

文章编号:1671-251X(2019)09-0075-05

DOI:10.13272/j.issn.1671-251x.2019030021

煤矿事故本体自动构建

桂冬冬¹, 王向前^{1,2}, 李慧宗¹

(1. 安徽理工大学 经济与管理学院, 安徽 淮南 232001;
2. 复旦大学 管理科学与工程博士后流动站, 上海 200433)



扫码移动阅读

摘要:针对通过人工方式构建煤矿事故本体存在耗时耗力、具有局限性等问题,提出了一种煤矿事故本体自动构建方法。首先,以煤矿事故案例文本作为数据源,采用BP神经网络自动提取本体概念。然后,通过层次聚类法和关联规则法分别自动提取本体概念间层次关系、非层次关系。最后,利用Protégé本体编辑器对煤矿事故本体概念、概念间关系和实例进行可视化表达,得到煤矿事故本体。自动构建的煤矿事故本体更加全面,可为煤矿安全生产工作提供有力支持。

关键词:煤矿事故; 本体; 概念提取; 关系提取; 层次聚类; 关联规则
中图分类号:TD67 **文献标志码:**A

Automatic construction of coal mine accident ontology

GUI Dongdong¹, WANG Xiangqian^{1,2}, LI Huizong¹

(1. School of Economics and Management, Anhui University of Science and Technology, Huainan 232001, China; 2. Post-doctoral Station of Management Science and Engineering, Fudan University, Shanghai 200433, China)

Abstract: Aiming at problems of time-consuming, labor-consuming and limitation in constructing coal mine accident ontology by manual means, an automatic construction method of coal mine accident ontology was proposed. Firstly, taking text of coal mine accident case as data source, BP neural network is used to automatically extract ontology concept. Then, hierarchical relationship and non-hierarchical relationship between ontology concepts are automatically extracted by adopting hierarchical clustering method and association rules method. Finally, ontology concepts, relationship between concepts and examples of coal mine accident are visualized by use of Protégé ontology editor, so as to obtain coal mine accident ontology. The coal mine accident ontology by automatic construction is more comprehensive and can provide strong support for coal mine safety production work.

Key words: coal mine accident; ontology; concept extract; relationship extract; hierarchical clustering; association rules

0 引言

作为世界范围内的产煤大国,我国煤炭开采产量一直处于较高水平,但同时我国煤矿事故发生率和人员死亡率也是最高的。煤矿事故发生后,如何在这些事故案例中寻找或发现特定的、有用的信息

来应用于以后的安全预警或救援行动中,其困难和复杂程度相对较高。本体作为一种可共享的、明确的、规范化的说明^[1],能够很好地实现领域知识的共享和重用。目前已有部分学者对煤矿事故本体进行了研究。刘婷^[2]通过FCA(Formal Concept Analysis,形式概念分析)技术构建采煤工作面动态

收稿日期:2019-03-08;修回日期:2019-08-14;责任编辑:盛男。
基金项目:国家自然科学基金资助项目(51474007);安徽省自然科学基金面上项目(1808085MG221)。
作者简介:桂冬冬(1994—),男,安徽安庆人,硕士研究生,主要研究方向为本体工程、管理信息系统,E-mail:943870725@qq.com。
引用格式:桂冬冬,王向前,李慧宗. 煤矿事故本体自动构建[J]. 工矿自动化,2019,45(9):75-79.
GUI Dongdong, WANG Xiangqian, LI Huizong. Automatic construction of coal mine accident ontology[J]. Industry and Mine Automation, 2019, 45(9): 75-79.

本体并结合 Jena 推理机进行推理,可及时发现煤矿井下不安全因素并采取措施,保障煤矿生产安全。郭晓黎等^[3]从煤矿事故发生前、发生中、发生后 3 个方面构建了煤矿安全事件本体模型并添加相应实例,为煤矿事故应急处理提供支持。桂红军^[4]构建了一种基于本体的煤矿案例库并应用于煤矿应急救援决策系统,为制定煤矿事故救援方案提供参考。孟现飞^[5]基于本体构建了煤矿事故预警知识库模型,应用结果表明该模型可有效提高安全预警准确率。药慧婷等^[6]构建了掘进工作面安全生产本体模型,可为智慧矿山的智能决策提供帮助。但上述本体在构建时,本体概念及概念间关系大多依靠人工进行提取,耗费大量人力、时间,同时受人的主观性影响,所构建本体具有局限性。鉴此,本文提出了一种煤矿事故本体自动构建方法,采用 BP 神经网络对煤矿事故本体概念进行自动提取,并在此基础上对概念间关系进行自动提取,从而大大减少煤矿事故本体构建的工作量,且所构建本体更全面,可为煤矿安全生产提供有力支持。

1 煤矿事故本体构建流程

煤矿事故本体构建流程如图 1 所示,具体步骤如下。

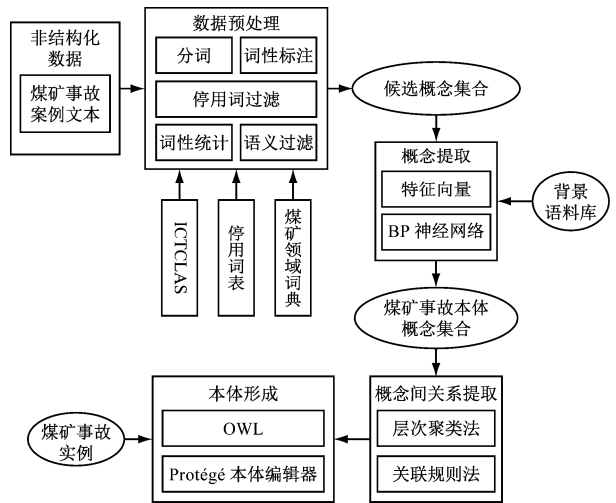


图 1 煤矿事故本体构建流程

Fig. 1 Process of construction of coal mine accident ontology

(1) 搜集煤矿事故案例文本这样的非结构化数据作为数据源,相比于结构化数据,可获得更多的概念和关系,使所构建煤矿事故本体更加全面、完整。由于大量文本杂乱无章、难以操作,需对数据进行预处理,利用 ICTCLAS 中文分词系统并结合停用词表和煤矿领域词典对数据进行分词、词性标注、停用词过滤、词性统计、语义过滤等一系列操作,得到候选概念集合。

(2) 通常利用特征来描述概念,但单一的特征通常对候选概念的区分度不高,为提高区分效果,添加背景语料库来获得概念的特征,并采用多个特征组成特征向量来描述概念。构建 BP 神经网络并以特征向量作为神经网络的输入,对神经网络进行训练,完成对煤矿事故本体的概念提取。

(3) 在获得煤矿事故本体概念集合后,将每一个概念视为一个类,采用层次聚类法^[7]并依据平均相似度进行迭代,得到概念间层次关系。利用关联规则法^[8]提取概念间非层次关系,从而完成本体概念间关系自动提取。

(4) 利用 Protégé 本体编辑器^[9]和 OWL(Web Ontology Language,网络本体语言)对本体的概念、概念间关系和实例进行可视化表达,得到煤矿事故本体。

2 煤矿事故本体概念自动提取

2.1 数据预处理

2.1.1 文本分词

ICTCLAS 是由中科院研发的汉语词法分析系统,其功能主要包括中文分词、词性标注、命名实体识别等^[10],且支持导入自定义词典,分词准确率较高,因此本文采用 ICTCLAS 对文本进行分词处理。但针对某个特定领域的概念时,ICTCLAS 有时会将一个概念拆分为多个词,从而影响分词效果,例如“工作面”可能被分为“工作/面”,“瓦斯涌出”被分为“瓦斯/涌出”,“瓦斯爆炸事故”被分为“瓦斯/爆炸/事故”等。因此根据煤矿事故领域词汇的特征,通过人工获取常用的煤矿领域词汇来构建自定义煤矿领域词典并导入 ICTCLAS 中,提高文本分词效果。

2.1.2 停用词过滤

经过分词处理后的文本中包含很多与煤矿事故领域无关的高频词汇,例如连词“然后”、“接着”,副词“一直”、“非常”等,还有大量标点符号。这些连词、副词、标点符号等停用词没有实际含义,增加了概念提取的工作量。因此选取哈尔滨工业大学停用词表和百度停用词表对分词处理后文本中所出现的停用词进行过滤,提高概念提取的精度。

2.2 基于 BP 神经网络的煤矿事故本体概念提取

BP 神经网络由多个非线性函数组成,具有高度的非线性运算能力和容错能力,本质上是一种黑箱建模方法,不必对事物的机理进行准确、详细的数学描述,常用于函数拟合、预测和分类等问题^[11-12]。对煤矿事故本体概念进行提取实际上就是概念分类问题,因此可通过 BP 神经网络进行模式识别,从而自动提取本体概念。

2.2.1 煤矿事故本体概念特征提取

在经过数据预处理得到候选概念集合后,需要定义特征来表示煤矿事故本体概念与非煤矿事故本体概念的差异。特征是对一个客体或一组客体特性进行抽象的结果,用来描述概念。为更好地区分煤矿事故本体概念与非煤矿事故本体概念,需要添加背景语料库来进行概念特征提取^[13],本文从艺术、教育、军事、体育、健康、金融 6 个领域选取 100 篇文本组成背景语料库。煤矿事故本体概念有众多特性,根据煤矿事故本体概念共有的特性抽象出某一特征,该特征就可用来区分是否为煤矿事故本体概念。同时,将这些特征通过向量的形式作为 BP 神经网络的输入,从而完成神经网络训练和概念分类。本文选用 TF (Term Frequency, 词频)、DF (Document Frequency, 文档频率)、IDF (Inverse Document Frequency, 逆文档频率)这 3 种特征组成特征向量^[14],部分候选概念的特征向量经归一化处理后结果见表 1。

表 1 部分候选概念的特征向量

候选概念	TF	IDF	DF
煤矿事故	1.000 000	0.694 345	1.0
瓦斯爆炸	0.132 742	0.792 343	0.5
有害气体	0.071 345	0.697 653	0.2
巷道	0.086 493	0.586 341	0.3
顶板事故	0.124 202	0.732 813	0.6
停风	0.056 443	0.485 233	0.1

2.2.2 煤矿事故本体概念提取

设训练样本集 $O = \{(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k)\}$, 其中 $x_i (i=1, 2, \dots, k, k \text{ 为样本数})$ 为第 i 个候选概念的特征, $x_i \in \mathbf{R}^m, m$ 为特征个数, y_i 为第 i 个候选概念所属分类, $y_i \in \{0, 1\}$, 设“0”表示非煤矿事故本体概念,“1”表示煤矿事故本体概念。因此煤矿事故本体概念的提取转换为寻找一个函数,满足当 x_i 为煤矿事故本体概念特征时 $y_i=1$, 当 x_i 为非煤矿事故本体概念特征时 $y_i=0$ 。

首先对候选概念进行人工标记,将煤矿事故本体概念即正样本标记为(1,0),非煤矿事故本体概念即负样本标记为(0,1)。然后将样本的特征向量作为神经网络的输入,标记后的样本向量作为神经网络的输出,进行 BP 神经网络的学习,学习特征向量与标记样本之间的关系。在 BP 神经网络的学习过程中不断调整网络的权值和阈值,将训练得到的神经网络模型用于概念提取,最后得到煤矿事故本体概念集合。煤矿事故本体概念提取流程如图 2

所示。

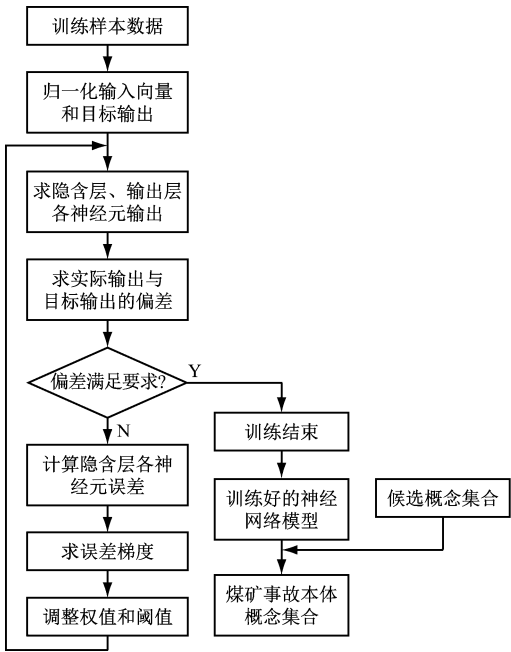


图 2 煤矿事故本体概念提取流程

Fig. 2 Process of concept extraction of coal mine accident ontology

3 煤矿事故本体概念间关系自动提取

在得到煤矿事故本体概念集合后,需要判断所提取的概念与概念之间是否存在关系及存在何种关系。本体概念间关系通常分为层次关系和非层次关系(也称分类关系和非分类关系)^[15];层次关系表示概念间的父子关系或上下位关系;非层次关系表示除层次关系之外的其他关系,主要包括整体与部分关系、属性关系等。

3.1 煤矿事故本体概念间层次关系提取

目前,研究者大多采用基于语言学的方法来提取本体概念间层次关系。但由于汉语语句组成形式的多样性,基于语言学的本体概念间层次关系提取没有统一的方法,过程繁琐。聚类方法是一种无监督学习方法,它将集合中的元素按照相似度组成多个类,不需要特定的汉语句式,操作方法简单且统一。因此采用层次聚类法对煤矿事故本体概念间层次关系进行自动提取,具体实施步骤如下。

(1) 将煤矿事故本体概念集合中各概念视为一个单独的类,使用余弦距离法计算各概念间的语义相似度。

(2) 选取相似度最大(距离最小)的 2 个类 E, F , 将 E, F 聚为一类形成一个新的类。重复聚类操作,直至所有类之间的相似度小于事先设置的阈值。2 个类 E, F 的相似度为

$$\text{sim}(E, F) = \frac{\sum_{C_e \in E, C_f \in F} \text{sim}(C_e, C_f)}{pq} \quad (1)$$

式中 $C_e, C_f (1 \leq e \leq p, 1 \leq f \leq q, p, q$ 分别为类 E, F 中概念的数量) 分别为类 E, F 中的概念。

(3) 当层次聚类完成后, 确定每一个类中的父概念。一个概念与类中其他所有概念的平均相似度越大, 说明其与类中其他概念联系越密切, 越具有广泛性, 因此取平均相似度最大的概念作为类的父概念。概念与类中其他概念的平均相似度为

$$\text{simavg}(C_g) = \frac{\sum_{h=1}^r \text{sim}(C_g, C_h)}{r} \tag{2}$$

式中 $C_g, C_h (1 \leq g \leq r, 1 \leq h \leq r, g \neq h, r$ 为类中概念的数量) 为类中的概念。

(4) 通过确定每个簇中的父概念和子概念, 得到煤矿事故本体概念间层次关系。

3.2 煤矿事故本体概念间非层次关系提取

非层次关系主要反映本体概念各个类之间的主要关系, 通过提取非层次关系, 不仅可增加知识表示的完备性和复杂性, 而且确保了对本体进行推理的完备性。概念间非层次关系提取主要是考虑有关联概念对的提取, 并判断概念对中 2 个概念之间的具体关系。

在煤矿事故案例文本中, 如果 2 个煤矿事故本体概念以较高频率出现在同一句子时, 说明这 2 个概念之间可能存在某种关系。当寻找出有关系的煤矿事故本体概念对后, 再筛选出与这 2 个煤矿事故本体概念关系最紧密的动词^[16], 将其作为概念对的关系标签, 从而得到煤矿事故本体概念间非层次关系。本文通过关联规则法自动提取煤矿事故本体概念间非层次关系, 具体实施步骤如下。

(1) 从煤矿事故本体概念集合中取出 2 个概念 C_i, C_z , 判断其是否属于已提取的煤矿事故本体概念间层次关系的同一类中, 若不是则计算这 2 个概念的支持度和置信度, 并判断是否大于或等于最小支持度和最小置信度, 若是, 则说明概念 C_i 和 C_z 具有关联关系, 将概念 C_i 和 C_z 提取出来形成 1 个概念对。

(2) 当提取出存在关系的概念对之后, 需要判断存在何种具体关系。提取煤矿事故案例文本中连接概念 C_i 和 C_z 的所有动词, 计算动词 V 与概念对的点互信息值(式(3)), 若大于或等于设置的点互信息阈值, 则该动词作为概念对的关系标签, 得到煤矿事故本体概念间非层次关系。部分煤矿事故本体概念间非层次关系提取结果见表 2。

$$\text{PMI}((C_i, C_z), V) = \log_2 \frac{P(C_i, C_z, V)}{P(C_i, C_z)P(V)} \tag{3}$$

式中 $P(\cdot)$ 为概率。

表 2 部分煤矿事故本体概念间非层次关系提取结果
Table 2 Extraction results of non-hierarchical relationship between concepts of partial coal mine accident ontologies

概念 C_i	概念 C_z	支持度	置信度	点互信息值	关系标签
瓦斯爆炸	有害气体	0.027 3	0.210 9	4.146 2	产生
顶板事故	回采工作面	0.015 1	0.188 6	5.231 3	发生在
瓦斯泄漏	瓦斯中毒	0.012 7	0.063 8	8.753 1	引发
停风	瓦斯积聚	0.011 5	0.136 1	7.584 1	造成

4 本体可视化

在获取煤矿事故本体概念和概念间关系后, 需要将本体以可视化形式表达, 便于用户直观了解煤矿事故知识, 因此需要对所获得的煤矿事故本体概念和概念间关系进行编辑。为了使煤矿事故本体更好地应用于煤矿领域, 则需要添加实例。目前比较成熟的本体构建工具有 OntoLearn, Text-To-Onto, OntoBuilder, Protégé 等。其中 Protégé 支持多种知识表示, 拥有图形化的用户界面, 易于操作者使用, 同时支持多种本体描述语言。另外 Protégé 拥有优秀的插件能力, 通过安装插件来扩展其功能, 增强所构建本体的应用能力。本文采用 Protégé 本体编辑器, 选取常用的本体描述语言 OWL 对所获得的煤矿事故本体概念、概念间关系和实例进行可视化表达, 得到煤矿事故本体, 如图 3 所示。



图 3 煤矿事故本体

Fig. 3 Coal mine accident ontology

5 结语

提出了一种煤矿事故本体自动构建方法。首先, 通过 BP 神经网络从煤矿事故案例文本中自动提取煤矿事故本体概念。然后, 通过层次聚类法自动提取煤矿事故本体概念间层次关系, 通过关联规则法提取煤矿事故本体概念间非层次关系。最后, 采用 Protégé 本体编辑器对获得的煤矿事故本体概念、概念间关系和实例进行可视化表达, 得到煤矿事故本体。相比以往的人工构建方式, 该方法大大减少了工作量, 同时可构建大规模本体。但在构建煤

矿事故本体的过程中仍存在一些不足:在选取概念特征作为BP神经网络的输入时,只选取了3个特征组成特征向量,不能更准确地描述领域概念,使得概念自动提取过程变得繁琐;煤矿事故本体概念间非层次关系并不全是由动词决定的,本文只考虑动词这种情况,影响非层次关系提取的准确性。因此,寻找更多的特征来提高概念自动提取精度,研究如何自动提取其他词性的词决定的非层次关系将是下一步研究重点。

参考文献(References):

[1] STUDER R, BENJAMINS V R, FENSEL D. Knowledge engineering: principles and methods[J]. Data and Knowledge Engineering, 1998, 25 (1/2): 161-197.

[2] 刘婷. 采煤工作面动态本体构建及推理规则研究[D]. 太原:太原科技大学,2017.

[3] 郭晓黎,王宇,刘瑞祥. 面向煤矿安全事件本体模型研究与应用[J]. 中国煤炭,2014,40(12):113-116.

GUO Xiaoli,WANG Yu,LIU Ruixiang. Research and application of event ontology model of coal mine accidents[J]. China Coal,2014,40(12):113-116.

[4] 桂红军. 煤矿事故应急救援案例推理系统研究[D]. 西安:西安科技大学,2017.

[5] 孟现飞. 基于本体的煤矿事故预警知识库模型及其应用[D]. 徐州:中国矿业大学,2014.

[6] 药慧婷,陈立潮,潘理虎. 掘进工作面本体模型研究[J]. 工矿自动化,2015,41(9):93-96.

YAO Huiting,CHEN Lichao,PAN Lihu. Research of ontology model of heading face[J]. Industry and Mine Automation,2015,41(9):93-96.

[7] 井津. 聚类分析在煤矿安全事故特征分析的应用[J]. 陕西煤炭,2015,34(1):95-98.

JING Jin. Application of cluster analysis in the research of mine safety accident characteristics[J]. Shaanxi Coal,2015,34(1):95-98.

[8] 贺海涛,郑山红,侯丽鑫,等. 基于中文文本的疾病领域本体学习的研究[J]. 吉林大学学报(信息科学版), 2014,32(1):76-81.

HE Haitao,ZHENG Shanhong, HOU Lixin, et al.

Research on disease ontology learning based Chinese text [J]. Journal of Jilin University (Information Science Edition),2014,32(1):76-81.

[9] NOY N F,FERGERSON R W,MUSEN M A. The knowledge model of Protégé-2000: combining interoperability and flexibility [C]//Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management, London, 2000:17-32.

[10] ZHANG Huaping,YU Hongkui,XIONG Deyi,et al. HHMM-based Chinese lexical analyzer ICTCLAS [C]//The 2nd SIGHAN Workshop on Chinese Language Processing,Sapporo,2003:758-759.

[11] 刘奕君,赵强,郝文利. 基于遗传算法优化 BP 神经网络的瓦斯浓度预测研究[J]. 矿业安全与环保,2015, 42(2):56-60.

LIU Yijun,ZHAO Qiang,HAO Wenli. Study of gas concentration prediction based on genetic algorithm and optimizing BP neural network[J]. Mining Safety & Environmental Protection,2015,42(2):56-60.

[12] 贺超峰,华心祝,马菁花,等. 基于 BP 神经网络的回采巷道围岩稳定性分类[J]. 矿业工程研究,2012, 27(3):6-9.

HE Chaofeng,HUA Xinzhu,MA Jinghua, et al. Classification of surrounding rock stability of roadway based on BP neural network[J]. Mineral Engineering Research,2012,27(3):6-9.

[13] 左红涛. 基于人工神经网络的水环境本体概念抽取与本体映射研究[D]. 武汉:华中科技大学,2016.

[14] 熊魏. 基于 TF-IDF 推荐算法的多样性研究[D]. 荆州:长江大学,2018.

[15] 李志义,李德惠,赵鹏武. 电子商务领域本体概念及概念间关系的自动抽取研究[J]. 情报科学,2018, 36(7):85-90.

LI Zhiyi,LI Dehui,ZHAO Pengwu. Research on automatic extraction of ontology concept and its relation in E-commerce [J]. Information Science, 2018,36(7):85-90.

[16] 唐涛. 基于文本挖掘的领域本体学习模型研究[J]. 图书情报工作,2010,54(增刊2):348-352.