文章编号:2096-1472(2023)08-0006-05
DOI:10.19644/j.cnki.issn2096-1472.2023.008.002

基于 TF-IDF 和 TextRank 结合的中文文本关键词提取方法——以体育新闻为例

兰晓芳[1]，刘 卓[2]，许志豪[1]，肖 毅[2]

(1.湖南农业大学东方科技学院，湖南 长沙 410128；
2.湖南农业大学信息与智能科学技术学院，湖南 长沙 410128)

✉ lanxf@stu.hunau.edu.cn; fetty_max@163.com; guapideyouxiang@stu.hunau.edu.cn; xiaoyi@hunau.edu.cn

摘 要:利用文本挖掘技术进行体育热点分析,可以为体育领域的发展提供更多有用的信息。文中提出了一种基于 TF-IDF(Term Frequency-Inverse Document Frequency,词频-逆文档频率)和 TextRank(文本排序)的中文文本关键词提取方法,该方法首先采用分词、去除停用词等对文本进行预处理;其次使用 TF-IDF 算法计算每个词的重要性并进行归一化处理,同时使用 TextRank 算法权衡单词之间的关系并计算每个单词的得分以进行归一化处理;最后将 TF-IDF 值和 TextRank 得分进行加权和得到每个词的综合权重值,最终获得权重值最高的 N 个关键词。应用 TF-IDF 和 TextRank 结合的方法在 F_1 值上选择 5 个关键词时取得了更好的结果,相较于只使用 TF-IDF 方法或 TextRank 方法,其关键词提取准确率分别提高约 40%和 32%。该方法有效提高了关键词提取的准确性和提取效率。

关键词:TF-IDF;TextRank;体育新闻;关键词提取
中图分类号:TP391.1 **文献标志码**:A

A Chinese Text Keyword Extraction Method Based on the Combination of TF-IDF and TextRank —— A Case Study of Sports News

LAN Xiaofang[1], LIU Zhuo[2], XU Zhihao[1], XIAO Yi[2]

(1.*Oriental College of Science and Technology, Hunan Agricultural University, Changsha* 410128, *China*;
2.*College of Information and Intelligent, Hunan Agricultural University, Changsha* 410128, *China*)

✉ lanxf@stu.hunau.edu.cn; fetty_max@163.com; guapideyouxiang@stu.hunau.edu.cn; xiaoyi@hunau.edu.cn

Abstract: Using text mining techniques for sports hot topic analysis can provide more useful information for the development of the sports field. This paper proposes a method for extracting Chinese text keywords based on TF-IDF and TextRank. This method preprocesses the text by tokenizing and removing stop words, and then calculates the importance of each word using the TF-IDF algorithm and normalizes the values. Fianlly, the TextRank algorithm is used to weigh the relationships between words and calculate scores for each word, which are also normalized. Finally, the TF-IDF values and TextRank scores are weighted to obtain a comprehensive weight for each word, ultimately obtaining the N keywords with the highest weight value. The method of combining TF-IDF and TextRank achieved better results when selecting 5 keywords on F_1 value, and compared to using only TF-IDF method or TextRank method, the accuracy of keyword extraction increases by about 40% and 32%, respectively. This method effectively improves the accuracy and efficiency of keyword extraction.

Key words: TF-IDF; TextRank; sports news; keyword extraction

0 引言(Introduction)

随着互联网的发展,人们可以方便地在互联网上获取各种类型的文本数据,而提取中文文本新闻的关键字有重大意义,新闻的关键字可以作为新闻标题和摘要的一部分出现,吸引更

收稿日期:2023-01-17

多读者点击阅读,进而促进新闻的传播和推广;还可以使读者更快速地了解文章的主要内容和重点,提高阅读效率。同时,它可以作为搜索引擎的关键词,提高搜索结果的精准度和效果[1]。此外,通过对新闻文本的关键字进行提取和分析,可以得到读者关注的相关信息,提高广告投放的精准性和效果[2]。近年来,基于 TF-IDF 和 TextRank 的关键词提取算法在中文文本领域得到了广泛应用。然而,由于中文语言的复杂性,传统的 TF-IDF 和 TextRank 算法在中文文本的关键词提取中存在一定的局限性[3]。因此本文提出了一种基于 TF-IDF 和 TextRank 的中文文本的体育新闻关键词提取方法,可以提高关键词提取的准确性和覆盖率。

1　相关工作(Related work)

关键词提取是一个广泛的研究领域,已经有许多算法被提出。中文文本的关键词提取与英文文本不同,主要因为中文词汇具有复杂性和多义性。因此,中文文本的关键词提取需要考虑词汇的语义、词频、文本结构等多方面因素。其中,基于频率的 TF-IDF 算法是最常用的一种方法,它通过计算词频和文档频率衡量词语的重要性。TextRank 算法是一种基于图的排序算法,它通过对文本中词语之间的关系进行建模,计算每个词语的重要性。这两种算法已经被证明在关键词提取任务中取得了良好的效果[4]。然而,这两种算法各自存在一些缺陷。TF-IDF 算法只考虑了单词的频率信息,忽略了单词之间的关系。TextRank 算法考虑了单词之间的关系,但是它没有考虑单词的频率信息。因此,结合应用两种算法可以克服它们各自的缺点,提高关键词提取的准确性。

本文方法首先对文本进行预处理,包括分词、去除停用词等操作;其次使用 TF-IDF 算法计算每个词的重要性并进行归一化处理,同时使用 TextRank 算法考虑单词之间的关系,计算每个单词的得分并进行归一化处理;最后,将 TF-IDF 值和 TextRank 得分进行加权和得到每个词的综合权重值,按照权重值从大到小排序后选择权重值最高的前 N 个单词作为关键词。关键词提取步骤如图 1 所示。

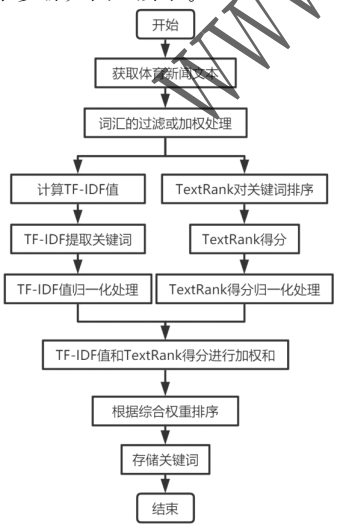

图 1　TF-IDF 和 TextRank 结合的关键词提取步骤图

Fig. 1 Diagram of keyword extraction steps using the method of combining TF-IDF and TextRank

2　方法实现(Method implementation)

2.1　数据集

为了评估本文所提方法的效果,使用来自清华大学自然语言处理实验室的 THUCNews 数据集中的 131 601 篇体育新闻数据集进行实验。数据集中都是 text 文件,为了方便数据处理与计算,将数据存入 MySQL 数据库中,数据集基本信息见表 1,数据库表设计见表 2。

表1　数据集基本信息表

Tab.1 Basic information table of the data set

数据来源	文档数/个	存储容量/MB	行格式	排序规则	引擎
THUCNews	131 601	379.24	Dynamic	utf8_general_ci	MyISAM

表2　数据库表设计

Tab.2 Design of the database tables

字段名	类型	长度/位	小数点	是否为空	虚拟	键	注释
tid	int	11	无	是	无	主键,自动递增	编号
tno	int	11	无	否	无	无	文件编号
title	varchar	255	无	否	无	无	标题
tcontent	text	65 536	无	否	无	无	内容

2.2　数据预处理

将文本从数据库中读取出来,使用 jieba.lcut()进行分词,同时使用百度停用词表过滤停用词等,方便后续处理。

(1)分词。使用分词工具(如 jieba)对给定的中文文本进行分词,将文本转化为词语序列。使用默认的精确模式 words = jieba.lcut(sentence)。虽然 Paddle 模式(飞桨模式)对机构团体名的解析更准确,但是对分词效果不大。使用 Paddle 模式非常耗时,性价比不高。通过实际测算,使用 Paddle 模式对 100 条语句进行分词的耗时,约是不使用 Paddle 模式的 103 倍,如表 3 所示。

表3　使用不同分词模式的耗时

Tab.3 Time consumption of different word segmentation patterns

模式	参数设置	100 条语句耗时/s
Paddle 模式	use_paddle=True	26.265 052 356 991 587
非 Paddle 模式	use_paddle=False	0.255 432 891 845 703 1

(2)去停用词。在进行新闻文本关键词提取前,需要做停用词处理,主要是为了去除一些无意义的高频词汇,如“的、是、了、而、和”等。这些词语出现的频率非常高,但它们本身并没有太多的语义信息,对于关键词提取没有太大的帮助。同时,去除这些无用的词汇也可以减少文本处理的时间和计算量。停用词处理的方法通常是通过建立一个停用词表,包含需要去除的无用词汇。在进行文本处理时,对于每一个词语都需要和停用词表中的词汇进行比对,如果该词语属于停用词,则将其去除,否则保留。这样可以去除一些无用的高频词汇,提高关键词提取的准确性和效率。本文对比三个常用的中文停用词表后,决定使用百度停用词表过滤停用词。停用词表适用类型见表 4。

表4 停用词表适用类型

Tab.4 Applicable types of stop word list

停用词表	文本类型
哈尔滨工业大学停用词表	文献期刊类文本
四川大学停用词表	邮件文献等类型
百度停用词表	新闻报道类文本

2.3 计算 TF-IDF 得分

TF-IDF 的中文名为“词频-逆文档频率”，是一种统计方法，用于评估一个词语在文档中的重要程度。由词频(Term Frequency，TF)和逆文档频率(Inverse Document Frequency，IDF)两个部分组成，它的核心思想是一个词语在一篇文档中出现的次数越多，同时在其他文档中出现的次数越少，那么就越能代表该文档[5]。

TF(词频)指的是某个词在一篇文档中出现的频率。TF越高，说明这个词在文档中出现的次数越多，越重要。

IDF(逆文档频率)指的是某个词在所有文档中出现的频率的倒数。如果一个词在所有文档中都频繁出现，那么它的 IDF 就会很低，说明这个词在区分文档时并没有太大的用处。相反，如果一个词只在少数文档中出现，那么它的 IDF 就会很高，说明这个词在区分文档时具有很大的作用。

综合考虑 TF 和 IDF，可以计算一个词的 TF-IDF 值，它越高就表示这个词在文档中越重要[6]。计算公式如下：

$$TF\text{-}IDF(w)=TF(w)\times IDF(w) \tag{1}$$

$TF(w)$表示给定词语 w 在文本中的出现次数，计算公式如下：

$$TF(w)=\frac{cout(w)}{|D_i|} \tag{2}$$

其中，$cout(w)$表示文本中词语 w 出现的次数，$|D_i|$ 表示文本 D_i 中所有词语的个数。

$IDF(w)$表示词语 w 的逆文档频率，计算公式如下：

$$IDF(w)=\lg\frac{N}{DF(w)} \tag{3}$$

其中，N 表示文本总数，$DF(w)$表示包含词语 w 的文本数。

2.4 计算 TextRank 相似度矩阵

TextRank 是一种基于图的排序算法，计算词语之间的相似度矩阵，可以用于文本关键词的提取[7]。对于给定的词语序列，构建一个无向图，其中每个词语对应一个节点，如果两个词语在文本中相邻，则在它们之间连一条边，并通过迭代计算每个节点(即词语)的权重，用于识别最重要的词语。使用 PageRank 算法对图中的节点进行排序，得到每个词语的得分[8]。TextRank 基于 PageRank 进行操作的流程如图 2 所示。

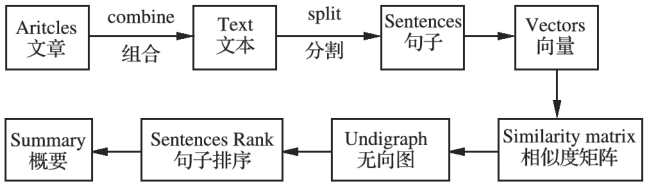

图 2 TextRank 基于 PageRank 操作流程图

Fig. 2 Operation flowchart of TextRank based on PageRank

在 TextRank 中，通过在文本中构建有向图表示文本中单词之间的关系，节点的重要性是根据它们在图中的连接情况和其他节点的连接情况确定的。对于关键词提取任务，TextRank 通过对排序后的节点进行截断，选取前几个作为关键词。

使用 TextRank 算法进行关键词提取的具体步骤如下。

(1)将文本表示为图：将文本中的每个单词表示为图中的节点，如果两个单词在文本中同时出现，则在它们之间连接一条边。

(2)计算每个节点的初始权重：为每个节点分配一个初始权重值，初始值设为 1。

(3)迭代计算每个节点的权重：迭代计算每个节点的权重，直至收敛迭代了 126 轮。PageRank 算法迭代公式如下：

$$PR(V_i)=(1-d)+d\times\sum_{j\in In(V_i)}\frac{1}{|Out(V_j)|}PR(V_j) \tag{4}$$

其中，$PR(V_i)$是词语 i 的重要性，$In(V_i)$是指向词语 i 的词语集合，$Out(V_j)$是从词语 j 指出去的词语集合，d 是阻尼系数。当指向网页 i 的个数为 0 时，公式(4)最右侧项为 0，阻尼系数避免了 $PR(V_i)$为 0。

(4)提取关键词：按照节点权重值从大到小的顺序提取关键词。

2.5 计算词语综合得分

将 TF-IDF 得分和 TextRank 相似度矩阵结合，计算每个词语的综合得分[8]。根据综合得分对词语进行排序，选取得分排名前 10 的词语作为关键词，采用如下公式计算词语得分：

$$score(\boldsymbol{w})=(1-alpha)\times tfidf(\boldsymbol{w})+alpha\times sum(sim(\boldsymbol{w},\boldsymbol{u})\times score(\boldsymbol{u})/sum(sim(\boldsymbol{u},\boldsymbol{v}))) \tag{5}$$

其中，$tfidf(\boldsymbol{w})$是词语 $\boldsymbol{w}$ 的 TF-IDF 得分，$alpha$ 是 TextRank 算法中的阻尼系数，这里取值为 0.85。$sim(\boldsymbol{w},\boldsymbol{u})$是词语 $\boldsymbol{w}$ 和词语 $\boldsymbol{u}$ 之间的相似度，使用余弦相似度等方法进行计算，有效平衡准确率和召回率。余弦相似度计算公式如下：

$$sim(\boldsymbol{w},\boldsymbol{u})=(\boldsymbol{w}\cdot\boldsymbol{u})/(\|\boldsymbol{w}\|\times\|\boldsymbol{u}\|) \tag{6}$$

其中，$(\boldsymbol{w}\cdot\boldsymbol{u})$表示词向量 $\boldsymbol{w}$ 和 $\boldsymbol{u}$ 的点积，$\|\boldsymbol{w}\|$ 和 $\|\boldsymbol{u}\|$ 分别表示词向量 $\boldsymbol{w}$ 和 $\boldsymbol{u}$ 的模。

3 结果分析(Result analysis)

实验中使用软件 PyCharm 2021.2，Runtime version 为 11.0.11+9-b1504.13 amd64，VM 为 OpenJDK 64-Bit Server VM by JetBrains s.r.o.，数据库软件为 Navicat Premium 15，计算机语言为 Python 3.10，以及 textrank4zh 0.3、jieba 0.42.1、re 2.2.1、pymysql 0.9.3 等包和模块。

将本文所提方法与传统的 TF-IDF 和 TextRank 算法进行对比，评估本文所提方法的准确率。

采用以下指标评估本文所提方法的准确率。

准确率 ACC(Accuracy)是用于评估分类模型性能的指标之一，它表示模型正确分类的样本数占总样本数的比例。计算公式如下：

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

其中，TP(True Positive)是真正例数，实际为正例，预测也为正例；FN(False Negative)是假负例数，实际为正例，预测为反例；FP(False Positive)是假正例数，实际为反例，预测为正例；TN(True Negative)是真负例数，实际为反例，预测也为反例。$TP-TN-FP-FN$ 类别区分见表 5。

表5　$TP-TN-FP-FN$ 类别区分

Tab.5 $TP-TN-FP-FN$ category differentiation

实际类别 \ 预测类别	正	反	总计
正	TP	FN	实际为正 $TP+FN$
反	FP	TN	实际为反 $FP+TN$
总计	预测为正 $TP+FP$	预测为反 $FN+TN$	$TP+TN+FP+FN$

精确率 P(Precision)是指提取出来的关键词中正确的个数与提取出来的关键词总数的比例，计算公式如下：

$$P=\frac{TP}{TP+FP} \tag{8}$$

召回率 R(Recall)是指提取出来的正确关键词的个数与正确关键词总数的比例，计算公式如下：

$$R=\frac{TP}{TP+FN} \tag{9}$$

F_1 值(F_1-score)是精确率和召回率的加权平均值，用来综合评估模型的性能，计算公式如下：

$$F_1=\frac{(\alpha^2+1)P\times R}{\alpha^2(P+R)} \tag{10}$$

当参数 $\alpha=1$ 时，即最常见的 F_1 如下：

$$F_1=\frac{2\times P\times R}{P+R} \tag{11}$$

使用 F_1 值作为评估指标比较 TF-IDF 和 TextRank 结合的方法与只使用 TF-IDF 和只使用 TextRank 的方法。实验中分别选择 3 个、5 个、7 个关键词进行分析。对比结果见表 6 和图 3 至图 5 所示。

表6　TF-IDF、TextRank 及两种方法结合后提取结果对比表

Tab.6 Extraction results comparison of TF-IDF, TextRank, and the combination of the two methods

算法	准确率(P 值)			召回率(R 值)			F_1 值		
	3 个	5 个	7 个	3 个	5 个	7 个	3 个	5 个	7 个
TF-IDF	0.523	0.492	0.465	0.352	0.369	0.396	0.421	0.422	0.428
TextRank	0.472	0.418	0.397	0.475	0.483	0.494	0.473	0.448	0.440
TF-IDF & TextRank	0.684	0.677	0.642	0.493	0.524	0.546	0.573	0.591	0.590

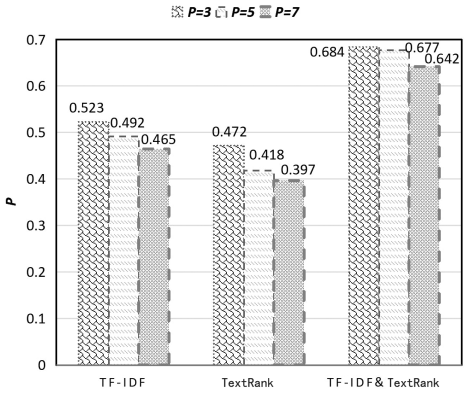

图 3　精确率 P 变化

Fig. 3 Changes of precision rate P

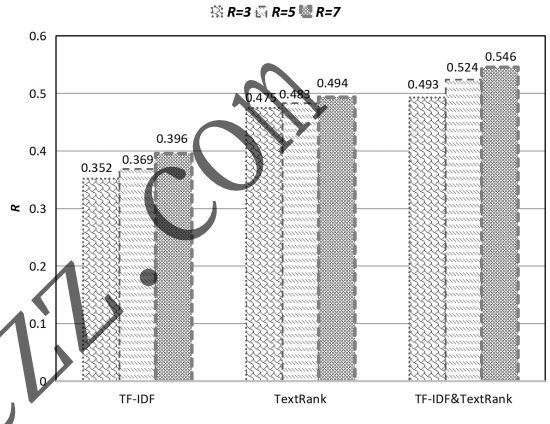

图 4　召回率 R 变化

Fig. 4 Changes of recall rate R

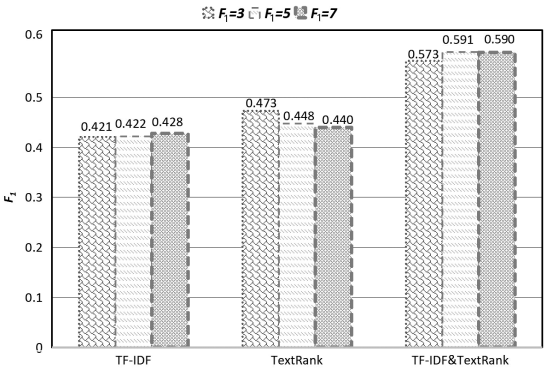

图 5　F_1 值变化

Fig. 5 Changes of F_1 value

4　结论(Conclusion)

本文提出了一种基于 TF-IDF 和 TextRank 的中文文本体育新闻的关键词提取方法。该方法结合了 TF-IDF 算法和 TextRank 算法的优点，提高了关键词提取的准确性。通过 TF-IDF 方法可以筛选出具有高重要性的单词，而 TextRank 方法则可以通过考虑单词之间的关系提高关键词的准确性。为验证该方法，使用体育新闻数据集进行实验，并与只使用 TF-IDF 和只使用 TextRank 进行提取关键词准确率结果的比

较。实验结果表明，TF-IDF 和 TextRank 结合的方法在 F_1 值上选择 5 个关键词时取得了更好的结果，相对于只使用 TF-IDF 方法准确性提高约 40%，相对于只使用 TextRank 方法准确性提高约 32%。对基于 TF-IDF 和 TextRank 的方法进行详细分析发现，使用 TextRank 算法可以捕捉单词之间的关系，有助于识别出一些相关的关键词，但是它也容易将一些不相关的单词也包含进来，而使用 TF-IDF 算法可以过滤一些常见的单词，提高了关键词的准确性。结合这两种算法可以更好地平衡准确性和召回率，提高关键词的提取效果。

基于 TF-IDF 和 TextRank 的方法还有一些改进的空间，如将词语的语义信息考虑进来，使用深度学习等方法进行建模等。在未来的研究中，可以继续改进基于 TF-IDF 和 TextRank 的方法，并在其他领域的关键词提取任务中应用。

参考文献(References)

[1] 孟庆麟. 我国新闻出版的热点关键词分析与发展对策究研究[D]. 大连：大连海事大学，2019.

[2] 蒋艳. 语料库方法在新闻传播研究中的发展应用分析[J]. 新闻研究导刊，2022，13(24)：23-26.

[3] 何传鹏，尹玲，黄勃，等. 基于 BERT 和 LightGBM 的文本关键词提取方法[J]. 电子科技，2023，36(3)：7-13.

[4] 张晓丽. 面向新闻领域的关键词提取方法研究及系统实现[D]. 太原：山西大学，2021.

[5] WANG Z H，WANG D，LI Q. Keyword extraction from scientific research projects based on SRP-TF-IDF[J]. Chinese Journal of Electronics，2021，30(4)：652-657.

[6] 张瑾. 基于改进 TF-IDF 算法的情报关键词提取方法[J]. 情报杂志，2014，33(4)：153-155.

[7] 赵占芳，刘鹏鹏，李雪山. 基于改进 TextRank 的铁路文献关键词抽取算法[J]. 北京交通大学学报，2021，45(2)：80-86.

[8] 李晨，赵燕清，于俊凤，等. 基于词向量与 TextRank 的政策文本关键词汇抽取方法研究[J]. 现代计算机，2023，29(2)：68-72.

作者简介：

兰晓芳(1998-)，女，本科生。研究领域：数据处理，推荐算法。

刘　卓(2002-)，男，本科生。研究领域：人工智能，数据处理。

许志豪(2001-)，男，本科生。研究领域：机器学习，数据处理。

肖　毅(1978-)，男，博士生，讲师。研究领域：数据处理，模式识别，本文通信作者。

(上接第 5 页)

参考文献(References)

[1] 臧志彭，解学芳. 中国特色元宇宙体系建设：理论构建与路径选择[J]. 南京社会科学，2022(10)：137-147，158.

[2] 苟尤钊，季雪庭，叶盈如，等. 元宇宙技术体系构建与展望[J]. 电子科技大学学报，2023，52(1)：74-84.

[3] RADOFF J. Building the metaverse and making it a place for everyone[EB/OL]. (2021-11-26)[2023-01-09]. https://www.metaverse.fm/jon-radoff.

[4] MYSTAKIDIS S. Metaverse[J]. Encyclopedia，2022，2(1)：486-497.

[5] 王文喜，周芳，万月亮，等. 元宇宙技术综述[J]. 工程科学学报，2022，44(4)：744-756.

[6] YANG Q L，ZHAO Y T，HUANG H W，et al. Fusing blockchain and AI with metaverse：a survey[EB/OL]. (2022-01-10)[2023-01-11]. https://arxiv.org/abs/2201.03201.

[7] 郭亚军，袁一鸣，李帅，等. 元宇宙场域下虚拟社区知识共享模式研究[J]. 情报理论与实践，2022，45(4)：52-57，40.

[8] GADEKALLU T R，HUYNH-THE T，WANG W Z，et al. Blockchain for the metaverse：a review[EB/OL]. (2022-03-18)[2023-01-14]. https://arxiv.org/abs/2203.09738.

[9] NGUYEN C T，HOANG D T，NGUYEN D N，et al. MetaChain：a novel blockchain-based framework for metaverse applications[EB/OL]. (2021-12-30)[2023-01-17]. https://arxiv.org/abs/2201.00759.

[10] 宋晓玲，刘勇，董景楠，等. 元宇宙中区块链的应用与展望[J]. 网络与信息安全学报，2022，8(4)：45-65.

[11] 沈鑫，裴庆祺，刘雪峰. 区块链技术综述[J]. 网络与信息安全学报，2016，2(11)：11-20.

[12] 靳世雄，张潇丹，葛敬国，等. 区块链共识算法研究综述[J]. 信息安全学报，2021，6(2)：85-100.

[13] 范吉立，李晓华，聂铁铮，等. 区块链系统中智能合约技术综述[J]. 计算机科学，2019，46(11)：1-10.

[14] 姚前，张大伟. 区块链系统中身份管理技术研究综述[J]. 软件学报，2021，32(7)：2260-2286.

[15] 李强，舒展翔，余祥，等. 区块链系统的认证机制研究[J]. 指挥与控制学报，2019，5(1)：1-17.

[16] 凯文·凯利. 失控：全人类的最终命运和结局[M]. 东西文库，译. 北京：新星出版社，2010：10-19.

[17] NAKAMOTO S. Bitcoin：a peer-to-peer electronic cash system[EB/OL]. (2008-10-31)[2022-12-11]. http://bitcoin.org/bitcoin.pdf.

作者简介：

王　迪(1988-)，男，硕士，工程师。研究领域：区块链，密码学。