



文章编号:1671-251X(2018)10-0010-05

DOI:10.13272/j.issn.1671-251x.2018050019

煤矿安全隐患信息自动分类方法

谢斌红¹, 马非¹, 潘理虎^{1,2}, 张英俊¹

(1. 太原科技大学 计算机科学与技术学院, 山西 太原 030024;

2. 中国科学院 地理科学与资源研究所, 北京 100101)

摘要:人工分类方式难以满足海量煤矿安全隐患信息的分类要求,而基于概率统计的文本自动分类方法分类准确率较低。针对上述问题,提出了一种基于 Word2vec 和卷积神经网络的煤矿安全隐患信息自动分类方法。首先对隐患信息进行分词、去停用词等预处理,然后应用 Word2vec 来表征词之间的语义相似性关系,最后利用卷积神经网络提取隐患信息的局部上下文高层特征,并使用 Softmax 分类器实现隐患信息的自动分类。实验结果表明,该方法实现了端到端的自动分类,可有效提升分类的准确性和全面性。

关键词:煤矿安全; 隐患信息自动分类; 文本分类; 卷积神经网络; Word2vec

中图分类号:TD67 文献标志码:A

网络出版地址:<http://kns.cnki.net/kcms/detail/32.1627.TP.20180919.1430.001.html>

Automatic classification method of coal mine safety hidden danger information

XIE Binhong¹, MA Fei¹, PAN Lihu^{1,2}, ZHANG Yingjun¹

(1. Department of Computer Science and Technology, Taiyuan University of Science and Technology,

Taiyuan 030024, China; 2. Institute of Geographic Science and Natural Resources Research,

Chinese Academy of Science, Beijing 100101, China)

Abstract: Manual classification method is difficult to meet classification requirements of massive coal mine safety hidden danger information, and automatic text classification method based on probability statistics has low classification accuracy rate. In view of the above problems, an automatic classification method of coal mine safety hidden danger information was proposed which was based on Word2vec and convolutional neural network. Firstly, hidden danger information is pre-processed through word segmentation and stop word deletion. Then semantic similarity between words is represented by employing Word2vec. Finally, local context high-level features of hidden danger information are extracted by use of convolutional neural network, and Softmax classifier is used to realize automatic classification of hidden danger information. The experimental results show that the method realizes end-to-end automatic classification and can effectively improve accuracy and comprehensiveness of classification.

Key words: coal mine safety; automatic classification of hidden danger information; text classification; convolutional neural network; Word2vec

收稿日期:2018-05-07;修回日期:2018-08-31;责任编辑:盛男。

基金项目:山西省中科院科技合作项目(20141101001);山西省社会发展科技攻关项目(20140313020-1)。

作者简介:谢斌红(1972—),男,山西万荣人,副教授,硕士,硕士研究生导师,主要研究方向为软件体系结构及服务计算,E-mail:binhongxie@tyust.edu.cn。通信作者:马非(1994—),男,山西临汾人,硕士研究生,主要研究方向为大数据软件工程,E-mail:358105693@qq.com。

引用格式:谢斌红,马非,潘理虎,等.煤矿安全隐患信息自动分类方法[J].工矿自动化,2018,44(10):10-14.

XIE Binhong, MA Fei, PAN Lihu, et al. Automatic classification method of coal mine safety hidden danger information[J]. Industry and Mine Automation, 2018, 44(10): 10-14.

0 引言

随着煤矿安全管理理念的不断发展、安全管理制度的日趋完善及信息化技术的发展,煤矿安全隐患数据量激增,据统计,2013年中国仅工矿企业就排查出523.1万项隐患,数据量约为127.3 GB^[1-2]。近年来,中国在煤矿安全隐患信息处理领域取得了一定成果,例如张长鲁^[3]使用六何分析法和对数线性模型的框架来揭示隐患之间的交互关系;陈运启^[4]利用改进的Apriori算法,从隐患责任部门、隐患种类、隐患等级和隐患发生地点4个维度进行了挖掘分析;张大伟^[5]利用数据立方体和OLAM(On-line Analytical Mining,联机分析挖掘)体系结构来挖掘煤矿隐患信息。但上述研究仅限于对结构化隐患信息的处理应用,缺乏对非结构化或半结构化隐患信息的挖掘分析,需要依靠人工方式进行分类标注,无法满足海量数据的分类要求。因此,如何实现煤矿安全隐患信息自动分类,进而挖掘海量数据隐藏的潜在知识,为管理者科学决策提供依据,已成为当前煤矿亟待解决的问题^[6]。

煤矿安全隐患信息自动分类本质上属于短文本分类问题^[7]。基于概率统计的文本自动分类方法应用较广泛。S. B. Kim等^[8]将朴素贝叶斯(Naive Bayes, NB)算法应用于文本分类领域,取得了很好的效果,但采用词袋模型来表征文本,一方面特征稀疏且维度高,导致计算较复杂,另一方面忽略了上下文的语义信息;黄章树等^[9]根据词频并利用卡方检验来选择文本特征,但短文本相较于长文本词频信息过少,导致分类精度不高。随着神经网络的发展, T. Mikolov等^[10]利用Word2vec模型计算词之间的语义相似性,在降低维度的同时可表征更多语义信息;Y. Kim^[11]通过使用静态的文本向量,运用卷积神经网络(Convolutional Neural Network, CNN)对短文本进行分类,一方面利用CNN权值共享的特性降低了计算复杂度,另一方面通过卷积核分析上下文的多元语义信息。本文在前人研究基础上,提出了一种基于Word2vec和CNN的煤矿安全隐患信息自动分类方法,该方法使用基于Word2vec的特征表示方式,可表征更多的语义相似性关系,利用CNN来提取局部上下文高层特征,可精确对文本进行自动分类。

1 方法原理

煤矿安全隐患信息自动分类方法分为训练和测

试2个部分,主要包含文本预处理模块、Word2vec模块、CNN模块,如图1所示。该方法具有如下优点:①可通过卷积核维度的选择来分析局部上下文的语义信息,更适用于短文本分类;②CNN具有权值共享的特性,可提高训练效率;③Word2vec训练的词向量维度固定,不受词典大小限制,且能很好地表示词之间的语义相似性关系;④无需手动提取文本特征,训练好权重即可得到很好的分类效果,真正实现端到端的自动分类。

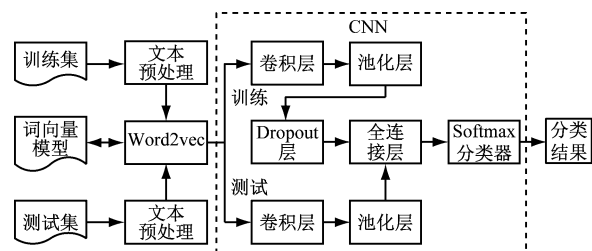


图1 煤矿安全隐患信息自动分类方法原理

Fig. 1 Principle of automatic classification method of coal mine safety hidden danger information

1.1 文本预处理模块

文本预处理主要包含分词和去停用词2个阶段。由于词粒度的特征表示方式远优于字粒度的特征表示方式(字粒度的特征表示方式会导致大量n-gram信息丢失),而中文的词语之间没有间隔符,所以处理中文文本时需要通过分词将中文词语间隔开。去停用词则是去除一些表征能力弱的虚词、副词等。

1.2 Word2vec模块

Word2vec模块通过训练将分词之后的每个字词转换为计算机可理解的稠密向量^[12],并能很好地表征词之间的相似性^[13]。其包含CBOW(Continuous Bag-of-Words,连续词袋)模型和Skip-gram模型2种训练模式^[14],如图2所示。

CBOW模型根据上下文来计算当前词的词向量,其输入为当前词 $W(t)$ 周围 n 个词的词向量,例如 $n=2$ 时,CBOW模型的输入为 $W(t)$ 前2个词的词向量和后2个词的词向量之和,输出为 $W(t)$ 的词向量;Skip-gram模型根据当前词计算上下文其他词的词向量,其输入为当前词 $W(t)$ 的词向量,输出为当前词 $W(t)$ 周围 n 个词的词向量。Word2vec训练参数见表1。

1.3 CNN模块

CNN模块的功能是特征提取和分类,主要由卷积层、池化层、全连接层、Dropout层和分类器组成,其工作流程如图3所示。

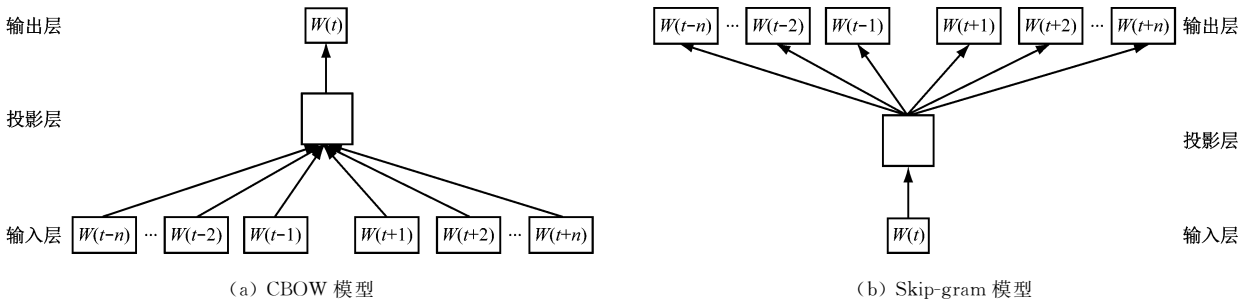


图 2 CBOW 和 Skip-gram 模型
Fig. 2 CBOW and Skip-gram model

表 1 Word2vec 训练参数

Table 1 Word2vec training parameters

参数	说明	参数	说明
train	需训练的文本数据路径	output	训练结果的输出文件
cbow	是否使用 CBOW 模型	size	输出词向量的维度
window	训练窗口的大小,即考虑的上下文词量	negative	是否使用负例采样法
hs	是否使用哈夫曼树	min-count	设置词语的最低频率
binary	是否保存为二进制文件	alpha	学习率

式中: $f()$ 为非线性激活函数; $\cdot$ 为卷积操作; b 为偏置。

在卷积过程中需要注意对于边缘的处理,在滑动过程中如果超出了文本矩阵的边界,则要补零,以保证不丢失边缘信息。虽然卷积过后得到的特征相较于原始输入数据的维度降低了不少,但维度还是相对较高的,特征也不明显。因此,需要利用池化层对特征图进行降维择优操作。池化层可选方案主要有最大值池化和均值池化 2 种,本文使用最大值池化,池化之后的特征图为

$$\hat{M} = \max M \tag{3}$$

池化层可在降维的同时保留更突出的特征,同时还可处理不同长度的特征图,保留其最关键的特征。

设 g 种不同维度的卷积核进行卷积、池化之后的输出为 $\hat{M}_1, \hat{M}_2, \dots, \hat{M}_g$, 则全连接层的输入为

$$S = \hat{M}_1 \oplus \hat{M}_2 \oplus \dots \oplus \hat{M}_g \tag{4}$$

全连接层为 BP 神经网络结构,其输出为

$$Y = f(\mathbf{GS} + b) \tag{5}$$

式中 $\mathbf{G}$ 为权重向量。

为避免在隐层神经元权重更新时出现过拟合现象,使用 Dropout 层减弱神经元间的联合适应性,以提高 CNN 模块的泛化能力。Dropout 层通过使用伯努利模型随机地将某些神经元的概率设置为 0,来达到随机剔除神经元的目的。当全连接层前加入 Dropout 层时,全连接层的输出为

$$Y_d = f(\mathbf{GS}r + b) \tag{6}$$

式中 r 为二值伯努利分布, $r = \text{Bernouli}(p)$, p 为 Dropout 率。

采用 Softmax 分类器可将分类结果转换为 $[0, 1]$ 之间的概率值,选择概率最大的一类作为最终分类结果。

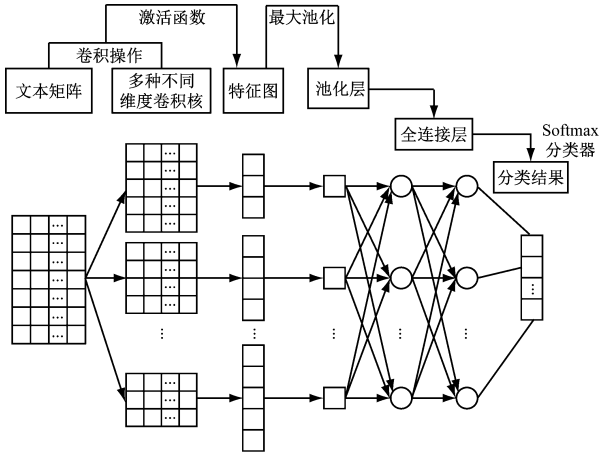


图 3 CNN 模块工作流程

Fig. 3 CNN module workflow

设输入数据为长度为 m 的句子, $\mathbf{x}_i$ 为句子中第 i ($i=1, 2, \dots, m$) 个词的词向量, $\mathbf{x}_i \in \mathbf{R}^q$, q 为词向量维度,则该句子的文本矩阵为

$$\mathbf{X}_{1:m} = \mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \dots \oplus \mathbf{x}_m \tag{1}$$

式中 $\oplus$ 为连接操作。

卷积层主要是对文本矩阵 $\mathbf{X}_{1:m}$ 与卷积核 k ($k \in \mathbf{R}^{h \times q}$, h 为卷积核窗口的高度) 进行卷积操作,并通过引入非线性激活函数精确表征局部上下文高层特征,从而形成特征图:

$$M = f(k \cdot \mathbf{X}_{1:m} + b) \tag{2}$$

2 实验分析

为验证本文方法的有效性,实验基于 Windows 操作系统,开发平台选择 Pycharm,开发语言为 python。煤矿安全隐患信息样本数据总量为 26 036 条,其中训练集占 90%,测试集占 10%,采用过抽样方式来均衡数据。煤矿安全隐患信息分类依据为《煤矿安全生产事故隐患分类分级标准》,实验主要目标是完成样本在“安全管理类”、“设备设施类”、“作业场所类”、“从业人员类”一级分类层面的自动分类任务。

2.1 文本预处理

文本预处理之前,首先要构建面向煤矿安全领域的专有词库,以提升分词的精度,专有词库的构建由人工完成。分词主要使用 NLPIR 开源工具包,其具有分词精度高、效率快的优势,并且可实现自适应分词,即发现词库中没有的新词时,可将新词应用于文本的分词。去停用词主要是根据停用词表,去除文本中一些没有实际作用的词,如“的”、“比如”等。例如,对“总回风巷与西轨道巷联络风门无语音报警装置,风门内放有风筒布等杂物”进行预处理后变成“总/回风巷/西/轨道巷/联络/风门/无/语音报警/装置/风门/放/风筒布/杂物”。

2.2 词向量训练

将预处理好的文本语料保存到文本文件中,使用 Cygwin64 Linux 模拟器运行 Word2vec 模块进行训练。实验发现,Skip-gram 模式相比 CBOW 模式可更加精确地表示词之间的语义相似性,因此实验中选用 Skip-gram 模型,词向量维度设置为 300,窗口大小设置为 5,结果存储为二进制文件。

2.3 CNN 模块构建

使用 Google 开源的 TensorFlow 深度学习框架构建 CNN 模块。在卷积层使用 3,4,5 维 3 种卷积核(每种 300 个)进行卷积,以提高分类精度;经过实验验证,Dropout 层的 Dropout 率设置为 0.5 时效果最佳;与传统的 Sigmoid 激活函数相比,ReLU 函数具有单侧抑制的作用^[15],可有效防止过拟合,提高数据分类收敛速度,因此实验中选取 ReLU 作为卷积之后的激活函数。

2.4 实验结果

采用精确率 P 、召回率 R 、 F_1 值 3 个指标来评价分类效果。精确率反映分类结果的准确程度,精

确率越高,表明分类结果越准确;召回率反映分类结果的全面程度,召回率越高,表明分类结果越全面; F_1 值为精确率和召回率的综合评价指标, F_1 值越高,表明分类的综合性能越好。

$$P = \frac{TP}{TP + FP}$$
(7)

$$R = \frac{TP}{TP + FN}$$
(8)

$$F_1 = \frac{2TP}{2TP + FP + FN}$$
(9)

式中:TP 为将正类分类为正类的个数;FP 为将负类分类为正类的个数;FN 为将正类分类为负类的个数。

针对煤矿安全隐患信息的分类评价结果见表 2、表 3。可看出本文方法可很好地对煤矿安全隐患信息进行自动分类,相较于文献[9]方法,其精确率提升了 13.54%,召回率提升了 13.58%, F_1 值提升了 13.35%。

表 2 煤矿安全隐患信息分类评价结果
Table 2 Classification evaluation result of coal mine safety hidden danger information

类别	指标		
	精确率	召回率	F_1 值
安全管理类	0.956 8	0.975 6	0.968 5
设备设施类	0.980 8	0.992 1	0.984 2
作业场所类	0.988 4	0.966 8	0.975 6
从业人员类	0.974 3	0.970 2	0.972 9

表 3 不同方法下分类评价结果
Table 3 Classification evaluation results under different methods

方法	精确率	召回率	F_1 值
文献[9]方法	0.839 7	0.840 4	0.841 8
本文方法	0.975 1	0.976 2	0.975 3

3 结语

煤矿安全隐患信息自动分类方法首先对数据进行分词、去停用词等预处理,然后使用 Word2vec 模块表征词之间的语义相似性关系,最后运用 CNN 抽取局部上下文高层特征进行分类,真正实现了端到端的自动分类,提升了分类的准确性和全面性。未来的研究工作主要是实现煤矿安全隐患信息在二级分类、三级分类层面的映射,完成煤矿安全隐患信息的多层次分类。

参考文献(References):

- [1] 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[J]. 中国安全科学学报, 2013,23(4):102-107.
MENG Xianfei, LI Keye, LIU Fei. Study on coal mine safety risk pre-control based on three-levels nested management mode[J]. China Safety Science Journal, 2013,23(4):102-107.
- [2] 贾立猛,丁淑艳,陈全. 煤矿隐患排查治理信息系统的研究和应用[J]. 煤矿安全, 2014,45(12):239-241.
JIA Limeng, DING Shuyan, CHEN Quan. Research and application of coal mine hidden dangers investigation and management information system [J]. Safety in Coal Mines, 2014,45(12):239-241.
- [3] 张长鲁. 煤矿事故隐患大数据处理与知识发现分析方法研究[J]. 中国安全生产科学技术, 2016,12(9):176-181.
ZHANG Changlu. Study on big data processing and knowledge discovery analysis method for safety hazard in coal mine[J]. Journal of Safety Science and Technology, 2016,12(9):176-181.
- [4] 陈运启. 数据挖掘技术在煤矿隐患管理中的应用[J]. 工矿自动化, 2016,42(2):27-30.
CHEN Yunqi. Application of data mining technology in coal mine hidden hazard management[J]. Industry and Mine Automation, 2016,42(2):27-30.
- [5] 张大伟. 基于 OLAM 的煤矿企业安全隐患趋势分析[J]. 煤炭工程, 2015,47(5):139-142.
ZHANG Dawei. Analysis of coal mine safety hidden danger trends based on OLAM[J]. Coal Engineering, 2015,47(5):139-142.
- [6] 周浩,宋保华. 大数据在安全生产中的作用[J]. 安全, 2015,36(10):8-11.
- [7] 谭章禄,陈晓,宋庆正,等. 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.
TAN Zhanglu, CHEN Xiao, SONG Qingzheng, et al. Analysis for the potential hazardous risks of the coal mines based on the so-called text mining[J]. Journal of Safety and Environment, 2017,17(4):1262-1266.
- [8] KIM S B, HAN K S, RIM H C, et al. Some effective techniques for naive Bayes text classification[J]. IEEE Transactions on Knowledge & Data Engineering, 2006,18(11):1457-1466.
- [9] 黄章树,叶志龙. 基于改进的 CHI 统计方法在文本分类中的应用[J]. 计算机系统应用, 2016,25(11):136-140.
HUANG Zhangshu, YE Zhilong. Application of text categorization based on improved CHI-square statistic method[J]. Computer Systems & Applications, 2016,25(11):136-140.
- [10] MIKOLOV T, YIH W T, ZWEIG G. Linguistic regularities in continuous space word representations [C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, 2013:746-751.
- [11] KIM Y. Convolutional neural networks for sentence classification [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, 2014:1746-1751.
- [12] XING Chao, WANG Dong, LIU Chao, et al. Normalized word embedding and orthogonal transform for bilingual word translation [C]//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, 2015:1006-1011.
- [13] LI Shuguo. The application of information technology in coal mine management[J]. Mechanical Management & Development, 2014,340(3):38-45.
- [14] LE Q, MIKOLOV T. Distributed representations of sentences and documents [C]//Proceedings of the 31st International Conference on Machine Learning, Beijing, 2014:1188-1196.
- [15] YAROTSKY D. Error bounds for approximations with deep ReLU networks[J]. Neural Networks: the Official Journal of the International Neural Network Society, 2017,94:103-114.