基于自动文摘的作文素材标签提取方法研究

朱晓亮[1,2] 吴逸尘[1] 殷 姿[1]

[1](华中师范大学国家数字化学习工程技术研究中心 湖北 武汉 430079)

[2](华中师范大学教育大数据应用技术国家工程实验室 湖北 武汉 430079)

摘 要 作文素材在小学语文作文辅助中的作用不可忽视。但当前存在的作文素材数量繁多却普遍缺乏个性化的推荐模式,极易引起小学生知识过载现象。作文素材的非结构化数据特征使得计算机对其存取操作过于复杂,进而会对数据的有效组织产生阻碍。通过对作文语料的文本自动摘要处理来去除冗余、提取文本中心内容,可以改善作文辅助中的信息过载问题以及素材语料中存在的大量冗余,对文本标签的获取产生干扰问题。在完善作文标签定义的前提下,提出一种基于文本自动摘要的小学语文作文标签提取方法。

关键词 作文素材 文本自动摘要 标签抽取 TextRank

中图分类号 TP391.1 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.02.040

EXTRACTION OF COMPOSITION MATERIAL LABEL BASED ON AUTOMATIC ABSTRACTS

Zhu Xiaoliang[1,2] Wu Yichen[1] Yin Zi[1]

[1](*National Engineering Research Center for E-learning, Central China Normal University, Wuhan 430079, Hubei, China*)

[2](*National Engineering Laboratory for Educational Big Data, Central China Normal University, Wuhan 430079, Hubei, China*)

Abstract The role of composition materials in primary school Chinese composition auxiliary cannot be ignored. However, there is a general lack of personalized recommendation mode in current composition materials, which easily leads to the phenomenon of knowledge overload of primary school students. The unstructured data features of the composition materials make it too complicated for computer to access operations, which will hinder the effective organization of the data. In order to improve the information overload in composition assistance and the interference of a large amount of redundancy in materials corpus on the acquisition of textual tags, this paper removed redundancy and extracted textual center content by automatically textual abstracting. On the premise of perfecting the definition of composition label, this paper presented a method of extracting primary school composition labels based on automatic text abstract.

Keywords Composition material Automatic text abstract Tag extraction TextRank

0 引 言

语文课程标准[1]以及语文学科研究者均明确强调,在小学语文作文中要加强写作素材的积累,并使学生能结合自身实际加以利用[2]。然而,当前语文作文教学中提供的素材数量庞杂且缺乏组织,在小学生有限的认知能力之下,若不经过加工直接推送,极易造成认知过载以至于引发学习迷航。因此,在当前小学中高年级语文写作教学过程中,缺少有效的作文素材资源辅助,这就要求在小学语文学科作文领域的信息化建设中,把握好优秀作文素材的存储管理工作,并且能够为个性化的作文素材提供数据支撑。

对小学语文作文素材的存储,借助知识图谱[3]的

收稿日期:2018 - 09 - 04。国家重点研发计划项目(2018YFB1004500,2018YFB1004504);湖北省技术创新专项重大项目(2017AKA191);教育部人文社会科学研究规划基金项目(18YJAZH152)。**朱晓亮**,副教授,主研领域:教育信息技术,教育大数据。**吴逸尘**,硕士生。**殷姿**,硕士生。

概念,将各个优秀作文语料视为本体的同时,必然会涉及到如何对本体进行有效的属性描述。而能够从作文语料本身抽取属性,即作文的标签,则是描述的核心内容。一方面,传统的标签抽取策略大部分采用关键词抽取方案,没有做到对作文语料的合理安排,即缺乏对小学语文作文标签的定义。另一方面,原始语料文本中包含的大量冗余信息也会对文本的关键信息抽取产生干扰。

综上所述,文本利用自然语言处理中文本自动摘要的方法去除冗余,并在定义了小学语文作文语料标签的前提下,提出了一种基于自动文摘技术的小学语文作文语料自动标签抽取方法。

1　相关技术研究

1.1　作文自动摘要方法

1.1.1　自动摘要

在对小学语文作文原始语料的处理过程中,为了使语料能更好地表达文章中心思想,则需要对原始语料进行去除冗余操作,其中,最常用的方法是文本自动摘要。自动文摘最早于 1958 年由 Luhn 提出,起初没有得到较高的关注度,但随着信息时代数据的爆炸式增长,人们逐渐意识到自动文摘对于文本去除冗余、提取中心的重要性。

自动文摘主要过程为文本分析、信息选取及文摘语言转换[4],从不同角度可以分为不同类型,从文摘的获取方式上可分为抽取型文摘和理解型文摘。抽取型文摘主要是从原文中选取合适的句子组成文摘,理解型文摘[5]则是通过对原文进行语义上的分析生成文摘,而理解型文摘因其较为深入的自然语言特征,一直处于技术攻坚阶段,故不具有实用性。本文采用抽取型文摘的技术方案对语料进行预处理。

图 1 为抽取型自动摘要的一般流程。其中,预处理环节主要对文本内容进行编码及断句处理。特征分析阶段则通过不同的分析方法获取原文中句子的权重,再通过对权重排序,选出适量句子,重新排序后输出。

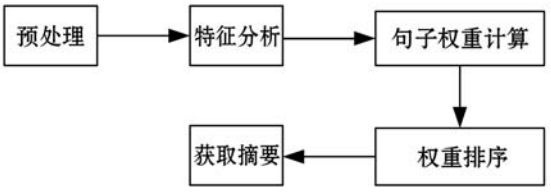

图 1　抽取型摘要的一般流程

1.1.2　作文自动摘要

目前主流的抽取型文摘方案来自于 Rada Mihalcea 和 Paul Tarau 提出的 TextRank 算法,是对谷歌基于图的网页权重计算方法 PageRank 加权重演变而来[6],主要用于关键词抽取和文摘句抽取[7]。

TextRank 算法分为计算相似度和排序两个部分,对于一篇语料而言,首先将句子分离出来,依据 Page Rank 的思路建立图。其中,图的节点就是句子,节点之间的权值就是句子之间的相似度。然而,选择不同的相似度算法,对 TankRank 产生的最终结果也有不同影响[8]。因此,本文比较了几种主流的相似度计算方法,进而选择出一种最适用于作文语料处理的自动文摘方法。

1)经典相似度算法。该方法借助于两个句子之间的共有词语来达到整体相似度计算的目的。

2)基于编辑距离的相似度算法。该方法的核心思想为计算从一个子串转移到另一个子串所需要的最小步骤,主要操作为“替换”、“插入”和“删除”。通常认为,两个子串的编辑距离越小,相似度越大。

3)基于 Word2Vec 的相似度算法。Word2Vec 可用于对文本进行词语聚类,获取关键词[9]。其主要思想是将自然语言中的词汇,映射到一个共同的维度内,使之成为一个个具有统一意义的短向量[10]。本文首先需要训练基于 Skip - Gram + HierarchySoftmax 的模型,然后获取词语之间的相似度关系,最后推算出句子间的相似度。

4)基于 BM25 的相似度算法。BM25 算法基于概率检索模型,其核心思想为解析搜索词,生成对应的语素信息,并将语素与文档进行比对,最后由每一个比对结果进行加权求和得到最终的相似度。

本文采用 ROUGE 对这四种相似度算法进行评价。ROUGE 是由 Chin-Yew Lin 在 2004 年提出的一种针对自然语言处理的自动评价方法[11],在其评价指标中,Precision 用于描述机器摘要的准确率,也被称为查准率;Recall 用于描述机器摘要的召回率,也被称为查全率;F-Score 是 Precision 与 Recall 的加权平均值,反映了机器摘要结合准确率和召回率的统一分数。通过对 1 410 条数据进行摘要处理,得到测试结果如表 1 所示。

表 1　四种自动摘要相似度算法的 ROUGE 评分结果

评价		算法			
		经典相似度	编辑距离	BM25	Word2Vec
ROUGE-1	Precision	0.781 4	0.865 1	0.823 2	0.595 3
	Recall	0.976 7	0.953 8	0.977 9	0.948 1
	F-Score	0.868 2	0.907 3	0.893 9	0.731 4

续表 1

评价		算法			
		经典相似度	编辑距离	BM25	Word2Vec
ROUGE-2	Precision	0.733 6	0.831 7	0.780 3	0.551 4
	Recall	0.918 1	0.917 5	0.927 7	0.880 6
	F-Score	0.815 5	0.872 5	0.847 7	0.678 1
ROUGE-W	Precision	0.314 5	0.338 7	0.322 3	0.246 7
	Recall	0.502 9	0.477 7	0.489 7	0.502 6
	F-Score	0.387 0	0.396 4	0.388 7	0.330 9

从表 1 中可以看出,基于 Word2Vec 的相似度算法在 F-Score 上的得分较低。基于编辑距离的计算方法在准确度上得分较高,基于 BM25 的计算方法在查全率上得分较高,并且,两者在最后的 F-Score 得分上相差不大。经典相似度计算方法在各个指标上均获得了较合理的分数,类似的情况也体现在 ROUGE-2、ROUGE-W 的评分方法中。

但在实际的标签抽取过程中,由于语料内容繁多,因此对时间效率也有一定的要求。本文在测试时,也对这 4 种算法的时间消耗做了记录,结果如表 2 所示。

表 2 四种自动摘要相似度算法的耗时

方法	耗时/ms
经典算法	28 085
编辑距离	38 896
BM25	1 687
Word2Vec	4 117

从表 2 中可以看出,BM25 算法虽然在 ROUGE 得分上不是最高的,但在计算的时间效率上领先较多。在 F-Score 相差甚微的前提下,本文最终选取 BM25 算法作为作文语料自动文摘预处理的相似度算法,并得到基于 TextRank 算法的作文语料抽取型自动文摘结果。

1.2 基于摘要的作文标签获取

1.2.1 分 词

在进行标签抽取之前,需要将句子以词汇组合的形式呈现,在自然语言处理中,通常采用中文分词的方法加以实现。目前主要的分词方法有基于词典的方法和基于统计的方法,由于基于词典的方法在算法复杂度以及分词速度上更具有优势,故本文选取基于词典的分词方法进行分词处理。

在基于词典的分词方法中,选择基于 N-最短路径的分词算法作为主要算法。其基本思想是根据词典,顺序匹配出在中文字串中所有可能的出现的词的集合[12]。相较于传统分词算法,其特性更适合发掘命名实体,故最适合本文分词方案。

1.2.2 命名实体识别

在小学语文作文标签抽取的过程中,需要对经过分词处理后的作文语料进行词性标注,识别出能够代表作文类型的词语,这就涉及命名实体识别。

目前对中文语料中普通的人名、地名等命名实体识别的研究中,中科院俞鸿魁等设计的一种层叠隐马尔可夫模型就能达到不错的效果。层叠隐马尔可夫模型由三层隐马尔可夫模型构成,自下而上分别是人名识别 HMM、地名识别 HMM 和机构名识别 HMM。

通过对作文语料的观察发现,机构名出现的频率并不高,因此本文将重点关注人名和地名的识别。基于层叠隐马尔可夫模型中关于人名和地名的部分标注角色见表 3。

表 3 人名及地名的部分角色标注

角色	人名识别中的含义	地名识别中的含义
A	人名的上文	地名的上文
B	人名的下文	地名的下文
C	中国人名的姓	中国地名的首部
D	双名的首部	中国地名的中部
E	双名的末字	无
F	单名	中国地名的末部
X	连接词	连接词
G	无	中国地名的后缀
Z	其他非人名成分	其他非地名成分

利用层叠隐马尔可夫模型可以高效地识别小学作文语料中的重要人名地名,从而协助标签抽取过程中的作文分类标签的获取。

1.2.3 词典设计

在实际处理作文语料时发现,一些特殊名词如“父亲”、“母亲”等,命名实体识别模型不会对其作出实体判断。但实际上,这些词语应归类于人物描写标签的范畴。另外,一些地名中包含的名词出现频率较高,分词模型可能会对其进行单独分类。为了避免这些情况,本文提出了一种自定义的词典内容来协助标签抽取。针对人物相关名词,本文结合实际经验,借助不同类型名词分类建立专属词典,与命名实体识别模型相结合达到更准确的结果。具体分类见表 4。

表 4　人物描写类型自建词典

分类	示例
亲属	爸爸、妈妈、爷爷……
职业	交警、医生、老师……
陌生人	叔叔、阿姨……
熟人	同桌、同学……

针对地名的情况,可依据词缀来进行相应识别,具体分类见表 5。通过对实际语料分析表明,小学语文作文中对家乡的描写一般以风景为主,故把其归类到景物描写中。

表 5　景物描写类型自建词典

分类	示例
自然风光	岛、湖……
行政划分	镇、乡……
季节	春、夏、秋、冬……
人工景观	公园、遗址……

2　基于自动摘要的作文标签抽取策略

前一节针对摘要的自动获取方法以及标签抽取过程中所需要的分词、命名实体识别以及词典设计等技术方法进行了比较与分析。本节将介绍基于自动摘要的作文标签抽取策略,该策略的实现方式如图 2 所示。

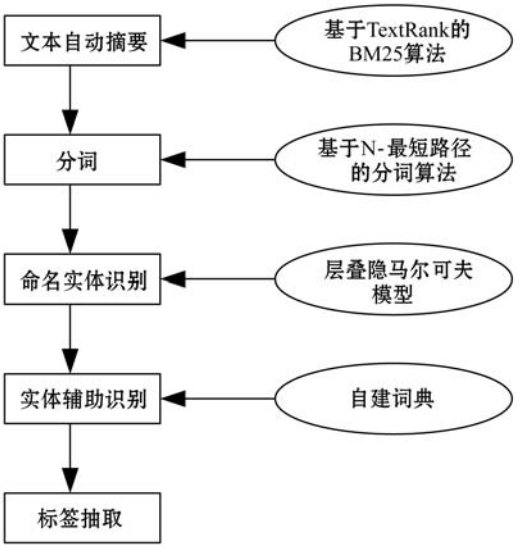

图 2　基于自动摘要的作文标签抽取方法

2.1　标签定义

通过对小学作文语料的分析,本文将文章标签总数限制为 6 个,同时根据标签涵盖的内容将其分为文章类型、核心实体、关键描述三个大类。其中,文章类型指的是小学语文作文的分类,由于小学语文作文的具体类别界限可以从不同维度、不同细分程度来划定,而作文分类并非本文的唯一目的,故后文对文章类型的阐述主要以人物描写和景物描写两个大类进行区分。

核心实体是从文章中获取的最核心命名实体。在人物描写分类中,核心实体是主要描写的人物;在景物描写分类中,核心实体是景物场景。关键描述是文章中频繁出现的形容词或文章中出现的俗语、成语等描述性词语。各分类限制词数见表 6。

表 6　标签三个分类的词数规定

分类	词数说明
文章类型	限制为 1 个,即景物描写或人物描写
核心实体	不超过 2 个,实际语料中一般为 1 个,如果有其它命名实体被发掘,则保证总数不超过 2 个
关键描述	除前两个类型之外,剩余的词数即为关键描述的词数,三类词数总和不超过 6 个

2.2　抽取实现

分词及命名实体识别基于开源自然语言处理框架 HanLP 实现,抽取过程如图 3 所示。

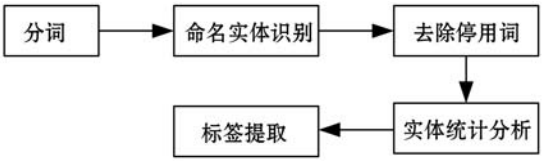

图 3　标签抽取整体步骤

命名实体识别在分词的结果上进行,最终以词性标注的方式显示。去除停用词目的在于去除抽取结果中的常用词语,减少对抽取策略的干扰,采用综合停用词表法实现。对于本文所需要的标签词语,其词性类型如表 7 所示。其中,nr、ns 代表最终获取的标签类型中的核心实体的词性,其他词性为关键描述中所涉及的词性。

表 7　标签抽取重点关注词性及实体

词性	含义
nr	人名实体
ns	地名实体
nz	专有名词
a	形容词
l	习惯用语,俗语
i	成语

其中,习惯用语、俗语一般为 4 字以上词语,是为了与文本中字数少且出现频繁的词语加以区分,从而更加精确地获取与文章核心实体相关的描述性词语。将经过命名实体识别后的结果按照词语-词性存储为列表,并按照降序从上而下排列,从高频词开始分析,

具体分析过程见图 4。

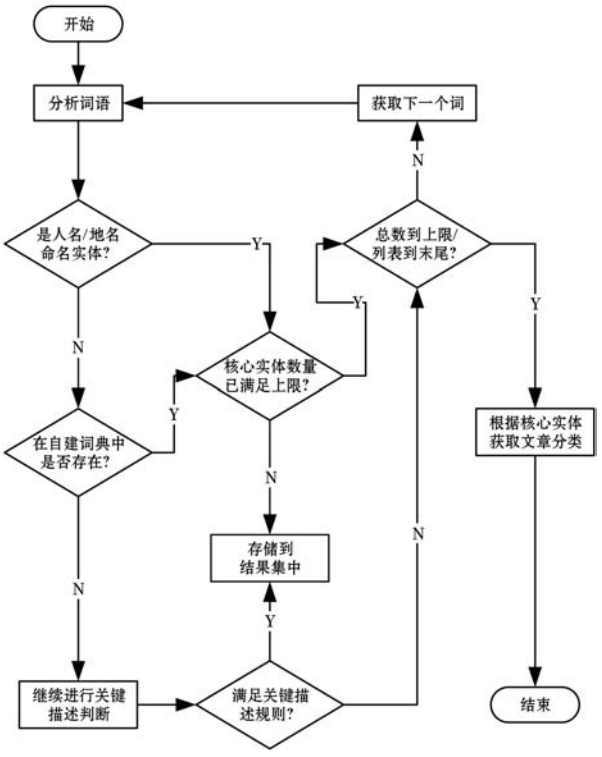

图 4 词语-词性列表分析流程

对于核心实体的获取过程主要分为两个步骤,第一,依据标准命名实体识别的方法进行识别,当词语词性为 nr/ns 时,检查该分类下计数器是否等于上限次数 2 个,若已达到上限则不作处理,反之将其加入到结果集中。第二,对于进行标准命名实体识别方法后没有标记出的词语,优先采用自建词典配对,若该词语存在于自建词语中,则进行与第一步类似的操作。

获取到的实体除了存储到结果集中以外,还需要记录词频来表示该实体所属类型占的权重,若结果集中核心实体的次数已满足,则后续识别到的词语继续计算权重但不添加到结果集中。若自建词典中也不存在当前分析的词语,则该词语进入到关键描述的判断中,关键描述判断的具体描述规则如下所示,依据以下规则,可以获取到标签分类中的关键描述部分。

关键描述判断流程:

(1) 当前词语长度是否超过 2,不满足则输出否;

(2) 当前词语的词频是否大于等于 2,不满足则输出否;

(3) 当前词语词性是否为所要求的描述性词语词性,不满足则输出否;

(4) 同时满足(1)、(2)、(3)条件的,判断词语属于关键描述;

(5) 特殊情况下,出现次数超过 3 次且长度大于等于 2 的一般名词(词性为 n)及专有名词(词性为 nz)将被判断属于关键描述;

(6) 当列表读取结束,总标签数仍不满足目标的情况下,对词语列表中词频为 1,但字数在 4 个以上的俗语及成语进行补充录入。

当从列表中获取的标签数已经满足需求或者词语-词性列表已经读取到末尾,则结束词语列表分析。此时,判断命名实体中的两个大类的权重对比,即判断 nr. weight 和 ns. weight,来确定文章类型标签。若 nr. weight > ns. weight,则该分类标签为人物描写;若 nr. weight < ns. weight,则该分类标签为景物描写;若两者相等,则将两个标签都添加到结果集中进行输出。至此,小学语文作文语料的标签抽取已经完成。

3 测试与评价

3.1 评价指标

对于抽取标签的结果,没有现行的统一标准。因此,本文设计三个维度来评价抽取的结果,这三个维度分别为分类准确度、实体准确度和形容词准确度,总分值设为 6 分。

1) 分类准确度用于描述标签中作文分类的正确与否。分类主要为人物描写和景物描写两个大类,正确得 1 分,错误得 0 分,该项评分总分 1 分。

2) 实体准确度用于描述标签中核心实体的正确程度,该项总分 2 分。由于核心实体标签数量为 1 ~ 2 个,故具体得分情况为:当核心实体标签数量为 1 个的时候,实体选取正确得 2 分,错误得 0 分;当核心实体标签数量为 2 个的时候,根据正确得数量获取得分。

3) 关键描述准确度用于描述标签中获取到的形容词或重要名词是否合适,该项总分 3 分。多个形容词情况下,从前至后按词频比求取加权平均值,共划分为 4 个层级进行评分工作,分别为:

(1) 所摘取关键描述与核心实体之间的关联度高,且对核心实体的描述十分贴切,如“险峻”之于“华山”,该情况得分为 3 分。

(2) 所摘取形容词与核心实体之间关联度一般,但对于核心实体所属类型而言,较为贴切。如“波澜不兴”之于“滇池”,该形容不具有代表性,但对于湖泊而言,相对通用,该情况得 2 分。

(3) 所摘取形容词于核心实体之间关联度较低,但对于分类标签而言,尚可利用。如“层峦叠翠”之于“太湖”,虽不具有较强关联性,但对于景物描写分类而言,尚有利用价值,该情况得 1 分。

(4) 所摘取形容词于核心实体之间无关联度,且于分类标签无价值。如“勤奋”之于“死海”,该情况得 0 分。

本文研究基于上述评分标准,以总分 6 分,3 个维度入手,通过人工评价的方式对标签抽取进行分数评估。通过严格设定各评分段位界定方法,可以有效降低人工评价中主观色彩过强引起的结果偏差。

3.2　对比方案

本文对比测试选择关键词抽取算法,由于对比项为关键词抽取,不具有文章分类能力,故选取 5 个关键词来与本文标签抽取策略获取的结果进行除分类外的对比。

TF-IDF 即词频-逆文件频率,是常用于资讯检索及资讯探勘的一种加权技术,其核心思想为一个词在一篇文档中出现频次高,而在其他文档中出现频次小,则具有代表价值。对应到单文档的关键词抽取中,则将对应句子视为评估单位。

TextRank 算法也常被用于关键词提取。为了计算句子之间的关联性,在 PageRank 的基础上引入了边的权值概念,并运用相应的相似度算法进行计算。而在获取关键词的过程中,若将词视为句子,则所有节点之间的权重变为 0,那么 TextRank 算法的计算就退变回了 PageRank 算法。

Word2Vec 可以将词语转换为语义向量,自然也能运用于关键词提取。本文对比方案利用朴素贝叶斯假设,将句子序列视为词语序列的集合。具体计算词语权重的方法为将序列集合中词语与词语之间的转移概率进行求和操作。

3.3　测试结果

测试数据为小学语文作文共 50 篇,涉及到的年级为三年级、四年级和五年级。采用人工盲评的方式进行打分,计算结果取平均分,保留小数点后三位最终的得分结果如表 8 所示。

表 8　测试结果

算法	分类准确度	实体准确度	关键描述准确度	得分总和
本文方案	0.806	1.755	0.869	2.625/3.431
TF-IDF 算法		1.755	0.467	2.222
TextRank 算法		1.735	0.418	2.153
Word2Vec		0.327	0.424	0.751

本文提出的标签自动抽取方案在不计算分类准确度的情况下最终得分为 2.625 分,计算分类准确度的情况下最终得分为 3.431 分,明显优于常用关键词算法。其中,分类准确度达到 80%,能够比较有效地区分人物描写和景物描写。在实体准确度上,本文方案、TF-IDF 算法以及 TextRank 算法在实体准确度上都取得了较好的效果,但 Word2Vec 的结果却差强人意,这可能是由于在关键词权重计算方法上缺少针对性。

在关键描述的得分结果上,本文得分明显优于其他三类得分,这说明本文所提出的标签抽取策略在该评价标准下取得了较好的效果。一方面是因为本文方案在选取关键描述的时候主动排除了动词的干扰,但是在另外三种算法中却没有体现。另一方面,由于本文方案采用了自动摘要去除冗余,因此能够更好地获取到中心内容,而另外三类算法没有对原始语料进行去除冗余操作,导致一般性动词的大量留存,最终干扰了关键描述的抽取结果。

最后,在对不同年级的评分结果进行比较时发现,高年级的平均得分明显高于低年级,这是因为高年级学生用词更加丰富,更加适合本文方案。当然,在对关键描述的获取上,本文仍有较大的进步空间,但这并不妨碍本方案的可用性。

4　结　语

本文围绕当前小学语文作文辅助中,作文素材的非结构化特征与语料信息化所需要的结构化数据之间的矛盾,提出了基于文本自动摘要的小学语文作文标签提取方法,实现作文语料的结构化组织。

本文对标签抽取过程中涉及的关键技术进行了对比分析,选取了抽取型自动文摘方法、基于词典的分词方案以及有监督的命名实体识别方法等作为主要技术框架。同时,本文尝试给出了小学语文作文标签的定义并根据应用场景设定了相应的评价指标。仿真实验表明,本文方法在小学语文作文领域的标签准确度评估中相较于传统的关键词算法有较大提升。

参　考　文　献

[1] 中华人民共和国教育部. 义务教育语文课程标准[M]. 2011 年版. 北京:北京师范大学出版社, 2011.

[2] 刘凤霞. 浅议小学语文作文教学方法的创新[J]. 语文教学通讯·D 刊(学术刊), 2013, 750(9): 54-55.

[3] 刘峤, 李杨, 段宏,等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582-600.

[4] 郭燕慧, 钟义信, 马志勇,等. 自动文摘综述[J]. 情报学报, 2002, 21(5): 582-591.

[5] Liu F, Flanigan J, Thomson S, et al. Toward abstractive summarization using semantic representations[EB]. eprint arXiv: 1805.10399, 2018.

[6] Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]//Emnlp, 2004: 404-411.

[7] 曹洋. 基于 TextRank 算法的单文档自动文摘研究[D]. 南京:南京大学, 2016.

(下转第 322 页)

从实验结果可以看出,LR 的准确率为 68.6%,KNN 的准确率达到 83.5%,而 RF 的准确率达到 90.3%,可见集成算法的效果要明显好于普通的机器学习算法。同时,基于参数优化的改进 RF 算法准确率为 91.6%,基于子分类器优化的改进 RF 算法准确率为 92.2%,其效果都要优于直接使用 RF 算法的效果,且 CGA-RF 的效果要优于 PGA-RF,可见选择性集成对模型效果有所提升,实验结果对比如图 9 所示。

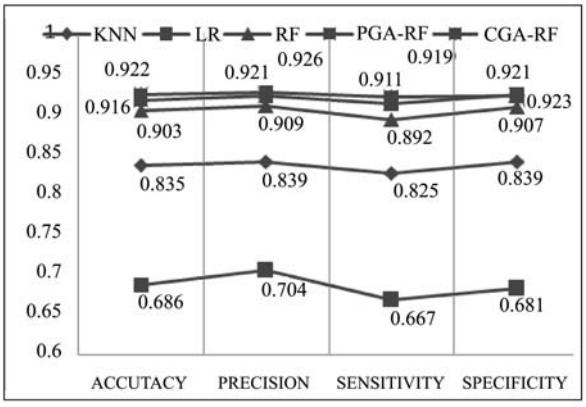

图 9 SSL VPN 识别结果

5 结 语

由于网络中的 SSL 加密流量越来越多,对于加密流量的监管也变得越来越重要。本文提出的混合方法,将指纹识别与机器学习方法相结合,实现了 SSL VPN 加密流量的识别。对于 SSL 流的识别,本文提出的改进的指纹识别方法对流漏识别的情况有所改善,使得识别效果更好。对于 SSL VPN 的识别,本文提出的改进 RF 算法也对准确率有所提升。实验结果表明,当前方法能够到达 92.2% 的准确率,实现了 SSL VPN 加密流量的有效识别。

参 考 文 献

[1] Paxson V. Empirically derived analytic models ofwide-area TCP connections[J]. IEEE/ACMTransactions on Networking,1994,2(4):316-336.

[2] Gabriel Gómez Sena, Belzarena P. Early trafficclassification using support vector machines[C]//International Latin American NETWORKING Conference, Lanc 2009, Pelotas, Brazil, September. DBLP,2009:60-66.

[3] Dainotti A, Pescapé A, Rossi P S, et al. Internet traffic modeling by means of Hidden Markov Models[J]. Computer Networks, 2008, 52(14):2645-2662.

[4] Schatzmann D, Spyropoulos T, Dimitropoulos X. Digging into HTTPS: flow-based classification ofwebmail traffic[C]// ACM SIGCOMM Conference onInternet Measurement. ACM, 2010:322-327.

[5] Sun G L,Xue Y, Dong Y, et al. An novel hybrid method for effectively classifying encrypted traffic[C]//Global Telecommunications Conference. IEEE, 2010:1-5.

[6] Lashkari A H, Draper-Gil G, Mamun M S I, et al. Characterization of encrypted and VPN traffic using time-related features[C]//The International Conference on Information Systems Security and Privacy,2016:94-98.

[7] Bagui S, Fang X, Kalaimannan E, et al. Comparison of machine-learning algorithms for classification of VPN network traffic flow using time-related features[J] Journal of Cyber Security Technology,2017,1(2):108-126.

[8] 苏皞昕, 施勇, 薛质. 基于 SSL 流量的指纹识别[J]. 信息安全与技术, 2015(11):58-60.

[9] Internet Engineering Task Force(IETF). The Secure Socket Layer(SSL) Protocol Version 3.0[OL]. 2011. http://tools.ietf.org/html/rfc6101.

[10] 周志华. 机器学习[M]. 北京:清华大学出版社, 2016.

[11] Zhou Z H, Wu J, Tang W. Ensembling neural networks: many could be better than all[J]. Artificial Intelligence, 2002,137(1/2):239-263.

[12] Ho T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1998, 20(8):832-844.

(上接第 227 页)

[8] 夏天. 词语位置加权 TextRank 的关键词抽取研究[J]. 现代图书情报技术, 2013, 29(9): 30-34.

[9] 李跃鹏, 金翠, 及俊川. 基于 word2vec 的关键词提取算法[J]. 科研信息化技术与应用, 2015,6(4): 54-59.

[10] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[EB]. eprint arXiv: 1301.3781, 2013.

[11] Flick C. ROUGE: A package for automatic evaluation of summaries[C]//The Workshop on Text Summarization Branches Out. 2004: 10.

[12] 吴晓倩, 胡学钢. 基于 N-最短路径的中文分词技术研究[J]. 安徽理工大学学报(自科版), 2014(1): 72-75.

(上接第 283 页)

[12] Fan Z, Su L, Liu X, et al. Multi-label Chinese question classification based on word2vec[C]//International Conference on Systems and Informatics. IEEE, 2018:546-550.

[13] 史兆鹏,邹徐熹,向润昭. 基于依存句法分析的多核心词义消歧[J]. 计算机工程,2017,43(9):210-213.

[14] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[15] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[16] 周练. Word2vec 的工作原理及应用探究[J]. 图书情报导刊, 2015(2):145-148.