的消耗也会造成判定时间过长的问题。而面对用于描述一整个海上平台隐患领域的文本信息，人工定义这些短语的工程量又太过庞大，几乎难以实现。为此，利用现有的隐患记录资源，实现无监督的 N-gram 生成是更为有效和可行的方式。

2.3. N-gram 剪枝生成算法

将每条文档的词与 N-gram 出现的次数作为特征值填入矩阵，选择适当的参数训练，构造朴素贝叶斯模型。最后，利用此模型就可以搭建一条隐患评估处理流水线。尽管，这样简单的剪枝算法达到了抑制维数灾难的目的，但也损失了相当多的信息，这些信息或许出现的次数并不多，但它对于隐患的评估却

很重要。并且随着文档数的增加，剪枝的阈值线无法变化的问题也会导致“过滤”失效。为了解决上述问题，本文提出一种 N-gram 剪枝生成算法。

N-gram 的剪枝生成算法步骤如下：首先扫描切分好的文档–词列表，得到每条文档的词项集 T_1 及对应的统计信息 $C(t_{1j})$，同时将 t_{ij} 按文本顺序拼接生成 T_2 至 T_N 并统计对应的统计信息 $c(t_{ij})$，构建一个不重复的键值对存储对象，其中键为词项，值为统计信息。将文档列表中重复出现的词项的统计值加和，直到扫描完所有训练集文档，该步骤仅需扫描一次数据集。然后读取该键值对存储对象计算出每个词项的支持度，设置正整数 M，计算出剪枝度量线 Cutline。以该值作为支持度过滤线，筛选出大于该线的词项，完成第一次剪枝。设置最小置信度，扫描剩下的词项，计算出置信度，取满足最小置信度的词项作为剪枝生成的最终词项集，算法结束。算法流程图如图 1 所示。

N-gram 剪枝生成算法可以无监督的生成 N-gram 特征单元且在生成步骤仅需一次扫描，剪枝度量线也克服了文档数对剪枝阈值的影响，算法经由两步剪枝得到最终的特征维度，其中第一次剪枝会快速收缩矩阵维度，第二次剪枝的最小置信度需要人工调整，是对特征维度的微调。

3. 词袋特征值增强

3.1. 词袋特征值构建

以文档中词和 N-gram 出现的次数来衡量它对评估的权重影响是可行的方法，但同样高频率这一数学特征也制约了评估模型精度的进一步上升。由于人拥有比机器更高级的语言能力，能够很轻易的分辨出一些在衡量隐患上更重要的信息，尽管这些信息出现的次数并不多。而对于机器和数学模型来说，它们并没有人在漫长的人生中不断积累并成长的外部语言模型来解决隐患评估这一问题。因此笔者提出一种基于逆 TF-IDF 值来进行词袋特征值增强的算法，进一步提升了模型的隐患评估能力。

3.2. 词袋特征值增强算法

笔者依旧使用词出现数作为特征值主体，转而使用逆 TF-IDF 值来增强这一信息。使得模型在隐患评估时，能够在某种程度上克服高频词的影响，使更多的 N-gram 参与到隐患评估任务中，实验结果也表明了该算法进一步提升了模型的评估准确度，计算公式如下：

$$\text{StrengthenFV}(t_{ij}) = -\log(TF(t_{ij}) \times IDF(t_{ij})) \times c(t_{ij}) \tag{1}$$

此处的 $c(t_{ij})$ 表示某一篇隐患文档中词项的出现次数，而非文档集词项出现的总次数。

4. 试验结果与分析

4.1. 文本隐患评估评价指标选择

本文使用机器学习中最常使用的分类任务评价指标，精确率(Precision)、召回率(Recall)、F1-score [16] 来衡量该模型对文本的识别和挖掘能力。考虑到模型解决的任务为四分类任务，而四种不同的隐患等级文本数量并非均衡，因此在计算以上三项指标时选择了权重(weighted)参数进行优化，见表 3。

Table 3. Text hidden danger evaluation indicators
表 3. 文本隐患评估评价指标

评价指标	含义
精确率 P	正确预测某类的样本数与预测某类总数的比例，衡量查准率
召回率 R	正确预测某类的样本数与实际数量的比例，衡量查全率
$F1-score(F1-值)$	精确率与召回率的加权平均值，F1-score 值越高说明模型越稳健

4.2. 实验结果及对照分析

本文以传统的词袋向量构造方法和 TF-IDF 值替换特征值作为基准模型进行实验对比。经过训练得到上述几种模型后将 1565 条数据作为测试集让模型进行隐患评估并输出结果，与实际结果比照并计算评价指标。在词袋重构步骤，不进行剪枝、传统剪枝方法及 N-gram 剪枝生成算法下词袋各词项集维度见表 4。在特征值填充步骤，根据不同的剪枝算法构造出的词袋，分以 TF-IDF 值为基、词项数为基和以本文中增强算法计算出的值作为特征值，实验得到的结果如表 5。

Table 4. Details of dimension generation of each item set in thesaurus
表 4. 词袋各词项集维度生成详情

词项集	未剪枝/词项维度	传统剪枝方法/词项维度	N-gram 剪枝生成算法/词项维度
words	3990	3990	3990
bi-gram	13,978	598	906
tri-gram	16,624	283	1332
four-gram	16,801	197	949

Table 5. Control experimental results
表 5. 对照实验结果

模型	P	R	*F*1 – *score*
传统剪枝 + 词项数	0.770	0.764	0.760
n-gram 剪枝 + TF-IDF	0.138	0.371	0.201
n-gram 剪枝 + 词项数	0.810	0.809	0.805
n-gram 剪枝 + 特征值增强算法	0.863	0.856	0.857

从表 4 的维数统计情况来看，传统依据词项数进行剪枝的方法与本文中提出的 N-gram 剪枝算法均能无监督地将词项集压缩到相同的数量级。与受词连接后数量不断减小这一现象影响的传统算法相比，新方法保留的维度数呈现高斯分布，能更好的保留有效特征短语。从实际的语言理解上来说，三个词组成的短语更容易出现特定的搭配，随着维数上升，这种性质又会被词出现的组合情况增加所淡化。从对照实验中也可以看到，仅在剪枝步环节进行优化，模型的性能约有 5%左右的提升。对于该任务，在特征值的选取上，TF-IDF 值并不是一个有效的选择。但从召回率大于 25%可以看出，以 TF-IDF 值为基的模型发现了某些不同于词项数的特征，而这部分特征贡献了 12%的召回率。因此将 TF-IDF 值作为一种权重调整手段，使用特征值增强算法填充特征值的模型又较仅使用词项数作特征的模型整体有了接近 5%的提升。

5. 结论

1) 本文提出的方法在不需要人工识别隐患文本的前提下取得了更好的效果，也可以将该方法作为特征筛选的初步处理，再对处理后的词项集进行人工筛选，进一步精简模型需要处理的词项维度。

2) 实验取得了 85%以上的准确度，多数错误出现在短文本以及难界定的隐患问题上。过短的隐患文本包含的短语特征太少，模型很难对其反应的现象作出准确的判断。这种情况在这些文本还有一半是干扰文字的影响下更为明显。

3) 本文基于浅层的语义信息，对于计算资源的需求并不高，在特征识别的过程中若出现近义词或错别字，并且在模型训练的过程中词袋并未收录这种情况，模型将会忽略掉这个词汇。在不影响评估精度的情况下提升模型的泛化能力是未来的研究方向。

参考文献

[1] 崔青. 海洋平台发展现状及前景[J]. 石化技术, 2018, 25(6): 213.

[2] 何沙, 陈东升, 朱林, 姬荣斌. 海上钻井平台安全风险预警模型应用研究[J]. 中国安全生产科学技术, 2012, 8(4): 148-154.

[3] 赵京胜, 宋梦雪, 高祥. 自然语言处理发展及应用综述[J]. 信息技术与信息化, 2019(7): 142-145.

[4] Zhi, Y.Z., Bo, F., Hang, Q., Yan, L.Z. and Xiao, B.L. (2017) Modeling Medical Texts for Distributed Representations Based on Skip-Gram Model. 2017 3*rd International Conference on Information Management* (*ICIM*), Chengdu, 21-23 April 2017, 279-283. https://doi.org/10.1109/INFOMAN.2017.7950392

[5] Yan, X.Y. (2017) Research and Realization of Internet Public Opinion Analysis Based on Improved TF-IDF Algorithm. 2017 16*th International Symposium on Distributed Computing and Applications to Business, Engineering and Science* (*DCABES*), Anyang, 13-16 October 2017, 80-83. https://doi.org/10.1109/DCABES.2017.24

[6] Gökçay, D., Işbilir, E. and Yildirim, G. (2012) Predicting the Sentiment in Sentences Based on Words: An Exploratory Study on ANEW and ANET. 2012 *IEEE* 3*rd International Conference on Cognitive Infocommunications* (*CogInfo-Com*), Kosice, 2-5 December 2012, 715-718. https://doi.org/10.1109/CogInfoCom.2012.6421945

[7] 谭章禄, 王兆刚, 胡翰, 姜萱, 彭胜男. 基于文本聚类的煤矿安全隐患类型挖掘研究[J]. 中国安全科学学报, 2019, 29(3): 145-148.

[8] 陈孝慈, 谭章禄, 单斐, 高青. 基于 Bigram 的安全隐患文本分类研究[J]. 中国安全科学学报, 2017, 27(8): 156-161.

[9] 胡瑾秋, 张曦月, 吴志强. 结合 TF-IDF 的企业生产隐患关联预警及可视化研究[J]. 中国安全科学学报, 2019, 29(7): 170-176.

[10] 黄春梅, 王松磊. 基于词袋模型和 TF-IDF 的短文本分类研究[J]. 软件工程, 2020, 23(3): 1-3.

[11] 孟涛, 王诚. 基于扩展短文本词特征向量的分类研究[J]. 计算机技术与发展, 2019, 29(4): 57-62.

[12] 韩天园, 田顺, 吕凯光, 李旋, 张佳涛, 魏朗. 基于文本挖掘的重特大交通事故成因网络分析[J]. 中国安全科学学报, 2021, 31(9): 150-156.

[13] 李然, 林政, 林海伦, 王伟平, 孟丹. 文本情绪分析综述[J]. 计算机研究与发展, 2018, 55(1): 30-52.

[14] 洪巍, 李敏. 文本情感分析方法研究综述[J]. 计算机工程与科学, 2019, 41(4): 750-757.

[15] 谭章禄, 陈晓, 宋庆正, 陈孝慈. 基于文本挖掘的煤矿安全隐患分析[J]. 安全与环境学报, 2017, 17(4): 1262-1266.

[16] 奉国和. 文本分类性能评价研究[J]. 情报杂志, 2011, 30(8): 66-70.