基于朴素贝叶斯模型和 PageRank 算法的电网事故主题爬虫算法

谢树泳[1],刘之亮[2]

(1.广东电网有限责任公司 惠州供电局,广东 惠州 516000;2.南方电网有限责任公司,广州 510000)

摘 要:为了减少电网人身安全事故,利用数据挖掘技术构建和分析事故多维数据,建立准确的预警模型十分必要.其中一个极具挑战性的问题是如何在海量网页中自动化采集人身事故样本数据.提出一种朴素贝叶斯模型与 PageRank 结合的主题爬虫算法.首先采用中文文本分割和设置关键词词频的方法对数据预处理,进行特征选择后,构建并训练朴素贝叶斯分类模型,从而实现电网事故分类准确度的显著提升.然后利用 PageRank 算法对精确分类后的网页进行主题相关性排序,有效避免普通爬虫方法中出现的主题漂移问题.实验结果表明,不论是在相同时间还是相同页面数的条件下,该方法的页面收获率均高于单独使用朴素贝叶斯分类器或 PageRank 的收获率,即本方法能够在大量网页中更高效、准确地爬取电网事故信息.

关键词:电网安全;人身事故;朴素贝叶斯模型;PageRank 算法;主题爬虫

中图分类号:TP391 **文献标志码**:A **文章编号**:1000-2367(2025)02-0124-07

电力事故的死亡率高于其他类型工伤事故死亡率,严重危害相关从业人员人身安全以及公司安全生产[1-2].电网人身事故样本数据少,使得研究人员无法通过大规模样本进行多维度分析.电网规模扩大,在线运行历史数据激增,因此,需要自动化采集电网事故样本的解决方案.其中的关键问题是如何快速准确地从海量网页中找到相关信息.使用传统搜索引擎进行检索是常用方法[3],但不精准检索,因此研究主题爬虫算法,根据网页的主题相似度进行有目的的爬取[4-5].

本文将主题判断视为文本分类问题,利用基于文本分类器计算主题相似度的方法.贝叶斯分类器因简单可靠被应用于文本分类[6],GAO 等[7]提出改进的分布式朴素贝叶斯自动分类算法提高了分类效率.利用朴素贝叶斯对待爬取链接的简介文本进行快速分类,有助于爬虫算法更精准快速地抓取相关信息.

在爬虫领域,如何提高爬虫的精准度和速度是重点研究方向.目前,主题爬虫研究方法主要分为 3 种[8]:基于智能优化、基于网页内容和基于链接分析的算法.基于智能优化算法的主题爬虫,其经典搜索策略有遗传算法[9]和退火算法.基于网页内容的主题爬虫经典搜索策略有鱼群搜索算法、鲨鱼搜索算法和向量空间模型(VSM).但基于智能优化、基于网页内容的方法在链接价值预测方面存在不足,容易忽视链接结构对结果的影响.为解决上述问题,基于链接分析的主题爬虫方法应运而生,其中 PageRank 是一种经典策略[10].董伟等[11]融合 PageRank 与评论情感倾向,利用 PageRank 和 SVM 等算法构建出在线健康社区用户影响力的测量方法,并从使用价值的角度,进一步计算用户的综合影响力,从而丰富健康社区用户信息行为.文献[12]分

收稿日期:2023-12-26;**修回日期**:2024-03-07.

基金项目:国家自然科学基金(52377103;52277148);南方电网科技项目(0313002023030103AJ0003;031300KK52222091).

作者简介(通信作者):谢树泳(1990—),男,广东潮州人,广东电网有限责任公司工程师,主要从事电力安全生产工作,E-mail:1529321557@qq.com.

引用本文:谢树泳,刘之亮.基于朴素贝叶斯模型和 PageRank 算法的电网事故主题爬虫算法[J].河南师范大学学报(自然科学版),2025,53(2):124-130.(Xie Shunyong,Liu Zhiliang.A focused crawler algorithm based on Naive Bayes model and PageRank on power grid accidents[J].Journal of Henan Normal University(Natural Science Edition),2025,53(2):124-130.DOI:10.16366/j.cnki.1000-2367.2023.12.26.0001.)

析了搜索引擎的工作原理以及有助于提高排名的因素，提出一种加权 PageRank 算法来解决偏见问题.基于链接分析的方法注重链接结构，侧重于抓取权威性强且质量高的网页，但较少考虑主题相关度.叶小榕等[13]利用朴素贝叶斯算法对抓取的网站用户数据进行清洗，再通过 PageRank 算法计算用户排名，实现了更贴合实际的用户社区划分.杨晶[14]用朴素贝叶斯分类算法对候选故障进行分类，并且使用改进后的 PageRank 算法实现故障重要性排序.

受上述工作的启发，本文提出基于朴素贝叶斯分类器和 PageRank 结合的主题爬虫算法.利用朴素贝叶斯分类简单、准确性高的优点，对网页先进行电网事故相关的精确过滤，使基于 PageRank 的主题爬虫爬取的网页更贴近主题，解决了普通爬虫的主题漂移问题.实验验证，该方法可以获取大量有效的事故样本信息，可靠性好.

1　基础理论

1.1　朴素贝叶斯分类器

贝叶斯分类是一种数据处理方法，其中朴素贝叶斯(Naive Bayes，NB)分类器是应用最广泛的模型之一.与贝叶斯分类不同的是，朴素贝叶斯假设特征相互独立，从而简化了分类模型.这种方法基于贝叶斯定理，通过假设特征独立来进行分类，具有原理简单、错误率低和分类效率高等优点[15].本文利用朴素贝叶斯算法对电网人身安全事故信息网页进行文本分类.朴素贝叶斯分类模型算法原理如下：

1)将数据集(内容特征和链接特征)表示为一个 n 维特征向量$(x_1,x_2,\cdots,x_n)$，分别描述给定的待分类数据样本 $\boldsymbol{X}$ 的 n 个属性.2) 假设有 2 个类 C_i，$i=1,2$，分别代表无 / 有电网人身事故信息.计算在 $\boldsymbol{X}$ 条件下 2 个类的后验概率，预测 $\boldsymbol{X}$ 属于哪一类.3) 计算类的先验概率，即统计得到网页中有无电网人身事故信息的概率 $P(C_i)=s_i/s$，其中，s_i 代表类中的训练样本实例数，s 是训练样本总数.4) 计算样本 $\boldsymbol{X}$ 的条件概率.由朴素贝叶斯算法假设特征之间相互独立可得 $p(\boldsymbol{X}\mid C_i)=\prod_{k=1}^{n}p(x_k\mid C_i)$，其中，$p(x_k/C_i)$ 可以由训练样本估计.5) 计算在 $\boldsymbol{X}$ 条件下 2 个类的后验概率 $P_{\mathrm{NB}}(\boldsymbol{X})=\mathrm{argmax}_{C_i\in C}P(C_i)\prod_{=1}^{n}p(x_k/C_i)$，样本 $\boldsymbol{X}$ 被分类到后验概率较大的类中，实现数据分类.

1.2　PageRank 算法

PageRank 算法是一种基于网页链接的主题爬虫方法，通过网页之间的指向来计算网页的 PageRank 值和排名.每个节点表示一个网页，由节点指出的箭头是前向链接，被指向该节点的箭头是反向链接.如网页 A 有一个箭头指向网页 B，则称 A 是 B 的一个前向链接，B 是 A 的一个反向链接.

该算法表示，一个网页的 PageRank 值取决于它所有反向链接页面贡献值的累加和，由所有链入它的页面的重要性经过递归算法得到，是一个迭代过程.一个网页的 PageRank，$R_P=\sum_{j\in Q(i)}R_P(j)/N(j)$，其中，$i$，$j$ 为网页序号；$Q(i)$ 表示网页 i 指向的所有链接的集合；$N(j)$ 表示网页 j 指向所有链接的数目.网页的反向链接越多表示该网页越重要，其 PageRank 值越高.

在实际应用中，网页链接的相互指向会出现网页出度或入度为 0 的情况，这时会产生等级泄露和等级下沉的问题.为了解决这一问题，需要先去掉链接网络中所有出度为 0 的节点，然后定义一个阻尼系数 $d(0<d<1)$，表示用户随机指向一个链接的概率.令其乘以 d 的 R_P 分配给其指向的前向链接，而乘以 $1-d$ 部分的 R_P 平均分配给链接网络其他所有节点，则 $R_P(i)=d\sum_{j\in Q(i)}\dfrac{R_P(j)}{N(j)}+\dfrac{1-d}{n}$，其中，$n$ 表示总体节点的个数.

通过 PageRank 算法得到的结果具有收敛性和可靠性，所以在对网页进行朴素贝叶斯分类后再利用 PageRank 算法计算网页的 R_P 并排序，从而优先爬取最相关的网页，获取最合适的信息.

2　基于朴素贝叶斯和 PageRank 的主题爬虫算法

为了解决普通主题爬虫算法存在的主题漂移问题，本文提出了一种基于朴素贝叶斯和 PageRank 算法

的主题爬虫改进方法.该方法继承朴素贝叶斯分类原理简单的优点,利用分类后网页的 R_P 进行主题相关度排序,避免主题漂移问题的发生.本算法的核心是网页分析模块和搜索调度模块,算法框架图如图 1 所示.

首先,通过常见搜索引擎(如 Firefox、Google)获取网页 URL.模拟客户端发送 HTTP 请求后,下载网页并解析其全文内容.接着,对内容进行预处理,提取简介和链接.网页分析模块通过主题相关度过滤网页,使用贝叶斯分类器将网页分为“电网事故安全”和“其他”类别.然后,搜索调度模块利用 PageRank 算法计算网页优先级并排序.URL 队列分为待爬取和已爬取队列,确保爬虫更精确、高效地访问 URL.最终,爬取得到的信息存储在数据库中.

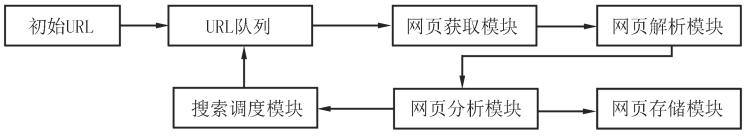

图1 算法框架图

Fig.1 Algorithm framework diagram

2.1 网页分析模块

该模块是主题爬虫的核心,其作用是通过对主题相关度的判断和预测,完成网页过滤工作,使接下来通过搜索调度,能够优先访问与主题相关的 URL.根据宽度优先规则,网页分析包含文本预处理、特征选择和朴素贝叶斯分类器模型构建与训练 3 个步骤.

2.1.1 文本预处理

为了将不符合输入规则的文本转换为朴素贝叶斯分类器能识别的内容,进行文本预处理是必要的.中文文本不像英文那样用空格分隔,因此需要分词来提高分类准确度.本文使用 jieba 中文分词库中的精确模式,通过确定汉字之间的关联概率进行文本分割,同时添加自定义词组以提高分词正确率[16].比如,处理“国家电网郑州供电公司发生一起人身伤亡事故,1 人死亡.违规进入高压开关间隔,导致郑州祥和集团工作负责人触电死亡”后的分词结果如下:

国家电网|郑州|供电|公司|发生|一起|人身|伤亡事故|,|1|人|死亡|.|违规|进入|高压|开关|间隔|,|导致|郑州|祥和|集团|工作|负责人|触电|死亡

2.1.2 特征选择

中文文本经过分割处理之后,余下部分可以看作有效词语,于是用余下部分作为初步提取文本的特征项.采用统计学思想,在提取特征前还需要给不同的关键词设定相应频率.电网相关的行业内特殊表达作为关键词,其重要性与出现频率正相关.本实验中关键词设置为“电力”、“电网”和“人身事故”等,通过提高关键词词频的方法进行分析,根据重要程度给相应关键词的词频乘以 10 到 100 不等.使用 TF 词频统计分类方法,具有易于计算的优点.

2.1.3 朴素贝叶斯分类器模型的构建

本文利用朴素贝叶斯分类器进行主题相关性的判断与预测,模型流程图如图 2 所示.文本经过数据处理和特征提取,输出一般是特征向量.特征向量经过学习处理,得到需要的分类信息,以此来构建朴素贝叶斯分类器[17].朴素贝叶斯分类器的具体实现步骤为:1)将数据集中的数据进行处理,分离出特征和标签,并且划分数据集为训练集和测试集;2)对训练集和测试集数据进行中文文本分割操作.初始化 TfidfVectorizer,用 TfidfVectorizer 向量化训练集和测试集数据;3)初始化 MultinomialNB 模型,使用向量化的训练集数据训练朴素贝叶斯分类模型,然后使用训练后的模型对向量化的测试集数据进行预测;4)计算并打印准确率和分类报告,返回训练好的模型和向量化器.

2.2 搜索调度模块

为确保爬虫对 URL 更有效、合理地访问,网络爬虫会根据网页制定合理的搜索规则,本算法结合基于链接分析的搜索策略 PageRank 算法.PageRank 算法的具体实现步骤为:

1)创建一个页面索引,将每个 URL 映射到一个索引,根据链接集的大小初始化一个零矩阵 $\boldsymbol{M}$;2)若链接集中的页面没有外链,就将该页面链接到所有页面;否则更新矩阵 $\boldsymbol{M}$ 以表示页面间的链接结构;3)对矩阵 $\boldsymbol{M}$ 按列进行归一化,初始化一个随机的 R_P 向量,迭代最大迭代次数,再计算 R_P;4)返回每个 URL 及其对

应的 R_P 的映射.

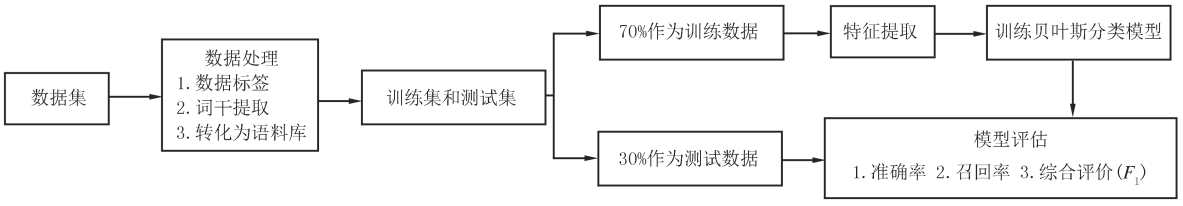

图2　朴素贝叶斯模型流程图

Fig. 2　Naive Bayes model flow char

3　主题爬虫算法实验与讨论

主题爬虫算法实验使用 Python 语言开发，搜索到相关网页元数据.对于电网人身事故这一主题，研究网页链接的相关性程度，是否可以保证集中抓取.本文分别实现了朴素贝叶斯分类模型训练、PageRank 算法实验以及本文主题爬虫算法，并对结果进行讨论与分析.

3.1　数据集

为了保证实验分析的客观性，本实验的数据集采用 Google、Firefox 等常用搜索引擎.几个知名的门户网站，如百度(http://www.baidu.cn)、国家电网有限公司(http://www.sgcc.com.cn)和国家能源局(http://www.nea.gov.cn)等作为爬虫的种子 URL，获取网页，共抓取 800 条 URL.然后对这些网页进行关于电网事故安全信息分类，其中 70%作为训练集，剩下的 30%作为测试集.

3.2　朴素贝叶斯分类模型训练

本次朴素贝叶斯分类模型训练使用 2 种方法：未进行文本分割的方法和经过文本分割的方法.每种方法分别训练 5 次，计算实验结果的平均值和方差，表 1 给出了 2 种方法的主题爬虫算法实验数据.模型将数据集里的所有网页分为“电网人身事故”和“其他”2 类，设置的关键词有“电力”“电网”“人身事故”等.判断主题爬虫抓取性能主要指标有准确率、召回率和综合评价(F_1).

表 1　2 种方法分类结果对比

Tab. 1　Comparison of the classification results between the two methods

方法	准确率/%	召回率/%	F_1/%
未进行文本分割	81.00±0.06	97.12±0.18	87.45±0.09
经过文本分割	**99.95±0.02**	**96.35±0.00**	**98.11±0.01**

网页内容比较详细地反映了网页的主题信息，实验结果表明，在选用的 2 种分类方法中，经过中文文本分割的朴素贝叶斯分类器准确率与 F_1 有很大程度上的提升，分别提高了 18.95%和 10.66%.虽然召回率相比于未经数据处理的方法低了不到 1.00%，但是综合评价 F_1 能达到 98.11%，表示经过文本分割处理的朴素贝叶斯分类模型，能够更准确地获得与电网人身事故相关的网页，提升了算法整体性能.并且经过文本分割方法的 3 项指标方差均减小了，表明实验结果趋于稳定，其中召回率的稳定性有最大的改善.

3.3　PageRank 算法实验

为验证 PageRank 算法的有效性，选取“电网人身事故”作为主题.由于本实验选择的初始网页来自电力相关网站，并且所查询关键字的含义较为简单，故根据算法得到的每个网页的 R_P 是可以信赖的.以“电网事故安全”为主题进行搜索，从返回结果网页中选取 6 个链接，分别标号为 1～6，表 2 为仅使用 PageRank 算法计算和将网页过滤后再使用 PageRank 计算得到各网页的 R_P.

由表 2 可以看出，6 个链接中，R_P 最高的是链接 4，其次是链接 1 和链接 2，与实际情况吻合.根据对查询结果进行分析可知，链接 4 在内容上与搜索主题“电网事故安全”关联密切.根据上述实验结果可知，基于 PageRank 算法能更全面地找出电网事故安全关键词的重要页面.

表 2 基于 PageRank 算法的计算结果

Tab. 2 Calculation results based on PageRank algorithm

网页	URL	R_P
1	https://news.bjx.com.cn/search? kw=电力安全 &type=5	0.003 7
2	https://news.bjx.com.cn/search? kw=电力安全生产 &type=5	0.003 6
3	https://news.bjx.com.cn/search? kw=国家能源局 &type=5	0.003 2
4	https://news.bjx.com.cn/html/20200211/1041538.shtml	0.004 0
5	https://news.bjx.com.cn/search? kw=电力安全监管 &type=5	0.003 1
6	https://news.bjx.com.cn/search? kw=电力设备事故 &type=5	0.002 7

3.4 实验结果讨论与分析

本次实验研究在相同时间或相同页面数的情况下,分别实现基于朴素贝叶斯分类器、基于 PageRank 算法以及本文提出的方法,并对这 3 种方法的电网事故网页收获率进行比较,收获率即有用网页占全体网页的比值.

图 3 为时间限制实验的结果.在 60 min 内,每隔 5 min 设计一次实验,设置爬虫的 max_pages 参数为 1 000 页,爬虫运行对应的时间,分别观察和记录在相应时间内爬取的页面数量.由图 4 可以看出,仅使用 PageRank 算法获取的有用页面占比只在 40%以下,并且随实验时间的增长,收获率的值下降较快.由于对网页简介进行了过滤,使用朴素贝叶斯分类的方法获取到的有效页面高于前者,在初级阶段达到了 63%.本方法在初期能达到最高收获率 70%,在 60 min 时,收获率也在 60%以上.

图 4 为页面限制实验的结果,设计页面数分别为 5 页、10 页、15 页、20 页、25 页和 30 页,对比 3 种方法电网事故网页获得率的实验结果.其中,主题爬虫算法爬取 5 页能够获取上千条链接,爬取 30 页获得的链接达到上万条.由图 4 可以看出,仅使用 PageRank 算法获取的有用页面占比只在 10%以下.由于对网页简介进行了过滤,使用朴素贝叶斯分类的方法获取到的有效页面高于前者,在初级阶段达到了 50%.而本文提出的算法是将两者结合起来,该方法爬取到的电网事故相关网页占比与前 2 种单一方法相比,性能上有了大幅提升.

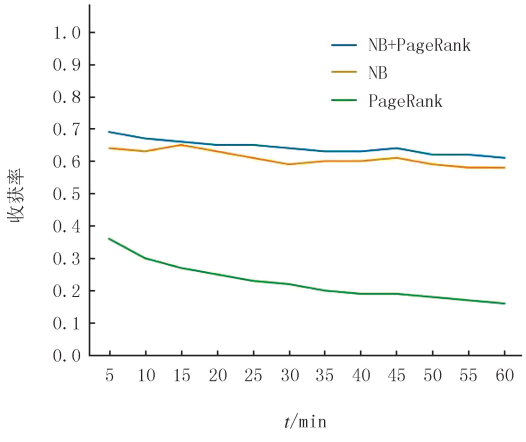

图3 不同时间下的网页收获率

Fig. 3 Page harvesting under different time

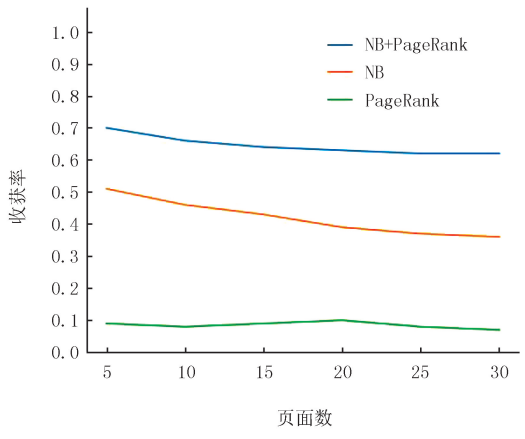

图4 不同页面数下的网页收获率

Fig. 4 Page harvesting under different pages

完成时间限制和页面限制实验后,评估所采用的 3 种方法的主题相关性.首先用 3 种方法分别爬取 1 000 条网页链接,然后标注出这 1 000 条链接中与电力、电网相关的链接,使用 3 种算法分别计算链接的相关性占比.本文算法可以获得很好的页面相关性,占比达到 68.0%,而朴素贝叶斯分类器、PageRank 算法分别为 59.5%、52.3%.

本文算法通过爬取电网事故通报等特定形式的网页提取事故信息,相同链接对应网页里可能有多个事故信息,通过正则表达式对网页文本内容的指定位置自动提取.爬取的维度包括电网事故的来源链接、发生日期、事故类型、死亡人数、发生地点、项目信息、损失情况、死亡人员信息、重伤人员信息、主要责任人员信息

以及不安全因素等信息.

表 3　爬虫获取的部分属性数据

Tab. 3　Partially retrieved attribute data by web crawlers

来源链接	发生时间	事故类型	死亡人数
https://www.nea.gov.cn/2016－11/24/c_135855744.htm	2016-10-23	坠落	1
https://www.nea.gov.cn/2016－11/24/c_135855744.htm	2016-10-02	高处坠落	1
https://www.nea.gov.cn/2020－06/01/c_139104451.htm	2020-04-06	其他	2
https://www.nea.gov.cn/2018－05/30/c_137217639.htm	2018-04-21	触电	1
https://www.nea.gov.cn/2020－08/27/c_139322243.htm	2020-07-02	其他	5

4　结　论

利用 Python 编写主题爬虫获取网页信息，在对信息进行预处理操作的基础上，构建和训练朴素贝叶斯分类器进行网页过滤，得到电网人身事故相关网页链接.在此基础上，根据 PageRank 算法计算链接的 R_P，得到链接的优先级.该方法继承了朴素贝叶斯分类简单可靠的优点，使 PageRank 算法在一定程度上避免了主题漂移现象的发生.最后，通过实验中的对比数据可视化本算法的有效性，为进一步分析电网人身事故发生规律提供了大量样本元数据.

该方法虽然在一定程度上削弱了主题漂移现象的产生，但并不能完全避免；而且在页面限制实验中，由于每个页面存在大量链接，部分页面的链接数量会达到 2 000 条，爬取所有链接的简介会耗时很久.因此，在未来的工作中，可以在本文提出的主题爬虫算法基础上，进一步改进朴素贝叶斯，达到更准确的分类效果；另外，可以继续对爬虫算法进行优化，以减少实验过程中的时间成本.

参　考　文　献

[1]　黄鸣宇，祁升龙，芦翔，等.面向配网保护的集分联合馈线自动化控制方法[J].河南师范大学学报(自然科学版)，2020，48(5)：49-54.
HUANG M Y，QI S L，LU X，et al.Combined centralized and distributed control method for distribution network protection[J].Journal of Henan Normal University(Natural Science Edition)，2020，48(5)：49-54.

[2]　严玉琼，张苏，梁志星，等.2016—2021 年我国电力企业人身事故统计与规律分析[J].安全，2023，44(4)：46-51.
YAN Y Q，ZHANG S，LIANG Z X，et al.Statistics and analysis of electric power enterprises personal accidents in China during 2016-2021[J].Safety & Security，2023，44(4)：46-51.

[3]　LI H Y，HU M M，LI G.Forecasting tourism demand with multisource big data[J].Annals of Tourism Research，2020，83：102912.

[4]　潘晓英，陈柳，余慧敏，等.主题爬虫技术研究综述[J].计算机应用研究，2020，37(4)：961-965.
PAN X Y，CHEN L，YU H M，et al.Survey on research of topic crawling technique[J].Application Research of Computers，2020，37(4)：961-965.

[5]　LIU J F，LI X，ZHANG Q S，et al.A novel focused crawler combining Web space evolution and domain ontology[J].Knowledge-Based Systems，2022，243：108495.

[6]　MARON M E，KUHNS J L.On relevance，probabilistic indexing and information retrieval[J].Journal of the ACM，1960，7(3)：216-244.

[7]　GAO H Y，ZENG X，YAO C H.Application of improved distributed naive Bayesian algorithms in text classification[J].The Journal of Supercomputing，2019，75(9)：5831-5847.

[8]　吴宗柠，狄增如，樊瑛.多层网络的结构与功能研究进展[J].电子科技大学学报，2021，50(1)：106-120.
WU Z N，DI Z R，FAN Y.The Structure and Function of Multilayer Networks：Progress and Prospects[J].Journal of University of Electronic Science and Technology of China，2021，50(1)：106-120.

[9]　YAN W，PAN L.Designing focused crawler based on improved genetic algorithm[C]//2018 Tenth International Conference on Advanced Computational Intelligence.Piscataway：IEEE Press，2018：319-323.

[10]　LI C L，BAI J P，ZHAO W J，et al.Community detection using hierarchical clustering based on edge-weighted similarity in cloud environment[J].Information Processing & Management，2019，56(1)：91-109.

[11]　董伟，陶金虎.融合 PageRank 与评论情感倾向的在线健康社区用户影响力研究[J].图书情报工作，2021，65(11)：14-23.
DONG W，TAO J H.Research on the user's influence in online health community based on PageRank and emotional tendency[J].Library and Information Service，2021，65(11)：14-23.

[12] SAMSUDEEN SHAFFI S,MUTHULAKSHMI I.Weighted PageRank algorithm search engine ranking model for web pages[J].Intelligent Automation & Soft Computing,2023,36(1):183-192.

[13] 叶小榕,邵晴.基于 Spark 的大规模社交网络社区发现原型系统[J].科技导报,2018,36(23):93-101.
YE X R,SHAO Q.A large scale social networking community detection prototype system based on Spark[J].Science & Technology Review,2018,36(23):93-101.

[14] 杨晶.基于数据的风电机组整机故障诊断研究[D].上海:上海交通大学,2020.

[15] 邓英杰.基于线性判别分析的公理模糊集分类模型研究[D].大连:大连理工大学,2020.

[16] 石凤贵.基于 jieba 中文分词的中文文本语料预处理模块实现[J].电脑知识与技术,2020,16(14):248-251.

[17] 陈娟,孙琪.惠民公共政策出台缘何遭"冷遇"?:基于新浪微博平台三孩政策转发评论的数据分析[J].河南师范大学学报(哲学社会科学版),2023,50(5):70-75.
CHEN J,SUN Q.Why Did the Introduction of Public Policies Benefiting the People Receive a Cold Shoulder?:Data Analysis Based on Comments Forwarded by Sina Weibo's Three-child Policy[J].Journal of Henan Normal University(Philosophy & Social Sciences),2023,50(5):70-75.

A focused crawler algorithm based on Naive Bayes model and PageRank on power grid accidents

Xie Shuyong[1], Liu Zhiliang[2]

(1. Huizhou Power Supply Bureau, Guangdong Power Grid Co. Ltd., Huizhou 516000, China;
2. China Southern Power Grid, Guangzhou 510000, China)

Abstract: In order to reduce the number of personal safety accidents in the power grid, it is necessary to construct and analyze multi-dimensional data of accidents to build precise early warning models by using data mining techniques. One of the challenging problems is to automate the collection of accident data in large-scale websites. In this paper, we propose a focused crawler algorithm that combines Naive Bayes model and PageRank algorithm. First, by adopting the Chinese text segmentation method and setting keyword frequency, data are preprocessed. After feature selection, a Naive Bayesian classification model is constructed and trained, so as to achieve a significant increase in the classification accuracy of power grid accidents. Then, the PageRank algorithm is used to sort the topic relevance of the accurately classified web pages, which effectively avoids the problem of topic drift that common crawler methods often suffer from. The experimental results show that the page harvesting rate of the proposed algorithm is higher than that of using the Naive Bayesian classifier or the PageRank algorithm alone, regardless of the same time budget or the same number of searched pages. Thus, this method is capable of crawling information about power grid accidents more efficiently and accurately among a large number of web pages.

Keywords: power grid accident; personal safety; Naive Bayes model; PageRank algorithm; focused crawler

[责任编校 杨浦 陈留院]