一种面向科技项目文本的相似度度量方法*

赵晓平[1],马 文[1],刘雪萍[2],陈 达[2]
(1. 云南电网有限责任公司 信息中心,云南 昆明 650011;2. 云南云电同方科技有限公司,云南 昆明 650220)

摘 要:现有的文本相似度度量方法主要采用TF-IDF方法,把文本建模为词频向量,但未考虑文本的结构特征。现将文本的结构特征和TF-IDF方法进行融合,提出了一种面向科技项目文本的相似度度量方法。该方法首先对文本进行预处理,其次根据文本的结构特征提取模块文本,然后使用TF-IDF方法提取每个模块文本的TOP-N关键词,作为模块文本的特征向量表示,最后使用余弦聚类计算文本的相似度。实验结果表明,在电力行业的科技项目文档数据集上,所提方法优于TF-IDF方法。

关键词:文本相似度;TF-IDF;文本聚类;自然语言处理
中图分类号:TP311 **文献标识码**:A **DOI**:10.16157/j.issn.0258-7998.191420

中文引用格式:赵晓平,马文,刘雪萍,等. 一种面向科技项目文本的相似度度量方法[J].电子技术应用,2020,46(5):31-34,39.
英文引用格式:Zhao Xiaoping,Ma Wen,Liu Xueping,et al. A similarity measurement method for science and technology project text[J]. Application of Electronic Technique,2020,46(5):31-34,39.

A similarity measurement method for science and technology project text

Zhao Xiaoping[1],Ma Wen[1],Liu Xueping[2],Chen Da[2]
(1.Information Center,Yunnan Power Grid Co.,Ltd.,Kunming 650011,China;
2.Yunnan Yundian Tongfang Technology Co.,Ltd.,Kunming 650220,China)

Abstract: Existing text similarity measurements often use the TF-IDF method to model texts as term frequency vectors without considering the structural features of texts. This paper combines the structural features of texts with the TF-IDF method and proposes a text similarity measurement for science and technology project texts. This approach firstly pre-processes a text and extracts module texts according to its structural features. After applying the TF-IDF method to these extracted module texts, this method extracts the top keywords of each module text, obtains its feature vector representation, and finally uses cosine formula to calculate the similarity of two texts. By comparing with the TF-IDF method, experimental results show that the proposed method can promote the evaluation metrics of F-measure.

Key words: text similarity;TF-IDF;text clustering;natural language process

0 引言

文本相似度度量是指将文本看成一组词的集合体,分析每个词在文本中出现的次数以及在整个文本集合中出现次数,进而利用这些词频信息将文本建模为一个向量,并利用向量间的余弦距离等计算文本之间的相似度[1-2]。

文本相似度度量被广泛应用于许多领域,例如:信息检索领域[3-4]、文本分类[5-8]、文本摘要的自动生成[9-10]、文本的查重检测[11-12]。本文关注的是在电力行业的科技项目查重中应用文本相似度度量。

现有的TF-IDF[13-15]方法主要将文本建模为词频向量,再使用余弦相似度来计算两个文本间的相似度。但是对于多数文本而言,这种采用词频向量模型的方法需要将文本表示为词项数目与文本数目大致相当的矩阵,矩阵中的行列向量都有着非常高的维度并且是极度稀疏的,从而最终导致非常低效的计算[1,16]。此外,这种方法也忽略了文本的的结构特征。

针对上述问题,本文提出一种既考虑了文本的结构特征,又能有效降低文本表示模型维度的文本相似度度量方法。给定两个文本,通过文本所提方法能够高效、准确地计算出两者间的相似度,为电力行业科技项目的查重提供有效支撑。

1 相关工作

TF-IDF方法的核心是将文本表示为n个关键词组成的向量[13-15]。其基本思想是:将词频和逆向文本频率相结合,当某个关键词在一篇文本中出现的频率(词频)越高而在其他文本中出现的次数(逆文本频率)越少,那么这个词对这篇文本的区分程度就越高,则该词的权重

*基金项目:国家自然科学基金项目(61702442)

值就应该越大。

基于上述思想，通常采用下述 3 个步骤来计算每个关键词的 TF-IDF 值。

(1)计算词频(Term Frequency，TF)。词频是指该词在文本中出现的频率，计算公式为：

$$TF_{ij}=\frac{n_{ij}}{\sum_k n_{k,j}} \tag{1}$$

其中，TF_{ij} 表示词 i 在文本 j 中的词频，n_{ij} 表示词 i 在文本 j 中出现的次数，$\sum_k n_{k,j}$ 表示文本 j 中包含的单词总数。

(2)计算逆文本频率(Inverse Document Frequency，IDF)。逆文本频率是指语料库中的文本总数除以包含该词的文本数，计算公式为：

$$IDF_i=\log\frac{|N|}{N(i)+1} \tag{2}$$

其中，IDF_i 表示词 i 在的逆文本词频，$|N|$ 表示语料库中文本总数，$N(i)$ 表示包含单词 i 的文本数。分母之所以要加 1，是为了避免分母为 0，即所有文本都不包含该词的情况。

(3)计算词的 TF-IDF 值，公式为：

$$IF-IDF_i=TF_{ij}\times IDF_i \tag{3}$$

TF-IDF 算法的优点是简单快速，缺点是忽略了词的语义特征，也未考虑文档的结构特征。

2 本文方法

图 1 给出了本文所提方法的概述，共包含 5 个步骤：文本预处理、模块文本提取、关键词提取、文本表示、文本相似度计算。

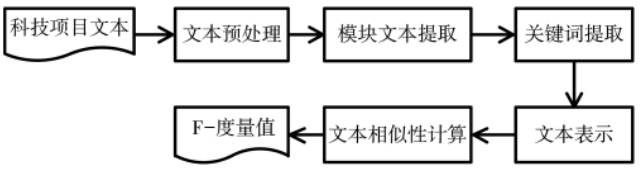

图 1 本文方法概述

2.1 文本预处理

在对文本分词后、关键词提取之前，需要对文本进行预处理，具体处理步骤如下：

(1)采用停用词列表，过滤掉文本中对应于停用词列表中的词项；

(2)对词项进行词性分析，因为最能表征科技项目文本含义的是名词和动词。对于形容词和副词，可以忽略这些词项，以尽可能降低词频向量模型的维度。

2.2 模块文本提取

电力行业领域的科技项目文本一般具有以下几个模块部分：项目名称、项目摘要、目的和意义、项目研究的背景、研究基础和条件、研究内容与实施方案、预期目标和成果形式，如图 2 所示。需要注意的是，各模块文本的长度不一且关注点不同，对项目科技文本进行相似度度量时，各模块文本所占权重也应该有所区别。

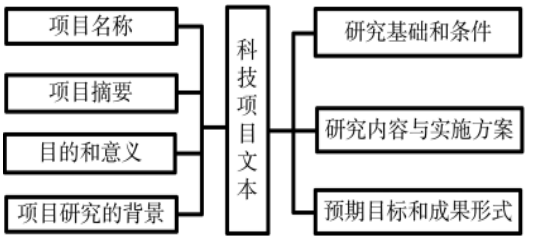

图 2 科技项目文本的结构

此外，科技项目文本中还有其他模块部分：配套科技应用示范项目落实情况，以及项目经费预算、经费预算详细说明、有关证明文件等内容，由于这些内容对检测文本相似度影响不大，因此本文主要提取图 2 所示模块文本作为相似度度量的主要对象。这样做的好处是可以尽量减少文本的长度，从而有效减少文本表示的维度。

2.3 关键词提取

文本特征词是指提取能够代替文本特征的词语。具体而言，针对每个模块文本，计算文本中各词项的 TF-IDF 值，并将各词项计算得到的 TF-IDF 值表示为一个向量。通常，这样计算得到的文本特征向量是高维且稀疏的。

为尽可能减少文本特征向量的维度，本文把每一个模块文本中计算得到的各词项的 TF-IDF 值进行排序，从中选取 TF-IDF 值 TOP-N 靠前的词项作为文本的特征词，其中，N 为百分比。与 TF-IDF 方法相比，本文所提方法的文本特征向量表示维度会降低 1-N，尽可能地减少文本特征向量的维度。

2.4 文本表示

针对电力行业科技项目文本的结构特征，本节将提出一种融合结构特征和 TF-IDF 方法的文本表示方法，将文本表示为特征和权重构成的矩阵，如表 1 所示。

表 1 文本的矩阵表示

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
t_1	w_{11}	w_{12}	w_{13}	w_{14}	w_{15}	w_{16}	w_{17}
t_2	w_{21}	w_{22}	w_{23}	w_{24}	w_{25}	w_{26}	w_{27}
…	…	…	…	…	…	…	…
t_n	w_{n1}	w_{n2}	w_{n3}	w_{n4}	w_{n5}	w_{n6}	w_{n7}

其基本思想是：首先，将一篇科技项目文本 Doc 表示为模块文本的集合(参见式(4))，每个模块文本对应项目文本中的某个结构，如项目名称、项目分类等；其次，使用 2.3 节所提方法，提取每个模块文本的 TOP-N 靠前的关键词；最后，将模板文本表示为式(5)，获得模块文本的特征向量表示。

$$\mathbf{Doc}=(d_1,d_2,d_3,d_4,d_5,d_6,d_7) \tag{4}$$

其中，d_1 表示项目名称，d_2 表示项目分类，d_3 表示项目摘要，d_4 表示目的和意义，d_5 表示研究基础和条件，d_6 表示研究内容与实施方案，d_7 表示预期目标和成果形式。

$$\boldsymbol{d}_i=[w_{1i},w_{2i},\cdots,w_{ni}] \tag{5}$$

其中，$\boldsymbol{d}_i$ 为模块文本，w_{ji} 为模块文本 $\boldsymbol{d}_i$ 对应的关键词 t_j 的 TF-IDF 值。

上述科技项目文本的表示充分利用了文本的结构特征，并有机地结合了 TF-IDF 方法，可以有效避免直接使用 TF-IDF 方法表示科技项目文本所带来的高维稀疏矩阵问题。

2.5 文本相似度计算

在计算得到文本的特征向量后，计算文本的相似度就可以转换为计算模块文本的相似度(参见式(6))，而计算模块文本的相似度进一步可转换为计算关键词向量间的相似度(参见式(7))。

$$\begin{aligned}\mathrm{TextSim}(\mathbf{Doc}_1,\mathbf{Doc}_2)=&\alpha_1\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_1,\mathbf{Doc}_2.\boldsymbol{d}_1)+\\&\alpha_2\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_2,\mathbf{Doc}_2.\boldsymbol{d}_2)+\alpha_3\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_3,\mathbf{Doc}_2.\boldsymbol{d}_3)+\\&\alpha_4\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_4,\mathbf{Doc}_2.\boldsymbol{d}_4)+\alpha_5\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_5,\mathbf{Doc}_2.\boldsymbol{d}_5)+\\&\alpha_6\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_6,\mathbf{Doc}_2.\boldsymbol{d}_6)+\alpha_7\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_7,\mathbf{Doc}_2.\boldsymbol{d}_7)\end{aligned} \tag{6}$$

其中，$\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_1,\mathbf{Doc}_2.\boldsymbol{d}_1)$ 表示文本 $\mathbf{Doc}_1$ 与文本 $\mathbf{Doc}_2$ 的项目名称相似度，$\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_2,\mathbf{Doc}_2.\boldsymbol{d}_2)$ 表示文本 $\mathbf{Doc}_1$ 与文本 $\mathbf{Doc}_2$ 的项目摘要相似度，$\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_3,\mathbf{Doc}_2.\boldsymbol{d}_3)$ 表示文本 $\mathbf{Doc}_1$ 与文本 $\mathbf{Doc}_2$ 的目的和意义相似度，$\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_4,\mathbf{Doc}_2.\boldsymbol{d}_4)$ 表示文本 $\mathbf{Doc}_1$ 与文本 $\mathbf{Doc}_2$ 的项目研究背景的相似度，$\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_5,\mathbf{Doc}_2.\boldsymbol{d}_5)$ 表示文本 $\mathbf{Doc}_1$ 与文本 $\mathbf{Doc}_2$ 的项目研究基础和条件的相似度，$\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_6,\mathbf{Doc}_2.\boldsymbol{d}_6)$ 表示文本 $\mathbf{Doc}_1$ 与文本 $\mathbf{Doc}_2$ 的项目内容与实施方案的相似度，$\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_7,\mathbf{Doc}_2.\boldsymbol{d}_7)$ 表示文本 $\mathbf{Doc}_1$ 与文本 $\mathbf{Doc}_2$ 的预期目标和成果形式的相似度；α_i 为模块文本 $\boldsymbol{d}_i$ 的权重值。

$$\mathrm{VecSim}(\mathbf{Doc}_1.\boldsymbol{d}_1,\mathbf{Doc}_2.\boldsymbol{d}_1)=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{7}$$

此外，由于科技项目文本中各个模块文本对于文本相似度度量所占的贡献比例不同。这比例不仅与模块文本的文字多少有关，还与抄袭剽窃最容易出现的概率有关。因此，需要对文本中各模块文本的权值进行分配，需要遵循式(8)的约束。

$$\sum_{1}^{7}\alpha_i=1 \tag{8}$$

3 实验

3.1 数据集及预处理

本文从电力行业的科技项目管理系统中提取出了900个科技项目文本作为入库数据集。数据集中每个本文都划分为3个类别中一个，分别是：基础研究、前瞻性研究和应用性研究。表2给出了入库数据集中每个类别的文本数量。其次，采用人工方式新造了10个抄袭科技项目文本组成测试数据集。表3给出了测试数据集中每个类别文本的数量。实验环境是：2核 CPU，2.50 GHz，8 GB 内存，操作系统 Windows 10，Python 3.7，开发工具 Spyder 3.2.4。

表2 入库数据集中每个类别的文本数量

类别	基础性研究	前瞻性研究	应用性研究
数量	200	300	400

表3 测试数据集中每个类别的文本数量

类别	基础性研究	前瞻性研究	应用性研究
数量	3	3	4

进一步对每个科技项目文本做如下处理：

(1)从每个科技项目文本中提取每个模块文本；

(2)使用 Python 语言的工具包 jieba 分词组件，对每个模块文本做分词和词性标注。由于电力行业涉及一些专用术语，因此本文在分词过程中加入了2 000多个专用术语组成的电力专用词典。表4显示了有无电力专用术语“空开”的分词结果；

(3)使用停用词表过滤停用词。本文在综合哈工大停用词词表、四川大学机器学习智能实验室停用词词表、百度停用词词表的基础上，建立了一个将近2 000个单词的停用词词表。

表4 分词结果对比

	分词结果
未加电力专用词典	客户 反映 空 开合 上无电，请 查处
加入电力专用词典	客户 反映 空开 合上 无电，请 查处

3.2 评价指标及参数设置

实验采用 F-度量值作为评价指标，用来衡量文本相似度。F-度量值是信息检索中一种组合查准率和召回率指标的平衡指标[17-19]。具体而言，文本相似度可以通过比较每一篇文本聚类后是否被划分为正确的类别以及同一类别下是否包含了正确类别的文本来进行衡量。

设 m_i 是类别 i 的文本数量，m_j 是聚类 j 的文本数量，m_{ij} 是聚类 j 中属于类别 i 的文本数目，则查准率 $P(i,j)$ 和查全率 $R(i,j)$ 可以分别定义如下：

$$P(i,j)=\frac{m_{ij}}{m_j} \tag{9}$$

$$R(i,j)=\frac{m_{ij}}{m_i} \tag{10}$$

对应的 F-度量值 $F(i,j)$ 定义为：

$$F(i,j)=\frac{2\times P(i,j)\times R(i,j)}{P(i,j)+R(i,j)} \tag{11}$$

全局聚类的 F-度量值定义为：

$$F=\sum_i\frac{m_i}{m}\max_j(F(i,j)) \tag{12}$$

其中，m 是数据集中文本总的数量。通常，F 越大，聚类效果越好，文本相似度度量也越好。

为了直观地将本文所提方法(定义为方法1)与基于TF-IDF算法的方法(定义为方法2)进行聚类效果的比较,本文采用经典聚类算法K-means,其中K值均设置为3。

实验首先要确定每个模块文本中TOP-N的值。图3给出了选取不同比例的TOP-N关键词在使用K-means聚类算法进行聚类条件下的实验结果。实验结果表明,TOP-N的取值设置为30%,聚类效果较为理想。如果低于比例,聚类效果都不太理想;如果高于这个比例(例如50%)也可以得到不错的聚类效果,但是增加了文本特征向量表示的维度。

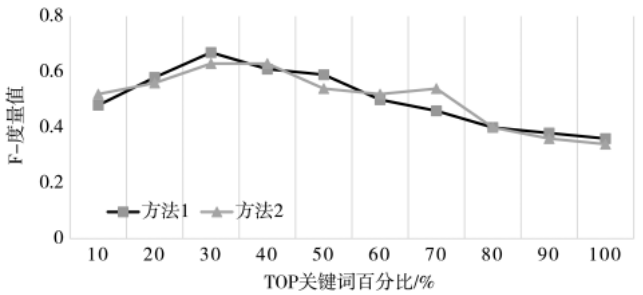

图3 TOP-N关键词百分比对聚类效果的影响

最后,需要确定各个模块文本的权重值。为此,在测试数据集上,本文通过比较模块文本在不同权值下得到的F-度量值选择权重值。最终,科技项目文本中各模块文本的权值设置如表5所示。

表5 模块文本的权重值

模块文本	d_1	d_2	d_3	d_4	d_5	d_6	d_7
权值	0.05	0.1	0.1	0.1	0.05	0.5	0.1

3.3 实验结果

在设定TOP-N为30%、权值如表5所示的情况下,采用本文所提方法与方法2在入库数据集上和测试数据集上进行文本相似度度量,对比结果如图4和图5所示。从图中可以看出,相比方法2,本文所提方法在两个数据集上具有更好的F-度量值,即聚类效果更好。这表明在对文本进行相似度度量上具有更好的效果。

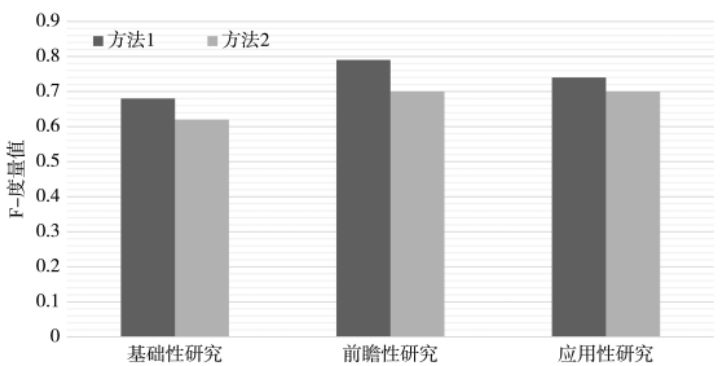

图4 入库数据集上的F-度量值对比

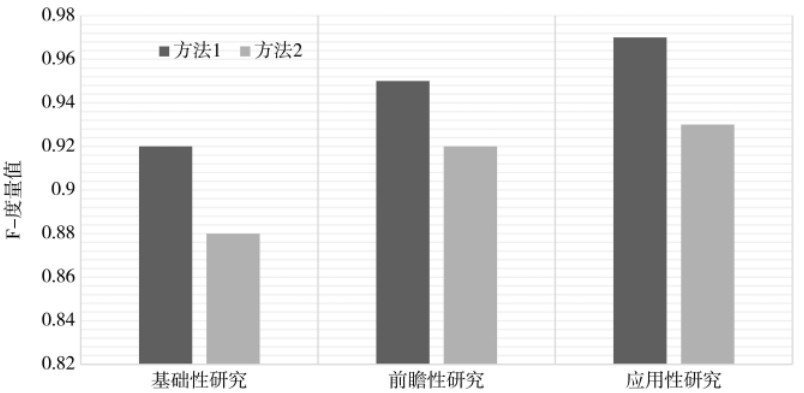

图5 测试数据集上的F-度量值对比

4 结论

本文融合文本结构特征和TF-IDF,提出了一种面向科技项目文本的相似度度量方法。与传统的TF-IDF方法相比,先通过提取文本中的模块文本,再从模块文本中选取TOP-N靠前的关键词作为文本的特征向量表示,可以有效降低TF-IDF方法带来的高维向量表示问题。实验结果表明这种方法是有效的。

本文的后续研究将考虑把文本的结构特征和文本的分布式表示进行融合,更好地提取文本的语义,进一步提高文本相似度度量的精度。

参考文献

[1] 黄承慧,印鉴,侯昉.一种结合词项语义信息和TF-IDF方法的文本相似度量方法[J].计算机学报,2011(5):98-106.

[2] ALTINEL B,GANIZ M.Semantic text classification:a survey of past and recent advances[J].Information Processing and Management,2018,54(6):1129-1153.

[3] HANNECH A,ADDA M,MCHEICK H.Recommendation model based on a contextual similarity measure[C].15th IEEE International Conference on Machine Learning and Applications(ICMLA),2016:394-401.

[4] RAJALAKSHMI A,SHAHNASSER H.Internet of Things using Node-Red and Alexa[C].17th International Symposium on Communications and Information Technologies(ISCIT),2017:1-4.

[5] 唐庄,王志舒,周爱,等.面向文本分类的transformer-capsule集成模型[J/OL].计算机工程与应用:1-7[2019-12-27].http://kns.cnki.net/kcms/detail/11.2127.TP.20191219.0859.002.html.

[6] JIN R,LU L F,LEE J,et al.Multi-representational convolutional neural networks for text classification[J].Computational Intelligence,2019,35(3):599-609.

[7] 曾祥坤,张俊辉,石拓,等.基于主题提取模型的交通违法行为文本数据的挖掘[J].电子技术应用,2019,45(6):41-45.

[8] 殷晓雨,阿力木江·艾沙,库尔班·吾布力.基于卷积

(下转第39页)

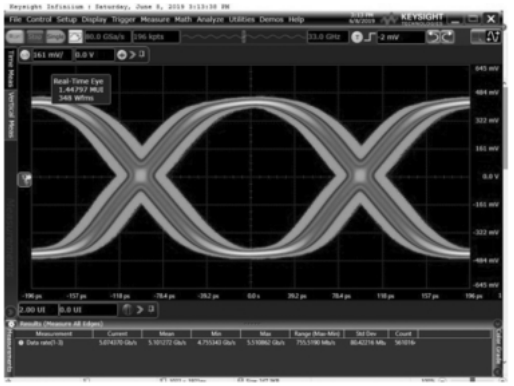

图 8 锁相环输出 2.5 GHz 时钟眼图

5 结论

本文实现了一种基于 Ring-VCO 结构的宽频带低抖动锁相环，采用与锁相环锁定频率强相关的环路带宽调整方法，即利用全局参考调节电路中比较器模块将锁定控制电压与参考电压比较来改变各模块电流，根据不同锁定频率调整环路参数，大大缩短了锁定时间，同时利用四级差分环形振荡器和占空比调整电路的差分对称结构，并且对敏感模块电源均采用 LDO 单独供电，有效地降低了电源噪声。电路采用 40 nm CMOS 工艺实现，测试结果表明锁相环锁定时间大大缩短，输出频率为 1.062 5 GHz~5 GHz，在时钟频率5 GHz 下眼图质量良好，时钟抖动为 39.6 ps。

参考文献

[1] ABEDI M, HASANI J Y. A fast locking phase-locked loop with low reference spur[C]. Iranian Conference on Electrical Engineering, 2018: 92-97.

[2] HUR C, CHOI Y S, CHOI H H, et al. A low jitter phase-lock loop based on a new adaptive bandwidth controller[J]. IEEE Asia-Pacific Conference on Circuit and Systems, 2004: 421-424.

[3] AMOURAN M, WHATELY M. A novel switched-capacitor-filter based low area and fast-locking PLL[C]. 2015 Custom Integrated Circuit Conference (CICC), 2015.

[4] AMOURAN M, KRISHNEGOWDA S, WHATELY M. A novel OTA-based fast lock PLL[C]. Proceeding of the IEEE 2013 Custom Integrated Circuit Conference, 2013.

[5] LOKE A L S, BARNES R K, WEE T T, et al. A versatile 90-nm CMOS charge-pump PLL for SerDes transmitter clocking[J]. IEEE Journal of Solid-State Circuits, 2006, 41(8): 1894-1907.

[6] Song Ying, Wang Yuan, Jia Song, et al. An adaptive-bandwidth CMOS PLL with low jitter and a wide tuning range[J]. Journal of Semiconductors, 2008, 29(5): 908-912.

[7] 丁志钊. 基于 PLL 频率合成器锁相环的降噪技术[J]. 电子测量技术, 2009, 32(5): 44-46.

[8] MOZHGAN M, Kong Yang C K. Jitter optimization based on phase-locked loop design parameters[C]. 2002 IEEE International Solid-State Circuits Conference, 2002.

[9] PIALIS T, PHANG K. Analysis of timing jitter in ring oscillators due to power supply noise[C]. Proceedings of the 2003 International Symposium on Circuits and Systems, 2003.

(收稿日期：2019-12-06)

作者简介：

刘颖(1988-)，女，硕士，工程师，主要研究方向：高速串行接口电路设计。

田泽(1965-)，男，博士，研究员，主要研究方向：集成电路设计、嵌入式系统开发。

吕俊盛(1986-)，男，博士，高级工程师，主要研究方向：高速串行接口电路设计。

(上接第 34 页)

递归模型的文本分类研究[J]. 电子技术应用, 2019, 45(10): 29-32, 36.

[9] 李文鹏，赵俊峰，谢冰. 基于 LDA 的软件代码主题摘要自动生成方法[J]. 计算机科学, 2017, 44(4): 35-38.

[10] 翟娟，汤震浩，李彬，等. 常用循环摘要的自动生成方法及其应用[J]. 软件学报, 2017, 28(5): 1051-1069.

[11] 李成龙，杨冬菊，韩燕波. 基于分词矩阵模型的模糊匹配查重算法研究[J]. 计算机科学, 2017, 44(S2): 55-60.

[12] 李成龙. 云环境下支持模糊匹配的文本查重技术研究与实现[D]. 北京：北方工业大学, 2018.

[13] 董蕊芳，柳长安，杨国田. 一种基于改进 TF-IDF 的 SLAM 回环检测算法[J]. 东南大学学报(自然科学版), 2019, 49(2): 251-258.

[14] 王杨，王非凡，张舒宜，等. 基于 TF-IDF 和改进 BP 神经网络的社交平台垃圾文本过滤[J]. 计算机系统应用, 2019, 28(3): 126-132.

[15] 叶雪梅，毛雪岷，夏锦春，等. 文本分类 TF-IDF 算法的改进研究[J]. 计算机工程与应用, 2019, 55(2): 104-109.

[16] 那海洋，杨庚，束晓伟. 基于 B~+树的多关键字密文排序检索方法[J]. 计算机科学, 2017, 44(1): 149-154.

[17] 杨传慧，吉根林，章志刚. AP 算法在图像聚类中的应用研究[J]. 计算机与数字工程, 2012(10): 125-127.

[18] 代飞，赵文卓，杨云，等. BPMN2.0 编排的形式语义和分析[J]. 软件学报, 2018, 29(4): 1094-1114.

[19] 代飞，陈凤强，莫启，等. 一种保持编排与参与者间行为一致的映射方法[J]. 软件学报, 2018, 29(5): 1451-1470.

(收稿日期：2019-12-27)

作者简介：

赵晓平(1984-)，男，硕士研究生，高级工程师，主要研究方向：自然语言处理、电网信息化管理和科技创新管理。

马文(1981-)，男，本科，高级工程师，主要研究方向：自然语言处理、信息化管理和大数据研究及应用。

刘雪萍(1990-)，女，本科，工程师，主要研究方向：自然语言处理、企业信息化咨询设计。

版权声明