工程科学学报
Chinese Journal of Engineering

基于文本挖掘的矿山安全隐患大数据分析与可视化

郭对明 李国清 胡乃联 侯杰

Big data analysis and visualization of potential hazardous risks of the mine based on text mining

GUO Dui-ming, LI Guo-qing, HU Nai-lian, HOU Jie

引用本文:
郭对明, 李国清, 胡乃联, 侯杰. 基于文本挖掘的矿山安全隐患大数据分析与可视化[J]. 工程科学学报, 2022, 44(3): 328-338. doi: 10.13374/j.issn2095-9389.2020.10.23.004
GUO Dui-ming, LI Guo-qing, HU Nai-lian, HOU Jie. Big data analysis and visualization of potential hazardous risks of the mine based on text mining[J]. *Chinese Journal of Engineering*, 2022, 44(3): 328-338. doi: 10.13374/j.issn2095-9389.2020.10.23.004

在线阅读 View online: https://doi.org/10.13374/j.issn2095-9389.2020.10.23.004

工程科学学报, 第 44 卷, 第 3 期: 328−338, 2022 年 3 月
Chinese Journal of Engineering, Vol. 44, No. 3: 328−338, March 2022
https://doi.org/10.13374/j.issn2095-9389.2020.10.23.004; http://cje.ustb.edu.cn

基于文本挖掘的矿山安全隐患大数据分析与可视化

郭对明[1,2)], 李国清[1,2)]✉, 胡乃联[1,2)], 侯　杰[1,2)]

1) 北京科技大学土木与资源工程学院, 北京 100083　2) 金属矿山高效开采与安全教育部重点实验室, 北京 100083
✉通信作者, E-mail: qqlee@ustb.edu.cn

摘　要　基于大数据分析技术, 构建了矿山安全隐患多维度分析模型, 分析了隐患在时间和空间两个维度上的分布规律; 利用主题挖掘模型将众多隐患信息归类, 得到了 13 个隐患主题; 利用关联规则挖掘模型探究了不同隐患之间的内在联系, 并利用 R 编程语言对上述结果进行可视化展示. 通过对安全隐患的分析研究不仅充分利用了矿山隐患数据, 避免了数据资源的浪费, 同时也对矿山井下事故预防有一定的指导价值.

关键词　矿山安全; 文本挖掘; 隐患数据; 大数据分析; 可视化
分类号　TD77.1

Big data analysis and visualization of potential hazardous risks of the mine based on text mining

GUO Dui-ming[1,2)], *LI Guo-qing*[1,2)]✉, *HU Nai-lian*[1,2)], *HOU Jie*[1,2)]

1) School of Civil and Resource Engineering, University of Science and Technology Beijing, Beijing 100083, China
2) Key Laboratory of High-Efficient Mining and Safety of Metal Mines, Ministry of Education, Beijing 100083, China
✉ Corresponding author, E-mail: qqlee@ustb.edu.cn

ABSTRACT　Compared with other production industries, metal mine is recognized as a high accident rate and the highest casualty rate due to the bad working environment. Therefore, safety production is the key concern of mining enterprises. With the attention of enterprises to safety problems and the increasing improvement of mine safety management system, many mines have established secure big data platform to effectively manage production and ensure the safety of underground operation, receiving the safety hazard information from daily safety inspection into the platform. However, due to the data of security risks are unstructured short texts with the operation of the enterprise, including the data recorded in the platform presents the characteristics of complex data content, large data scale, and non-standard data records. Moreover, due to the lack of an effective text analysis model, a small part of the security risk data is only used for simple analysis such as report analysis and data statistics, whereas more data is stored in a secure big data platform. Thus, the data did not play a guiding role in production, resulting in a waste of these valuable data resources. In order to explore the internal relationship between hidden danger data and the rule of hidden danger occurrence, based on big data analysis technology, this paper constructed a multi-dimensional analysis model of mine safety hidden danger. We analyzed the distribution law of hidden danger in two dimensions of time and space, used the topic mining model to classify hidden danger information, and obtained 13 hidden danger topics, using association rules to mine hidden danger. The model explores the internal relationship between different hidden dangers and uses an R programming language to visualize the above results. The results made full use of the mine hidden danger data and avoided the waste of data resources through the analysis and research of the hidden danger with a certain guiding value for preventing mine accidents.

KEY WORDS　mine safety; text mining; data of hidden danger; data analysis; data visualization

收稿日期: 2020−10−23
基金项目: 国家自然科学基金资助项目(52074022); 中央高校基本科研业务费专项资金资助项目(FRF-TP-20-001A1)

金属矿山由于作业条件复杂、劳动环境恶劣，被公认为是事故高发且伤亡率最高的行业之一[1]. 因此，安全生产成为矿山企业永恒的主题，安全隐患管理也受到了国家和企业的重视[2−3]. 随着矿山安全管理体系日益完善，针对事故发生机理在人类可控范围内对安全隐患进行及时的辨识、处理与监控是矿山安全生产管理的重要手段. 对不同的隐患数据有不同的辨识分析方法，Martin 和 Morris 提出建立被控过程模型，通过数学模型将研究对象的可测信息和通过模型表达的先验信息进行比较，对残差结果进行分析处理，完成了对安全隐患的确定[4]，通过利用故障关系的先验模型建立知识模型，利用被监控对象的定性描述建立定性模型，从而完成了对安全隐患的定位与识别[5]. Dunia 等[6] 提出在描述对象的精确性及建模的可行性上，介于以上两种方法之间，通过相关的频谱分析、主元分析、小波变换等工具，直接分析可测信号，提取诸如方差、幅值、频率等特征值，从而检测安全隐患的存在. 李季等[7] 提出了完整、科学的危险源信息和隐患辨识数据库，然后结合矿山监测系统和人工监测提供的实时数据，完成了安全隐患的捕捉与辨别. 秦文静[8] 通过事故树原理，建立煤矿井下瓦斯爆炸危险源事故树，对煤矿瓦斯爆炸危险源进行辨识. 张宝隆等[9] 提出了基于本体的隐患辨识排查系统构建的方法，通过对煤矿隐患知识分析，建立了隐患本体层次结构，定义类的对象和属性，构建了煤矿事故隐患辨识排查系统模型，从而解决了煤矿事故排查效率低，排查不到位等问题.

为了有效分析安全隐患信息，有学者尝试了将大数据分析技术应用到矿山安全管理中. 马小平和代伟[10] 通过总结大数据技术在煤炭工业中的应用，分析了大数据在煤矿设备故障诊断、灾害事故预警与防治等方面的可行性. 孙继平[11] 运用大数据技术实现了煤矿事故灾害的超前预警. 谭章禄等[12] 借助文本分析方法，通过对隐患信息的预处理，得到隐患事故高频词，进一步指导隐患治理. 钱宇虹[13]、石记斌和石记红[14]、雷煜斌等[15] 采用数据挖掘技术，应用 Apriori 算法和 FP-growth 算法分析瓦斯与地质构造、煤结构等因素间的关联关系.

随着计算机的发展，很多矿山搭建了安全大数据平台或相应的管理系统[16−18]，并将安全检查中发现的隐患信息录入到大数据平台中. 随着企业运行，平台中会积累海量以安全检查信息为主的非结构化文本数据，而且数据具有内容繁杂、规模大、不规范等特征. 据数据显示，矿山一年的数据量可达上百 GB[19−20]. 虽然大数据平台为安全隐患数据提供了存储平台，但是由于缺少安全隐患分析模型，在数据的分析利用方面存在短板[21]，大量安全数据只是用于完成简单的问题处理、报表分析和数据统计，导致这些有价值的信息生命周期很短暂，在完成隐患排查后即以分散化、无序化的形式存储，成为历史数据，未能发挥这些数据对安全生产的指导作用，从而导致上述海量数据的浪费. 另外，矿山安全隐患数据的记录内容较短，每条数据的有效信息少，具有明显的短文本特征，所以选择适用于短文本挖掘的分析方法构建数据挖掘模型，从多角度探究隐患数据的内在联系，借助可视化手段对挖掘结果进行可视化展示，指导矿山安全隐患排查治理是当前矿山企业隐患治理中亟待解决的问题.

因此，本文在数据预处理的基础上对隐患信息进行多维度辨识，得到隐患在时间和空间两个维度上的分布规律；针对矿山隐患信息的短文本特征，采用双词主题模型（Biterm topic model, BTM）对安全隐患进行主题挖掘，得到了 13 个隐患主题，有效避免了潜在狄利克雷分配模型（Latent Dirichlet allocation, LDA）算法不适用于短文本挖掘的不足；最后通过 Apriori 算法对隐患数据进行了关联规则挖掘，得到了多条有效的关联规则，并对其进行了可视化展示.

1 基于大数据的安全隐患分析模型

构建安全隐患分析模型，首先对隐患数据进行预处理，然后基于大数据分析方对隐患信息进行多维度分析、主题挖掘、关联规则挖掘等，具体流程如图 1 所示.

1.1 数据预处理

由于矿山安全隐患数据记录的内容繁杂且在记录过程中缺乏规范性，因此为了保证文本挖掘的效果，在进行文本挖掘之前需要对数据进行清洗. 从矿山安全管理系统中导出的数据包含很多内容，比如责任人、责任单位等内容对文本挖掘不产生影响，因此将这些信息删除，仅保留数据中时间、地点、隐患问题部分，用以降低文本挖掘维度，提高文本挖掘处理的速度. 同时对记录中不规范格式及错别字进行纠正. 数据清洗完成后用 R 语言自带的 jiebaR 包对数据进行分词，分词过程可以理解为根据词库将文本分割成零碎的词汇，而这些词汇就是数据文本的特征项，由矿山安全隐患数据包含大量的采矿专业词汇，而这些专业

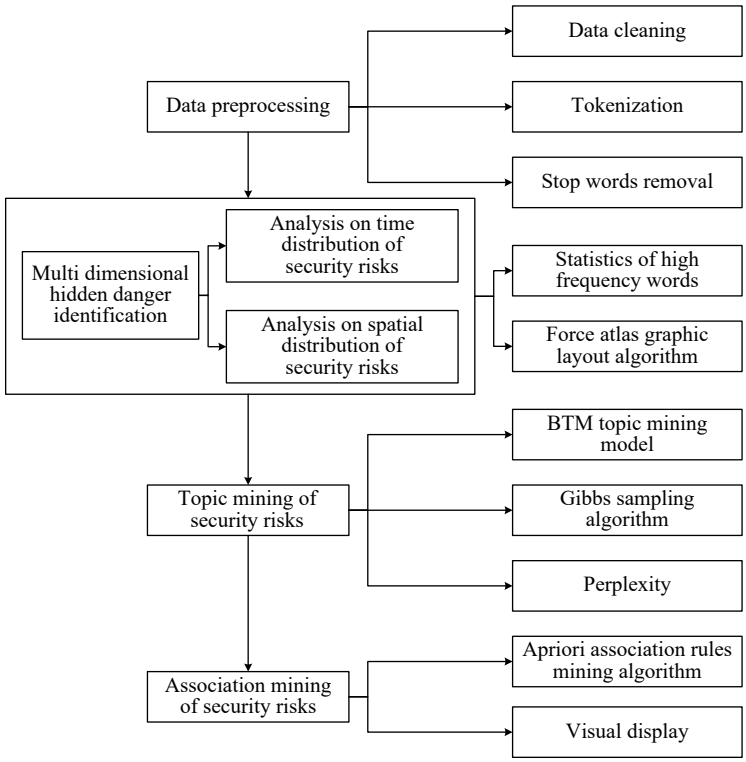

图 1　基于大数据分析的隐患分析模型流程

Fig.1　Hidden danger analysis model process based on big data analysis

词汇并不包含在 R 语言的词库中, 为了提高数据分词的准确性, 再分词前添加自定义词库, 词库内容来源于采矿工程、矿井通风、矿山工程、矿山应急救援等专业词汇. 分词结束后数据中依然存在频率较高但没有实际意义的词, 即停用词, 因此需要对分词结束的数据去停用词, 目的是对文本的特征项降维减噪, 提高文本挖掘工具的处理速度与处理能力.

1.2　隐患信息多维度辨识

地下矿山安全隐患的发生不是杂乱无序的, 随着企业对隐患的治理, 造成某些安全隐患数量随着时间的延长呈现出一定的波动起伏规律. 为了分析安全隐患的变化, 在数据预处理的基础上, 按照年份统计该年内出现频率较高的隐患词, 对高频词按时间顺序进行分析, 得到历年隐患数量随时间变化的曲线, 从而在时间维度上对隐患的发展变化进行把握. 同样, 隐患并不是在井下任何地点都存在, 一些特定的隐患会在某些关键地点出现且出现的次数较高, 为了明确井下隐患发生的重点区域, 指导企业对危险区域的排查, 通过统计隐患数据中的地点并提取频繁出现的地点, 对高频地点对应的隐患信息进行分析, 得到该地点可能发生的相关隐患.

为了更加直观的展示安全隐患在时间和空间两个维度上的分布规律, 利用大数据分析方法中的 Force Atlas 图形布局算法, 对隐患进行可视化表示, 得到安全隐患时间分布图和空间分布图.

1.3　隐患信息主题挖掘

矿山安全隐患数据的数量庞大而且所涉及的种类复杂多样, 在实际管理过程中很难通过人工完成对隐患数据按照隐患类别进行分类统计, 更难以发现安全隐患问题中隐藏的隐患主题. 因此, 通过大数据分析中的主题挖掘算法构建矿山隐患信息的主题挖掘模型, 对井下安全隐患数据进行深层次分析, 通过将众多的隐患归类, 获得能够反映井下生产安全问题的隐患主题, 更加有针对性的指导安全管理工作的开展.

文本的主题挖掘是大数据分析中重要的组成部分, 该方法可以将众多的数据按照一定的规则进行高度概括, 按照不同的隐患内容划分为不同的隐患主题. BTM 主题挖掘模型[22]与传统的 LDA 主题挖掘模型[23]的相似点在于, 两种主题算法的先验分布均服从狄利克雷分布 (Dirichlet distribution, Dir(α)), 区别在于 BTM 是对词对进行建模而不是单独的词语, 然后利用共轭分布对主题模型进行推理. 该模型通过对短文本语料进行词对扩充, 改善了短文本建模的稀疏问题. 该模型的概率模型如图 2 所示.

上图 2 中, Z 为一个主题, k 为维度, θ 为短文本集合中 k 个主题的分布, φ_k 为主题维度 k 的词汇

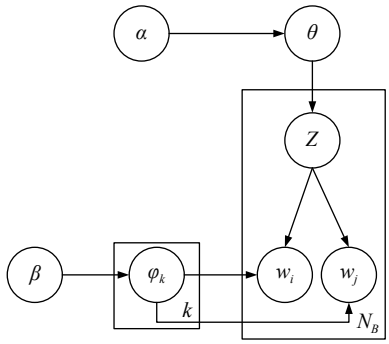

图 2 BTM 概率图模型

Fig.2 BTM probability graph model

分布，N_B 为文档数目，α、β 为词对与主题分布和主题与单词分布的狄利克雷分布的超参数. 模型的计算过程如下所示：

① 对于每一个主题 Z，其主题维度下的词分布为 φ=Dir(β)；

② 对于短文本语料集，确定一个全局的主题分布 θ=Dir(β)；

③ 对于词对 W<w_i, w_j>W 中每一个词，执行：

从全局主题分布 θ 中，抽取一个主题 Z，即 Z=Mult(θ)；

从被抽取的主题中，抽取两个词 w_i 和 w_j；其中每个词对都是从一个独立主题中产生，即 w_i, w_j=Mult(φ).

BTM 模型采用 Gibbs 抽样算法进行采样，主题分布的最终化简如下式所示.

$$P(Z|Z_{-w},W,\alpha,\beta)\propto(n_z+\alpha)\frac{(n_{w_i|z}+\beta)(n_{w_j|z}+\beta)}{(\sum wn_{w|z}+\beta M)^2} \quad (1)$$

式中：Z_{-w} 为除去当前词对的主题分布，Z 为主题，n_z 为分配到主题 Z 的次数，$n_{w|z}$、$n_{w_i|z}$、$n_{w_j|z}$分别为词对 w、w_i、w_j 分配到主题 Z 的次数，M 为语料集中不同的词语数.

之后根据公式 (2) 和公式 (3) 对超参数进行估计.

$$\theta=\frac{n_z+\beta}{B+K\alpha} \quad (2)$$

$$\varphi=\frac{n_{w|z}+\beta}{\sum_w n_{w|z}+M\beta} \quad (3)$$

1.4 隐患信息关联规则挖掘

导致矿山安全事故产生的原因往往不止一种，多种安全隐患的出现增加了矿山安全事故发生的概率，这表明安全隐患并不是孤立存在，他们往往存在着一定的联系. 因此分析不同隐患之间的内在联系，理清隐患之间的因果关系，对有效治理隐患，预防隐患产生起到事半功倍的作用.

Apriori 算法是挖掘布尔关联规则频繁项目集的经典算法之一[24−25]，该算法通过构建候选集和建立规则挖掘频繁项集，其核心是基于两阶段频集思想的递推算法. Apriori 算法对关联规则的挖掘主要分为两个步骤，首先要构建一组最小支持度的频繁项，然后根据所建立的频繁项集构造关联规则，具体步骤如图 3 所示.

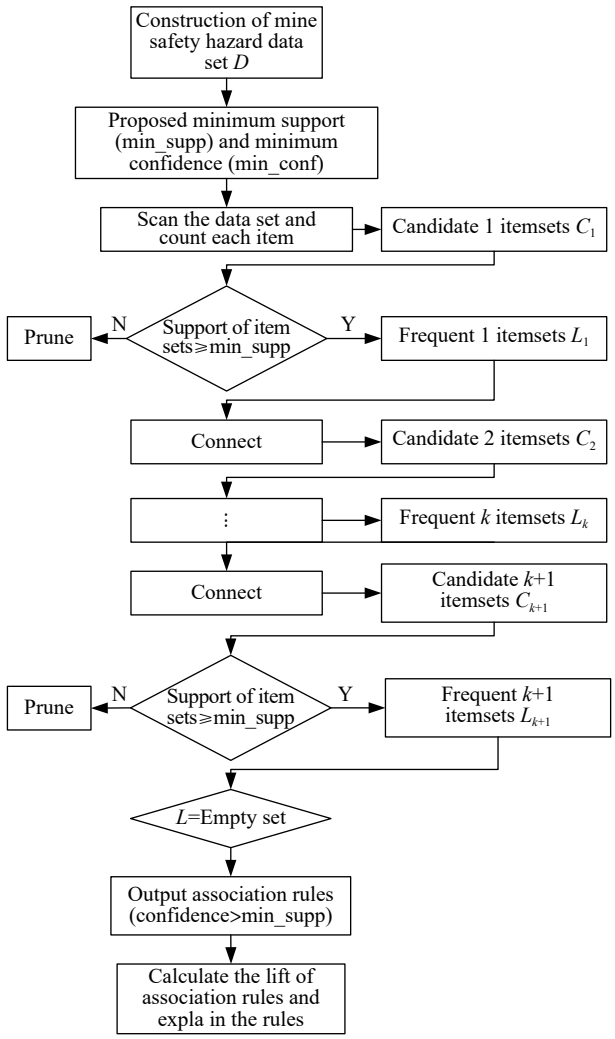

图 3 基于 Apriori 算法矿山安全隐患关联规则挖掘流程

Fig.3 Mining process of association rules for mine safety hazard based on Apriori algorithm

2 基于大数据隐患分析模型应用与结果分析

2.1 数据描述与处理

本文以某矿山的 GIS 安全管理系统中抽取得到安全隐患排查数据为研究对象，该数据从 2013 年至 2019 年，共计 34000 条，记录了隐患发生的时间、位置、隐患单位、具体问题、整改意见等内容.

对收集的隐患数据进行预处理，经过分词和去停用词后提取词频前 200 的高频词汇作为有效结果 (最小词频大于 200)，部分高频词如表 1 所示.

表 1　安全隐患高频词 (部分)

Table 1　High frequency words of hidden danger (part)

Number	Hidden danger vocabulary	Word frequency	Proportion/%	Number	Hidden danger vocabulary	Word frequency	Proportion/%
1	Support	9493	5.27	11	Civilized production	2440	1.36
2	Roof	9174	5.10	12	Pavement	2327	1.29
3	Pumice	8756	4.86	13	Roadway's sides	2232	1.24
4	Illumination	6145	3.41	14	Not in place	2190	1.22
5	Head-on	5237	2.91	15	Fan	2112	1.17
6	Much more	4931	2.74	16	Work	2099	1.17
7	Hydrops	2909	1.62	17	Distribution box	2011	1.12
8	Roof and sidewalls	2773	1.54	18	Fracture	1900	1.06
9	Facilities	2659	1.48	19	Explosive	1798	1.00
10	Rock bolt	2456	1.36	20	Jeep	1538	0.85

通过上表的词频统计可以清楚的看出在矿山安全隐患中支护的相关问题出现的频率最高, 其次是顶板和浮石问题. 为了更直观的看出隐患文本的分词结果, 对表 2 中的数据进行可视化展示, 通过 R 语言自带的词云展示工具对分词结果进行词云展示. 在词云展示过程中, 词汇的字体越大, 代表着该词在分词结果中出现的频率越高, 这样可以让读者对分析结果有快速、直观的理解[26]. 如图 4 所示, 在例如, 支护、顶板、浮石等词在分词结果中出现的频率较高, 因此在图中的字体大小较大.

表 2　不同年份共有隐患词汇统计表 (部分)

Table 2　Statistical table of common hidden danger vocabulary (part)

Hidden danger vocabulary	Word frequency						
	2013	2014	2015	2016	2017	2018	2019
Roof	605	872	818	1080	1358	1716	1246
Illumination	451	547	405	489	938	1176	1106
Rock bolt	161	220	360	259	321	322	235
Pumice	593	850	1014	1326	1317	1748	1234
Distribution box	176	226	156	237	333	391	303
Head-on	484	477	387	618	765	1242	794
Support	704	781	849	1152	1274	2116	1687
Fan	221	254	167	210	363	302	313
Hydrops	280	280	278	296	459	592	484
…	…	…	…	…	…	…	…

2.2　安全隐患时空分布规律分析

在数据预处理的基础上, 进行隐患信息的多维度辨识, 从时间和空间两个维度上对数据进行分析, 得到隐患在时间和地点上的分布规律. 具体维度划分如图 5 所示.

为了推测矿山安全隐患出现的趋势, 分析矿山安全隐患随时间变化的规律, 将安全隐患数据按照时间进行分类统计词频. 选取各年中相同的隐患词汇进行分析, 得到 30 种以上共有的隐患, 如表 2 所示为部分共有隐患.

图 4　矿山安全隐患词云图

Fig.4　Cloud chart of mine safety hidden danger

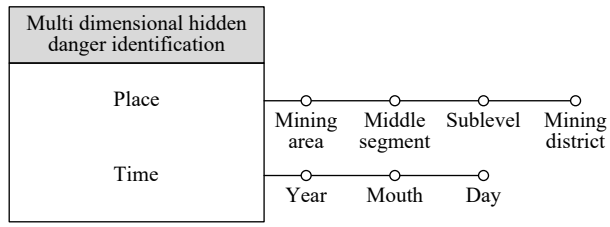

图 5　隐患数据维度分析图

Fig.5　Dimension analysis chart of hidden danger data

为了分析上述共有隐患词汇随时间的变化规律, 对上述隐患词汇绘制隐患发生频率随时间的变换曲线, 如图 6 所示.

由图 6 可以看出大部分隐患呈现出从 2013 年开始先增长, 到 2018 年达到最大值, 之后减少的趋势. 其中支护、浮石、顶板、迎头问题出远高于其他隐患, 且从 2013 ~ 2018 年有明显增加, 但 2018 年之后出现下降, 表明四种问题得到了一定程度的改善, 但整体出现频率依然很高, 表明依然是威胁员工井下生产的主要隐患, 需要矿山开采过程中重点关注. 配电箱、风机、安全背甲、漏电等隐

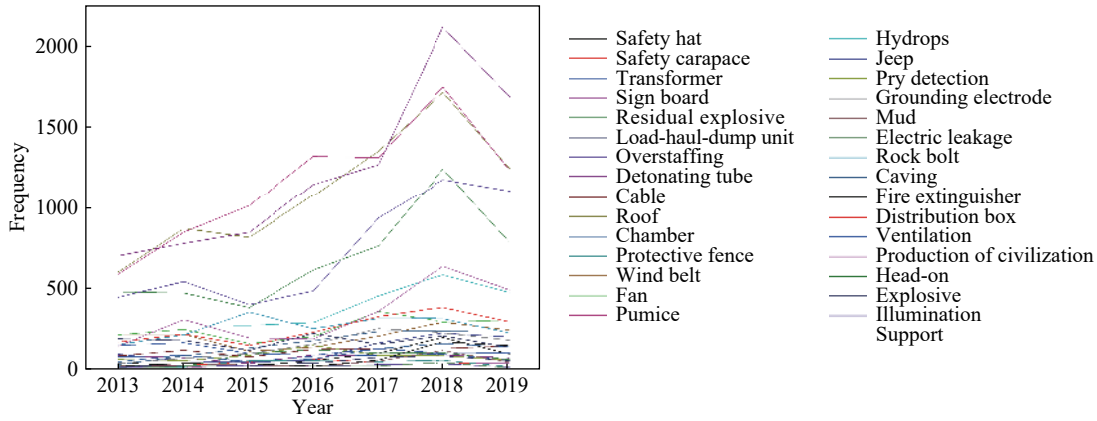

图 6 隐患−时间变化图

Fig.6 Hidden danger−time

患问题出现的频率相对较低，且随着时间变化表现为波动增长，但是增长幅度不明显，表明虽然在矿山开采过程中对上述三种隐患控制的比较好，但矿山企业仍需要加强对此类隐患进行的管理. 其他隐患数量相对较少，且变化不大，说明这些隐患在矿山生产中普遍存在，但对生产威胁较小.

为了更加直观显示隐患在时间维度上的分布，运用大数据分析方法中的 Force Atlas 图形布局算法绘制矿山安全隐患与时间规律分布图，如图 7 所示.

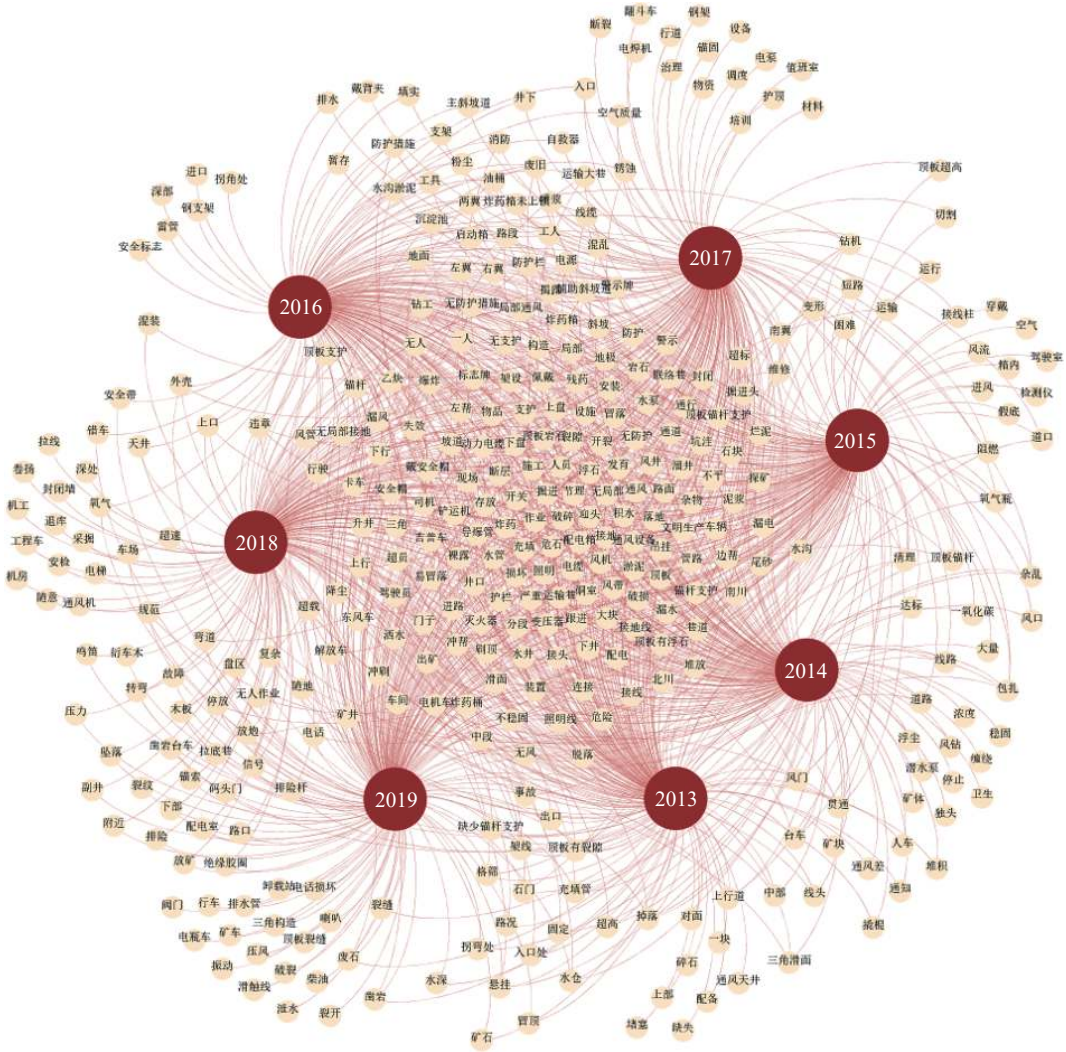

图 7 矿山安全隐患与时间规律分布

Fig.7 Hidden danger of mine safety and time distribution

由图 7 可以清晰的看出隐患随时间的分布情况, 中间部分的隐患词表示该隐患为 2013 ~ 2019 年间共有的隐患, 例如顶板问题、通风问题、照明问题等. 年份节点外部边缘部分隐患词表示该年份特有的隐患问题, 例如 2013 年电话损坏、顶板裂缝、排水管问题、卸载站等词出现频率较高、说明这些隐患在 2013 年出现角度且问题突出.

对隐患数据按照隐患发生地点进行统计, 提取隐患数量排名前 20 的隐患地点并分别进行分析和词频统计, 如表 3 所示, 选取各隐患地点排名前 100 的高频词, 利用 Force Atlas 图形布局算法绘制隐患与空间规律分布图, 如图 8 所示.

如表中所示, 隐患地点编码守卫代表矿区, 即 X 矿区和 S 矿区, 中间两位代表中段, 后三位代表勘探线, 取中间勘探线 (三位数), 不足三位前面补 0. 由表可以看出高频隐患地点中有 7 个属于 S 矿区, 9 个属于 X 矿区.

表 3　高频隐患地点统计表 (前 20)

Table 3　Statistical table of high frequency hidden danger location (top 20)

Hidden danger location	Quantity	Hidden danger location	Number
Slope mouth	509	S13155	149
S12186	254	X06111	144
S14186	239	X08059	141
X07097	228	S18156	140
X07087	226	X08055	132
X07105	225	X05103	123
S13186	202	S15186	122
Assistant ramp	197	S10167	115
X09105	170	X05111	108
Main ramp	164	West ventilating shaft	105

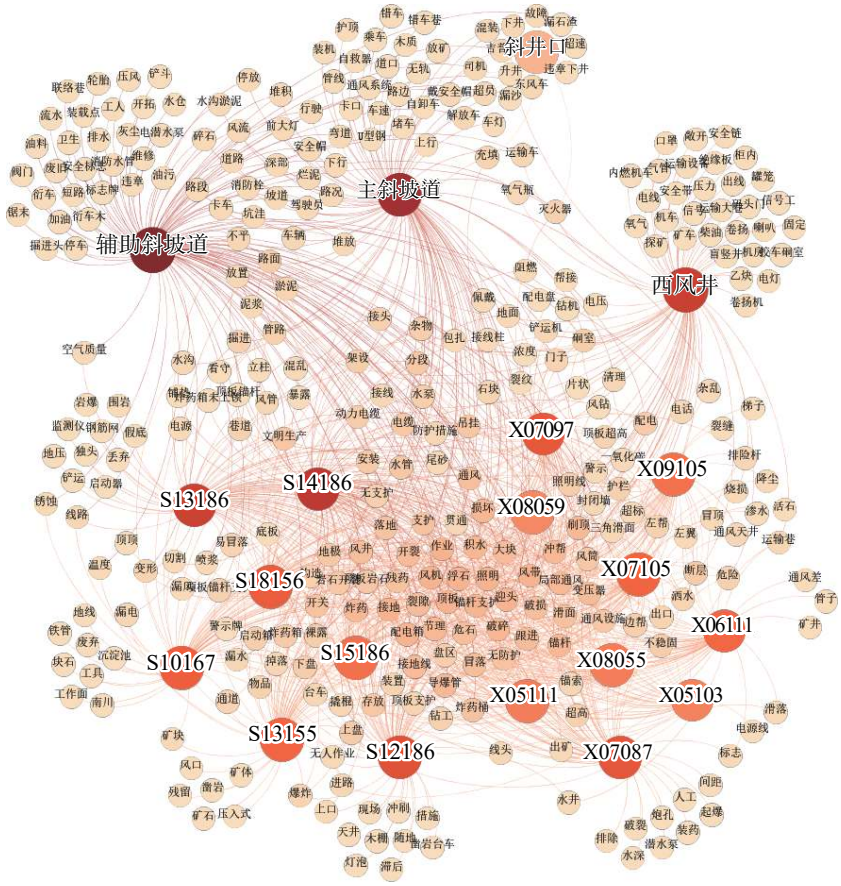

图 8　矿山安全隐患与空间规律分布

Fig.8　Hidden danger of mine safety and its spatial distribution

从上图可以看出, 隐患地点构成了四个主要的群, 其中两个较为突出的地点群分别有 S 矿区的 7 个采场区域和 X 矿区的 9 个采场区域构成, 另外两个较小的地点群分别由辅助斜坡道、主斜坡道和西风井构成. 图中隐患词在地点群中交叉的部分为各区域共有隐患, 没有交叉部分的隐患为各区域特有的隐患. 地点节点对应的颜色越深代表该地点隐患问题越多. 例如, S13186 地点岩爆、围岩、地压等隐患词出现频繁, 说明该地点采场地压问题严重, 因此为保证井下作业安全, 应及时对井下进行支护.

2.3　隐患数据主题挖掘结果分析

利用 BTM 主题挖掘模型对矿山井下安全隐

患数据进行分析，首先利用困惑度对最优主题数目进行判断，困惑度越小，主题挖掘的质量越好，相反越差．将主题范围设置为 2 ~ 100，步长设置为 5，迭代次数 1000 次，得到困惑度与主题数目的关系图，如图 9 所示．

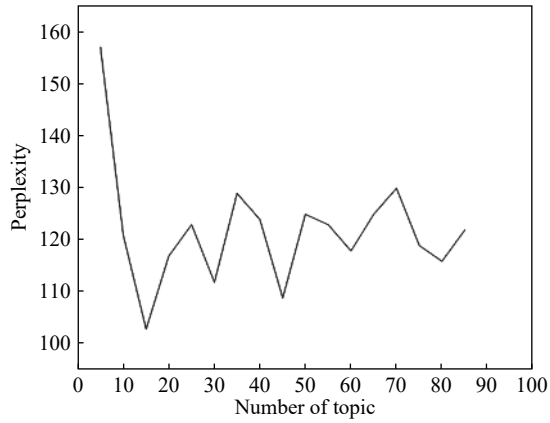

图 9 困惑度−主题数目关系图

Fig.9 Perplexity−topic number graph

由图 9 可以观察到随着主题数目的增加，困惑度表现出明显的波动，在主题数目为 15 时困惑度最低，表明此时为最佳主题数目．为了保证对隐患主题分类的准确性和标准性，参考张勇[27]对生产安全事故隐患分类的方法和《安全生产事故隐患排查治理暂行规定》[28]、《金属非金属矿山重大生产安全事故隐患判定标准 (试行)》[29]，在对各个主题词归纳统计的基础上对主题进行命名，并将相近的主题合并，最终得到 13 个隐患主题．为了对各个隐患主题有直观的了解，确定主题数目之后，对隐患文本做进一步分析，针对每个隐患主题提取对应的特征词，剔除隐患主题中的噪声词后即可得到各个主题对应的隐患主题词，如表 4 所示．

表 4 BTM 矿山安全隐患主题与隐患主题词表

Table 4 BTM mine safety hidden danger theme and hidden danger keywords list

Number	Safety hidden danger theme	Hidden danger keywords
1	Hidden danger of support	Support, roof, roadway's sides, network degree, measures, not in place, invalid, fracture
2	Hidden danger of roof	Roof, joint, caving, fragment, pumice, dangerous rock, crack, development
3	Hidden danger of transport	Overload, ramp, violation, jeep, down, fire extinguisher, load-haul-dump unit
4	Hidden danger of rock bolt	Rock bolt, network degree, not in time, follow-up, lack, long- cable, too long
5	Hidden danger of pipeline	Wind belt, cable, set up, follow-up, damaged, hang, stringing, drop, water pipe
6	Hidden danger of ventilation and three prevention	Fire extinguisher, fire water pipe, fire box, dust, airflow, oxygen, air quality, local ventilation
7	Hidden danger of operation	Operation, grouting, excavation, scene, top brush, people, construction, not completely
8	Hidden danger of safety protection	Safety hat, protect, protective fence, sign, carapace, measures, sign
9	Hidden danger of electromechanical	Fan, distribution box, transformer, switch, ground wire, grounding electrode, cable
10	Hidden danger of blasting operation	Smooth blasting, explosive, detonating tube, explosive box, lock, lying around
11	Hidden danger of road	Pavement, out-of-flatness, silt, potholes, sundries, hydrops
12	Hidden danger of water disaster	Hydrops, too much, deeper, ditch, water pump, puddles, drain
13	Hidden danger of environmental	Silt, mud, clean up, poor, hydrops, sundries, purling

通过表 4 可以清楚地了解到每个隐患主题包含的主要内容．例如，支护隐患主题中主要包含了支护、顶帮、两帮、网度、不到位等隐患主题词，这些主题词既描述了支护隐患容易发生的地点，同时也描述了该隐患的主要表现形式，能够为矿山井下安全检查和隐患排查工作提供必要的指导，使现场安全检查过程中更具精确性和针对性，可以迅速捕捉生产中的隐患，从而提高对安全隐患危险源的排查效率和隐患治理效率．

2.4 隐患数据关联规则挖掘分析

利用 R 语言中“arules”工具包对矿山隐患文本进行关联规则挖掘，R 语言中默认的最小支持度与最小置信度分别为 0.1 和 0.8，该支持度对于本文分析相对过小，导致得到的规则过于侧重顶板、浮石、支护，因此本文设置最小支持度和最小置信度分别为 0.005 和 0.5，执行算法后得到了 296 条关联规则，由于过低的提升度不具备现实意义，因此将关联规则按照提升度进行排序，去除提升度小于 3 规则后作为最终的有效规则，最终得到了超过 237 条有效关联规则，表 5 中列举了有效关联规则中典型的 10 条规则．

从表 5 可以看出隐患数据之间存在一定的联系，通过文本挖掘得到的关联规则能够切实提高井下安全隐患检查工作的效率．例如：井下从事运输工作的司机更多的安全隐患是不佩戴或者不正确佩戴安全帽，该类安全隐患占全部隐患数据的 0.504%，参照该关联规则在对井下四级进行隐患排查过程中与随机排查相比效率可以提高 50 倍

表 5　矿山安全隐患关联规则挖掘 (部分)

Table 5　Mining association rules of mine hidden danger (part)

Number	Association rules	Support	Confidence	Lift	Count
1	{driver}=> {safety hat}	0.0050427	0.7208333	51.200060	173
2	{pry detection}=> {top brush}	0.0158860	0.9663121	49.332244	545
3	{pavement}=> {potholes, uneven}	0.0124173	0.9487751	41.891410	426
4	{network degree} =>{bigger}	0.0123299	0.9276316	22.348495	423
5	{roof and sidewalls, head-on}=>{pumice}	0.0139039	0.9173077	4.151725	477
6	{roadway's sides, illumination, facility}=>{pumice}	0.0102603	0.9048843	4.095497	352
7	{lying around}=> {explosive}	0.0091235	0.8743017	22.317461	313
8	{landing}=>{fan}	0.0063544	0.5561224	11.322785	218
9	{pumice, ventilation facilities} =>{illumination}	0.0061503	0.5926966	4.800199	211
10	{residual explosive}=>{roof}	0.0123590	0.7054908	3.445306	424

以上; 矿山生产过程中炸药的使用具有较大安全隐患, 也是企业重点关注的对象, 通过对隐患数据的挖掘得出了在对炸药的处理过程中经常出现不按规定放置的现象, 比如炸药裸放. 这种隐患占比达到了隐患总数的 0.9%, 根据该规则对炸药隐患进行检查能够提高隐患排查效率 20 倍以上.

为了更加直观的分析得到的关联规则, 利用 R 语言中的 arulesViz 包对关联规则进行可视化展示, 如下图 10 ~ 图 11 所示.

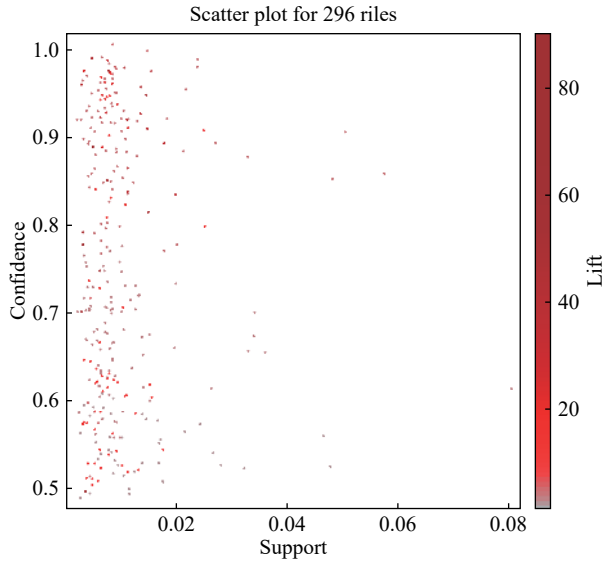

图 10　矿山井下安全隐患关联规则散点图

Fig.10　Scatter diagram of association rules for underground safety hazards

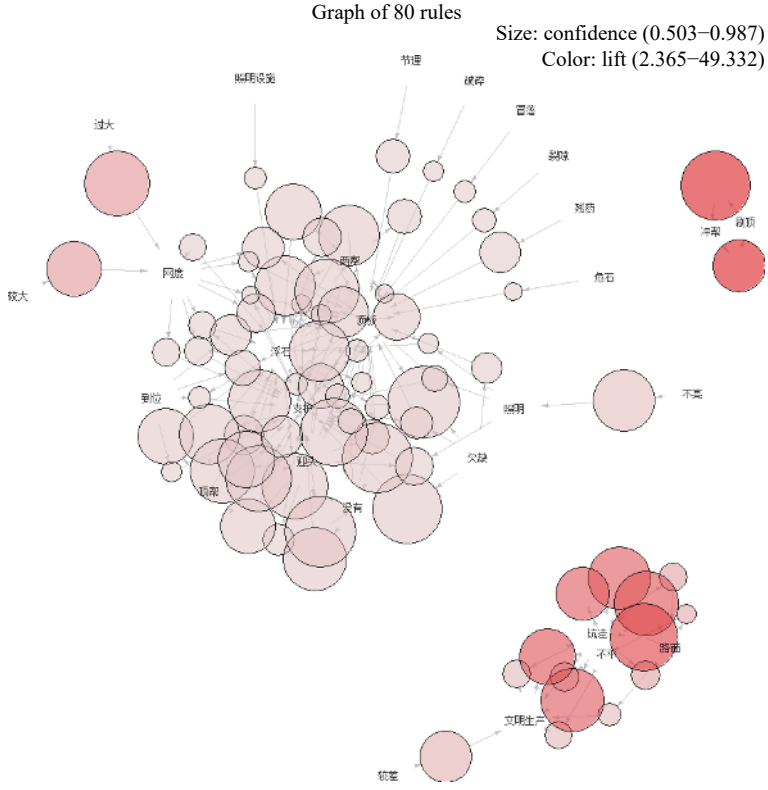

图 11　基于图的矿山安全隐患关联规则可视化

Fig.11　Visualization of mine hidden danger association rules based on graph

从图 10 可以看出通过文本挖掘得到的 296 条关联规则的支持度主要分布在 0 ~ 0.02 之间, 置信度从 0.5 ~ 1.0 呈现均匀分布, 图中每个点的颜色代表了不同的提升度, 颜色越深表示提升度越高. 从

图中可以看出大部分散点颜色呈现红色，表明大部分规则的提升度较高，通过这些规则可以提高隐患的排查效果.

图 11 中展示了按照支持度排序的前 80 条关联规则，图中圆圈大小代表了置信度，颜色代表提升度，从图中可以看出上述规则主要分为了两个集合，一个主要以顶板、浮石、支护等隐患为主，另一个以路面相关隐患为主. 通过上述关联规则的可视化展示可以使矿山工作人员更加直观清楚的对矿山安全隐患情况有所了解，为隐患治理提供可视化的决策支持.

3 结论

本文针对矿山具有短文本特性的安全隐患数据开展研究，从不同角度出发建立分析模型对非结构化的隐患数据进行挖掘，首先通过构建多维度分析模型分析了安全隐患随时间和空间维度的变化分布规律. 之后针对短文本特点通过 BTM 算法构建主题挖掘模型，通过分析得到了 13 个隐患主题，避免了传统 LDA 算法不适用于短文本建模的不足. 最后通过 Apriori 算法建立了关联规则挖掘模型，对数据进行分析，得到了隐患之间的内在联系. 通过研究表明矿山安全隐患数据之间存在一定内在联系，基于文本挖掘的大数据分析技术是解决文本数据不易分析的可行途径，且本文所构建的隐患分析模型适用于具有短文本特性的非结构化隐患数据的挖掘分析，所得到的结果可为矿山企业治理排除井下安全隐患提供指导.

参 考 文 献

[1] Bi H T, Li G Q. Research on the closed circuit management system of safety production in gold mines. *Gold*, 2014, 35(8): 1
(毕洪涛, 李国清. 黄金矿山安全生产闭环管理体系研究. 黄金, 2014, 35(8): 1)

[2] Meng X F, Li K Y, Liu F. Study on coal mine safety risk pre-control based on three-levels nested management mode. *China Saf Sci J*, 2013, 23(4): 102
(孟现飞, 李克业, 刘飞. 基于3级嵌套安全管理模式的煤矿安全风险预控研究. 中国安全科学学报, 2013, 23(4): 102)

[3] Zhao D F, Shen Y Q, Zhao Z Q, et al. Risk classification method for accident potential based on development and control measures of accident. *China Saf Sci J*, 2012, 22(4): 71
(赵东风, 申玉琪, 赵志强, 等. 基于事故发展与控制的隐患分级方法. 中国安全科学学报, 2012, 22(4): 71)

[4] Martin E B, Morris A J. Non-parametric confidence bounds for process performance monitoring charts. *J Process Control*, 1996, 6(6): 349

[5] Martin E B, Morris A J. An overview of multivariate statistical process control in continuous and batch process performance monitoring. *Trans Inst Meas Control*, 1996, 18(1): 51

[6] Dunia R, Qin S J, Edgar T F. Identification of faulty sensors using principal component analysis. *AIChE J*, 1996, 42(10): 2797

[7] Li J, Cui B, Wang Q, et al. Information system for identification and control of major coal mine disasters // *Innovative coal enterprise development and informatization Summit Forum*. Lanzhou, 2010: 56
(李季, 翟勃, 汪群, 等. 煤矿重大灾害辩识和控制信息系统 // 创新型煤炭企业发展与信息化高峰论坛论文集. 兰州, 2010: 56)

[8] Qin W J. *The Research and Application of the Three-Dimensional Method in Coalmine Gas Explosion Hazard Identifiction* [Dissertation]. Taiyuan: Taiyuan University of Technology, 2015
(秦文静. 煤矿瓦斯爆炸危险源三维辨识研究及应用[学位论文]. 太原: 太原理工大学, 2015)

[9] Zhang B L, Wang X Q, He Y R, et al. Construction of safety hazard identification and investigation system of coal mine based on ontology. *Saf Coal Mines*, 2018, 49(5): 239
(张宝隆, 王向前, 何叶荣, 等. 基于本体的煤矿事故隐患辨识排查系统构建. 煤矿安全, 2018, 49(5): 239)

[10] Ma X P, Dai W. Research status and application prospect of big data technology in coal industry. *Ind Mine Autom*, 2018, 44(1): 50
(马小平, 代伟. 大数据技术在煤炭工业中的研究现状与应用展望. 工矿自动化, 2018, 44(1): 50)

[11] Sun J P. Accident analysis and big data and Internet of Things in coal mine. *Ind Mine Autom*, 2015, 41(3): 1
(孙继平. 煤矿事故分析与煤矿大数据和物联网. 工矿自动化, 2015, 41(3): 1)

[12] Tan Z L, Chen X, Song Q Z, et al. Analysis for the potential hazardous risks of the coal mines based on the socalled text mining. *J Saf Environ*, 2017, 17(4): 1262
(谭章禄, 陈晓, 宋庆正, 等. 基于文本挖掘的煤矿安全隐患分析. 安全与环境学报, 2017, 17(4): 1262)

[13] Qian Y H. Application of data mining algorithm in gas safety prediction. *Coal Technol*, 2018, 37(5): 207
(钱宇虹. 数据挖掘算法在瓦斯安全预测中的应用. 煤炭技术, 2018, 37(5): 207)

[14] Shi J B, Shi J H. Application of correlation analysis on data mining in coal mine gas safety monitoring and early warning. *China Energy Environ Prot*, 2017, 39(8): 1
(石记斌, 石记红. 关联分析数据挖掘在煤矿瓦斯安全监测预警中的应用. 能源与环保, 2017, 39(8): 1)

[15] Lei Y B, Chen Z B, Zeng J C, et al. Research on causal chain of coal mine gas accidents based on association rule. *Saf Coal Mines*, 2016, 47(8): 240
(雷煜斌, 陈兆波, 曾建潮, 等. 基于关联规则的煤矿瓦斯事故致因链研究. 煤矿安全, 2016, 47(8): 240)

[16] Liu H B, Li C H. Research on occupational health and safety management information system in intelligent mine. *Coal Sci Technol*, 2019, 47(3): 87

(刘海滨, 李春贺. 智慧矿山职业健康安全监管信息系统研究. 煤炭科学技术, 2019, 47(3): 87)

[17] Henriques V, Malekian R. Mine safety system using wireless sensor network. *IEEE Access*, 2016, 4: 3511

[18] Zhou X T. Mine safety monitoring, early warning and integrated management information system. *World Nonferrous Met*, 2017(19): 26
(周雪田. 矿山安全监测预警与综合管理信息系统. 世界有色金属, 2017(19): 26)

[19] Zhang D W. Analysis of coal mine safety hidden danger trends based on OLAM. *Coal Eng*, 2015, 47(5): 139
(张大伟. 基于OLAM的煤矿企业安全隐患趋势分析. 煤炭工程, 2015, 47(5): 139)

[20] Yao Q G, Zhao L X, Zhang X M. Fuzzy comprehensive evaluation model of coal mine safety management information system. *Min Saf Environ Prot*, 2017, 44(6): 120
(姚庆国, 赵丽霞, 张学睦. 煤矿安全管理信息系统模糊综合评价模型. 矿业安全与环保, 2017, 44(6): 120)

[21] Cheng J W, Yang S Q. Data mining applications in evaluating mine ventilation system. *Saf Sci*, 2012, 50(4): 918

[22] Tang X B, Xiang K. Hotspot mining based on LDA model and microblog heat. *Libr Inf Serv*, 2014, 58(5): 58
(唐晓波, 向坤. 基于LDA模型和微博热度的热点挖掘. 图书情报工作, 2014, 58(5): 58)

[23] Blei D M, Ng A Y, Jordan M I, Lafferty J. Latent Dirichlet Allocation. *J Mach Learn Res*, 2003, 3: 993

[24] Ma T B, Xu F. Research and improvement for apriori algorithm of association rule mining. *J Lanzhou Polytech Coll*, 2010, 17(1): 13
(马廷斌, 徐芬. 关联规则挖掘中Apriori算法的研究与改进. 兰州工业高等专科学校学报, 2010, 17(1): 13)

[25] Cui Y, Bao Z Q. Survey of association rule mining. *Appl Res Comput*, 2016, 33(2): 330
(崔妍, 包志强. 关联规则挖掘综述. 计算机应用研究, 2016, 33(2): 330)

[26] Krause S, Busch F. New insights into road accident analysis through the use of text mining methods // 2019 6*th International Conference on Models and Technologies for Intelligent Transportation Systems* (*MT-ITS*). Cracow, 2019: 1

[27] Zhang Y. Establishment and application of hidden danger automatic investigation and automatic report system to coal mine safety production accident. *Coal Eng*, 2014, 46(11): 150
(张勇. 煤矿安全生产事故隐患自查自报系统建立与应用. 煤炭工程, 2014, 46(11): 150)

[28] Ministry of Emergency Management of People's Republic of China. Interim Provisions on the Investigation and Treatment of Hidden Dangers of Work Safety [EB/OL]. *Ministry of Emergency Management of the People's Republic of China* (2008-01-10) [2020-10-23]. http://www.mem.gov.cn/gk/gwgg/zjl_01/200801/t20080110_233738.shtml
(中华人民共和国应急管理部. 安全生产事故隐患排查治理暂行规定[EB/OL]. 中华人民共和国应急管理部 (2008-01-10) [2020-10-23]. http://www.mem.gov.cn/gk/gwgg/zjl_01/200801/t20080110_233738.shtml)

[29] Ministry of Emergency Management of People's Republic of China. Judgment Standards for Hidden Dangers of Major Production Safety Accidents in Metal and Nonmetal Mine (Trial) [EB/OL]. *Ministry of Emergency Management of People's Republic of China* (2017-09-05) [2020-10-23].https://www.mem.gov.cn/gk/gwgg/201709/t20170905_241758.shtml
(中华人民共和国应急管理部. 金属非金属矿山重大生产安全事故隐患判定标准(试行)[EB/OL]. 中华人民共和国应急管理部 (2017-09-05) [2020-10-23].https://www.mem.gov.cn/gk/gwgg/201709/t20170905_241758.shtml)