文章编号：2095-2163(2021)05-0229-07 中图分类号：TP391 文献标志码：A

基于词频-逆文档频率和法律本体的相似案例检索算法

张云婷，叶 麟，方滨兴，张宏莉
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：智慧检务是近年来研究的热点问题，而相似案例检索是智慧检务中公共法律服务模块的基本需求。传统的基于关键词的检索方式使案例的相似性仅局限在浅显的词语层面上，无法满足用户在文章和语义层面上的检索需求。针对公共法律服务中的相似案例检索问题，该文以公共法律服务案例为研究对象，引入能够突出法律语义的案例要素，并以其为依据为案例建模，提出了一种基于语义的相似案例检索算法。该算法首先结合词频-逆文档频率和法律本体，提取出语料库中全部案例要素，再基于向量空间模型，通过欧氏距离计算出用户输入案例和语料库中各案例的相似程度，从而实现语义层面上的相似案例检索。通过对 12348 中国法网司法行政（法律服务）案例库中案例的分类实验可知，与传统的词频-逆文档频率提取关键词方法相比，该算法在监狱教改类案例分类上，其 F1 值提高了 36.36%。

关键词：语义检索；文本相似度计算；词频-逆文档频率；本体知识；案例要素

A similar case retrieval algorithm based on TF-IDF and law ontology

ZHANG Yunting, YE Lin, FANG Binxing, ZHANG Hongli
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Intelligent procuratorial is a challenging task in recent years, and similar case retrieval is the basic demand of public legal service module in intelligent procuratorial. However, the traditional retrieval methods based on key words make the similarity of cases limited to the level of words, unable to meet the retrieval needs of users at the level of articles and semantic. To solve the problem of similar case retrieval in public legal service, this paper takes the public legal service cases as the research object, introduces the case elements which can highlight the legal semantics, and according to them, proposes a similar case retrieval algorithm based on semantics. The algorithm first extracts the case elements of all cases in the corpus by combining term frequency-inverse document frequency (TF-IDF) and law ontology, and then calculates the similarity degree between user input cases and cases in the corpus through Euclidean distance based on the vector space model (VSM), so as to achieve similar case retrieval at the semantic level. Through the classification experiment of cases in Public Legal Services of China, it can be seen that compared with the traditional TF-IDF extraction method, the proposed algorithm increases the F1 by 36.36% for prison education reform cases.

【Key words】 semantic retrieval; text similarity computing; TF-IDF; ontology; case element

0 引 言

随着普法率的不断提高，普通民众的法律意识日益增强，法务系统的建设变得愈发重要，人们对公共法律服务的需求也呈上升趋势。由于现阶段面向公共法律服务的专业队伍规模有限，且分布在偏远地区的专业人员相对较少，很多民众对公共法律服务的需求无法得到很好的满足。在此情况下，智慧法务系统和智慧检务系统应运而生。无论是普通民众还是法官、律师等法律从业人员，均可从中得到所需的服务，这将为缓解专业队伍规模不够及分布不均的问题，提供有效的系统和技术支撑。其中，相似案例检索又是法务系统中一项基础且必要的需求。相似案例检索是指检索出与用户输入案例相似的案例，其在法官判案、民众普法、案件分类等方面均起着重要的作用。但是，传统的基于关键词的检索方式只能实现字面意义上的相似，并不能实现语义层面上的相似；而由于法律案例本身涉及很多知识领域，导致其和一般文章有所不同，因此仅依靠基于关键词的检索方式，无法实现非专业人员的精确检索。为此，本文提出了一种基于词频-逆文档频率（TF-IDF）和法律本体的相似案例检索算法。该方法引入案例要素替代传统的关键词，使其能够实现语义上的相似文本匹配。其中，案例要素是指法律案例

基金项目：国家重点研发项目（2018YFC0830900）。

作者简介：张云婷（1997-），女，博士研究生，主要研究方向：自然语言处理、文本对抗；叶 麟（1982-），男，博士，副教授，主要研究方向：P2P 网络、网络安全、网络测量等；张宏莉（1973-），女，博士，教授，博士生导师，主要研究方向：网络与信息安全、云安全与隐私保护等。

通讯作者：张宏莉 Email：zhanghongli@ hit.edu.cn

收稿日期：2021-02-07

中的关键元素,如嫌疑人的个人背景、性格特点、心理特征以及行为表现等。基于案例要素的检索方式将法律案例的研究重点从关键词转移到了案例要素上,由于案例要素的法律语义远远强于关键词,在专业性上必定优于基于关键词的查找方式,进而更能满足非专业人员对相似案例更加精确的检索需求。

本文主要贡献如下:

(1) 提出了一种基于语义的相似案例检索算法,提升了相似案例检索结果的准确性。

(2) 引入了基于法条的案例要素,以公共法律服务中的监狱教改案例为例,搭建了案件模型,并构建了相关字典,提升了相似案例检索结果的专业性。

1 相关工作

1.1 基于本体知识的文本相似度计算

基于 Berners-Lee 在 1998 年于国际万维网联盟提出的语义网的概念,本体这一概念逐渐从中衍生出来。本体是一种重要的知识表示手段,文本相似中的本体不仅包括狭义上的本体,也包括通用词典、词汇表等具有本体知识的知识库[1]。本体可以根据其通用性分为通用本体和领域本体。研究中常用的通用本体包括《知网》[2](HowNet)、WordNet[3]等;领域本体包括医学本体[4]、法律本体[5]等。

本体知识一般用树状结构表示,相关算法大都基于"IS_A"关系树状分类体系。学者们通常将基于本体的文本相似度算法分为基于距离(Edge Counting Measures)、基于信息内容(Information Content Measures)、基于属性(Feature - based Measures)以及混合式(Hybrid Measures)文本相似度计算[6]。

基于距离的文本相似度计算的基本思想为,利用概念之间的路径长度来度量概念之间的语义距离。其最基础的算法为 Shortest Path 法[7],在其基础上,加入权重、最近公共父节点、树的深度、路径方向的改变次数等方面的考量,衍生出了 Weighted Links[8]、Wu and Palmer[9]等方法。

基于信息内容的文本相似度计算的基本思想是,利用两概念词共享的信息量,来度量其之间的语义相似性。其最具代表性的算法为 Lord 等人[10]提出的基于最近公共父节点计算共享信息量的算法;在此基础上,加入对其它公共父节点、自身结点的考量,分别衍生出 Resnik[11]和 Lin[12]法。

基于属性的文本相似度计算的基本思想为,利用两概念词的公共属性数,来度量其之间的语义相似性。其最具代表性的算法为 Tversky 算法[13]。在此基础上,衍生出了基于概念释词的方法[14],此类算法通过在释词(gloss)集合中提取公共属性或划分属性的相似程度,来计算 2 个概念的语义相似程度。

混合式文本相似度计算的基本思想为,综合以上 3 种方法,进行文本相似度的计算。目前研究出的文本相似度算法,大部分都为混合式算法。

1.2 词频-逆文档频率技术

TF-IDF 技术是一种被广泛使用的特征词提取技术,也是生成词向量的主要手段之一。TF-IDF 技术最早在文献[15]中提出,该技术用于评估词语对于文档集或语料库中文本的重要程度,是自然语言处理中提取文档主题或关键词的重要技术。其基本思想是:如果某个词语在某篇文档中出现频率很高,而在语料库里其它文档中出现频率很低,则这个词语在某种程度上可以作为该文档的特征词。因此,该技术可以用作文档分类、文本相似度计算以及信息检索等用途。

词频(TF)指的是某个词语 w 在某篇文档 d 中出现的次数 ($count(w, d)$) 与文档 d 中总词语数 ($size(d)$) 的比值,可用式(1) 来进行计算:

$$tf(w,d) = \frac{count(w,d)}{size(d)}. \tag{1}$$

逆文档频率(IDF)指的是语料库中的文档总数 N 与词语 w 所出现文件数 $docs(w)$ 比值的对数,可以用式(2) 来进行计算:

$$idf(w) = \log \frac{N}{docs(w)}. \tag{2}$$

而词语 w 在文档 d 中的 TF-IDF 值可以用式(3)计算:

$$tf - idf(w,d) = tf(w,d) \times idf(w). \tag{3}$$

为了生成每篇文档的词向量,需要对语料库中的所有文档进行特征词的抽取,总结出一串由 n 个特征词组成的特征词串 $w_0, w_1, \ldots, w_n$; 再针对每篇文档,依次计算这些特征词在该文档中的 TF-IDF 值,这些值就组成了该文档的词向量。

例如,对于文档 d 来说,其词向量就为 $(tf - idf(w_0,d), tf - idf(w_1,d), \ldots, tf - idf(w_n,d))$。利用这种方法,对语料库中的每篇文档,计算其对应的词向量,即可生成语料库中所有文档的词向量集合。

2 案例要素及其选择依据

案例要素是能够描述案例特征的关键元素。由

于公共法律服务案例基本都是用自然语言描述的，若想利用计算机对其进行处理，就需要将其中的各案例要素进行抽象化表示，再将抽象化表示后的案例要素进行量化，继而抽取每件案例中的案例要素，最终形成计算机可以处理的词向量（即本文引入的案例要素向量）。

案例要素的选取与算法的准确率密切相关。由于案例要素需要有一定的专业性，因此，笔者依据相关法律法规、对应领域内的专业知识及近千篇具体案例来选择合适的案例要素。

以监狱教改类案例为例，根据司法部 2003 年 6 月 13 日发布的《监狱教育改造工作规定》第四条："监狱教育改造工作，应当根据罪犯的犯罪类型、犯罪原因、恶性程度及其思想、行为、心理特征，坚持因人施教、以理服人、循序渐进、注重实效的原则。"，笔者共选择了 4 类案例要素，分别为犯人的个人背景、犯人的性格特点、犯人的心理特征以及犯人的行为表现。其中由于罪犯犯罪类型过于繁杂，且对于监狱教改案例而言代表性较弱，因此未将犯人的犯罪类型加入到案例要素类别中。而犯人的个人背景很大程度上决定了犯罪原因及恶性程度，犯人的性格特点与其思想行为有很紧密的联系，犯人的心理特征及行为表现也与该法律条款的心理特征和行为一一对应。因此，笔者所选取的案例要素类别非常具有代表性，能将整个案例的关键要素全部表征出来。

表 1 列出了这 4 个案例要素类别中各案例要素的选取依据。

表 1　案例要素选取依据

Tab. 1　Case elements selection basis

案例要素类别	案例要素	案例要素选取依据
个人背景	犯人是否为未成年	《监狱教育改造工作规定》第八条、第六十一条
	犯人是否患病	《监狱法》第十七条
	犯人的文化程度	《监狱教育改造工作规定》第二十六条
	犯人是否为少数民族	《监狱教育改造工作规定》第八条
	犯人是否为多次犯	《监狱教育改造工作规定》第二十条
	犯人的父母是否离异	《监狱教育改造工作规定》第十七条
	犯人的家庭经济是否困难	通过阅读大量案例总结
	犯人是否被家人溺爱	《监狱教育改造工作规定》第十七条
	犯人是否缺乏家人关爱	《监狱教育改造工作规定》第十七条
	犯人的感情经历是否不顺	《监狱教育改造工作规定》第十七条
性格特点	犯人的情绪是否不稳定	
	犯人是否自卑	
	犯人是否自负	
	犯人是否悲观	
	犯人是否冷漠	
	犯人是否心智不成熟	
	犯人是否内向	
	犯人是否外向	《监狱教育改造工作规定》第二十一条结合阅读大量案例总结
心理特征	犯人是否焦虑	
	犯人是否抑郁	
	犯人是否人际敏感	
	犯人是否恐惧	
	犯人是否偏执	
	犯人是否敌对	
	犯人是否神经病性	
	犯人是否强迫	
	犯人是否躯体化	
	犯人的睡眠饮食是否正常	《监狱教育改造工作规定》第四十三条、症状自评量表 SCL90、通过阅读大量案例总结
	犯人是否猜疑	
行为表现	犯人是否消极改造	
	犯人是否违规违纪	
	犯人是否有自杀自残倾向	《监狱教育改造工作规定》第二十一条

3 基于语义的相似案例检索算法

由于基于 TF-IDF 技术提取出的案例要素法律语义较弱,并不能真正做到语义查询。因此,还需要结合法律本体提取一部分案例要素,以达到增强法律语义的目的。本文将通过 TF-IDF 和法律本体提取出的案例要素相结合,再利用量化后的案例要素,计算输入案例和语料库中案例的欧氏距离,即可得到两者间的相似度。

3.1 基于法律本体的案例建模

在结合法律本体提取案例要素的过程中,首先需要进行法律本体的案例建模。案例要素的抽象化表示和量化过程即为案例建模的过程。在案例要素的抽象化表示过程中,需要以法律法规及大量案例为基础,将某类案例的特点分层抽象出来,进而建立案例要素表示体系。以监狱教改案例为例,对该类案例进行建模,每个监狱教改案例向量 $\boldsymbol{C}_{jyjg}$ 均可用式(4)的形式表示:

$$\boldsymbol{C}_{jyjg} = (\boldsymbol{B},\boldsymbol{P},\boldsymbol{H},\boldsymbol{A}). \tag{4}$$

其中,$\boldsymbol{B}$、$\boldsymbol{P}$、$\boldsymbol{H}$、$\boldsymbol{A}$ 分别代表犯人的个人背景、性格特点、心理特征以及行为表现。

为了使案例要素的表示体系更加充实,对案例要素四维特征中的每个特征进行了二次抽象。通过这种分层抽象的方式使得建立的案件模型更加完整,抽取出的词向量不会太稀疏。在对监狱教改案例模型的四类特征进行二次抽象后,监狱教改案例的具体模型可以表示为式(5)-式(8)所示:

$$\boldsymbol{B} = (b_1,b_2,b_3,b_4,b_5,b_6,b_7,b_8,b_9,b_{10}), \tag{5}$$

$$\boldsymbol{P} = (p_1,p_2,p_3,p_4,p_5,p_6,p_7,p_8), \tag{6}$$

$$\boldsymbol{H} = (h_1,h_2,h_3,h_4,h_5,h_6,h_7,h_8,h_9,h_{10},h_{11}), \tag{7}$$

$$\boldsymbol{A} = (a_1,a_2,a_3). \tag{8}$$

其中,各维度的含义及值域见表 2。

表中值域为{0,1},0 表示犯人没有该特征,1 表示犯人有该特征;犯人的文化程度 b_3 的值域为{0,1,2,3,4,5},0-5 的数字分别表示犯人文化程度为小学文化、初中文化、高中文化、中专文化以及接受过高等教育。

3.2 基于法律本体的案例要素提取

在建立某类案例模型后,需要在该类所有案例中一一提取所有的案例要素,将每个案例用词向量的形式表示。下文利用监狱教改类案例进行举例,介绍基于法律本体的案例要素提取方式。

由于每个案例要素都有各自的特点,无法一概而论,因此对于不同的案例要素,需要用不同的方法来进行提取。基于法律本体的案例要素提取主要有 4 种方法:基于正则表达式的案例要素提取方法、基于常识字典的案例要素提取方法、基于专业字典的案例要素提取方法、基于混合方法的案例要素提取方法。

表 2 监狱教改案例要素量化表

Tab. 2 Quantification of case elements for prison education reform cases

特征向量	维度	含义	值域
B	b_1	犯人是否为未成年	{0,1}
	b_2	犯人是否患病	{0,1}
	b_3	犯人的文化程度	{0,1,2,3,4,5}
	b_4	犯人是否为少数民族	{0,1}
	b_5	犯人是否为多次犯	{0,1}
	b_6	犯人的父母是否离异	{0,1}
	b_7	犯人的家庭经济是否困难	{0,1}
	b_8	犯人是否被家人溺爱	{0,1}
	b_9	犯人是否缺乏家人关爱	{0,1}
	b_{10}	犯人的感情经历是否不顺	{0,1}
P	p_1	犯人的情绪是否不稳定	{0,1}
	p_2	犯人是否自卑	{0,1}
	p_3	犯人是否自负	{0,1}
	p_4	犯人是否悲观	{0,1}
	p_5	犯人是否冷漠	{0,1}
	p_6	犯人是否心智不成熟	{0,1}
	p_7	犯人是否内向	{0,1}
	p_8	犯人是否外向	{0,1}
H	h_1	犯人是否焦虑	{0,1}
	h_2	犯人是否抑郁	{0,1}
	h_3	犯人是否人际敏感	{0,1}
	h_4	犯人是否恐惧	{0,1}
	h_5	犯人是否偏执	{0,1}
	h_6	犯人是否敌对	{0,1}
	h_7	犯人是否猜疑	{0,1}
	h_8	犯人是否强迫	{0,1}
	h_9	犯人是否躯体化	{0,1}
	h_{10}	犯人的睡眠饮食是否正常	{0,1}
	h_{11}	犯人是否神经病性	{0,1}
A	a_1	犯人是否消极改造	{0,1}
	a_2	犯人是否违规违纪	{0,1}
	a_3	犯人是否有自杀自残倾向	{0,1}

3.2.1 基于正则表达式的案例要素提取方法

利用正则表达式提取案例要素的两种情况:

(1)被提取的案例要素在所有案例中的描述虽

然不一致，但都遵循一定规律。例如，在提取“犯人是否为未成年”这一案例要素时，每个犯人的出生时间并不一样，且对出生时间的描述也不同。如，甲犯的出生时间描述为 1989 年生；乙犯的出生时间描述为 1989 年 2 月出生；丙犯的出生时间描述为 1989 年 2 月 5 日生等。从中可以看出，虽然每个案例对出生时间的描述不同，但是其遵循的格式规律是一致的，其格式都是最前面是 4 个数字，最后是一个“生”字。因此，可以通过相应的正则表达式抽取出所需的出生年份的信息。同理，如果有些案例中没有出生日期信息，但是有犯人年龄的信息，也可以用正则表达式将犯人的年龄抽取出来。

(2)被提取的案例要素在所有案例中的描述，无法通过一个或几个关键词判断，还需要考虑关键词的上下文，过滤掉不符合条件的情况。例如，在提取“犯人是否强迫”这一案例要素时，很容易知道“强迫”一词肯定是判断犯人是否强迫的必要不充分条件。因为很多案例中有“强迫”一词，但却并不能说明犯人就是有强迫倾向的。如，在案例描述中，该犯人因强迫妇女卖淫罪而被捕，该描述中也有“强迫”一词，但并不能说明该犯人有强迫倾向。因此，需要通过正则表达式将不符合强迫条件的信息过滤掉，剩下的就是所需信息。另外，该情况通常与基于常识的案例要素提取结合使用。

3.2.2　基于常识字典的案例要素提取方法

有些案例要素在所有案例中的描述可以通过一个或几个关键词来进行判断，即在某一案例中，只要匹配到所有对应关键词中的一个，就能够判断该案例拥有相应的案例要素。而这些关键词所构成的字典，则需要在阅读大量监狱教改类案例的基础上，结合常识进行总结，这种方式即为基于常识字典的案例要素提取。此种方式适用于关键词较少、在文中的描述较为规范且无需考虑上下文的情况。例如，在提取“犯人的文化程度”这一案例要素时，由于该案例要素在文中的描述大都为“小学文化”、“初中文化”、“高中一年级文化”等，这样的描述形式规范且无需考虑上下文，而文化程度的范围只是小学到高等教育，相应的关键词较少，常识字典很好建立，因此可以用此方法来对这一案例要素进行提取。

3.2.3　基于专业字典的案例要素提取方法

与常识字典相似，专业字典也是由关键词构成的，运用方式也与常识字典相同。而与常识字典不同的是，专业字典所包含的关键词数量极大，且专业性很高，通常为互联网中可获取的专业性细胞词库。例如，在提取“犯人是否患病”这一案例要素时，就需要收集所有疾病的名称。而这些疾病的名称显然无法用常识总结出来，因此笔者从互联网中下载了搜狗细胞词库中关于疾病名称的词库，通过与该词库中的疾病名称进行匹配，即可完整抽取出这一案件要素。

3.2.4　基于混合方法的案例要素提取方法

该方法是将上述 3 种方式中的 2 种方式进行混合使用，以达到更精确地提取案例要素的目的。如在基于正则表达式的案例要素提取方式的第二种情况中的举例，即为基于正则表达式的案例要素提取方式与基于常识字典或专业字典的案例要素提取方式的混合使用。

为了便于理解，将基于混合方法的案例要素提取进行如下伪代码表示。

```
输入：案例文本、常识(专业)字典
输出：对应抽取元素标志位(flag)
  对不同案例要素，进行对应的前期处理
  flag ← 0
  for 字典中的每一个词语 do
    if 案例文本中能找到该词语 then
      flag ← 1
    end if
  end for
  利用正则表达式过滤不符合条件的案例
  if 匹配成功 then
    flag ← 0
  end if
  对不同案例要素，进行对应的后期处理
  return flag
```

3.3　人工增加停用词表

在基于 TF-IDF 的案例要素提取算法中，停用词表没有经过任何的人工改动。这样虽然减少了人工操作，但会使得一些没有区分能力的词语被抽取到特征词串中。这些词语通常是一些普遍出现在公共法律服务案例中，但却无法作为案例要素的词语。例如，“监狱”、“民警”等。因此，需要通过人工的方式，将它们添加到停用词表中，这样可以在很大程度上排除非特征词的干扰，使提取的案例要素更具有代表性。

值得注意的是，虽然非特征词可以通过调节 *max_df* 参数(该参数可忽略在阈值以上的文档数量中出现过的词语)进行去除，然而这种去除方式也会过滤掉那些普遍出现在各文档中、却仍能作为一

篇案例的案例要素的词语。例如,“焦虑”、“抑郁”等词语。因此,人工增加停用词在去掉非特征词的基础上,不会去掉真正有用的案例要素,从而使得后续计算相似度时得到的结果更加准确。

3.4 计算文本相似度

通过上文的方法,将用自然语言描述的案例量化为词向量后,即可进行文本相似度的计算。该思想来自于向量空间模型(VSM)。VSM 的基本思想是:假设词与词之间是不相关的,以向量来表示文本,从而简化了文本中关键词之间的复杂关系,使得模型具备了可计算性[16]。当模型具备可计算性之后,即可利用数学中向量的计算方法,计算 2 个向量之间的距离。利用欧氏距离计算空间中 2 个向量之间的距离。若 2 个向量之间的欧氏距离越小,则 2 个向量在空间坐标系中就越近。具体的计算方法如下:

设:向量 $\boldsymbol{A}$ 为 $(a_1,a_2,\ldots,a_n)$,向量 $\boldsymbol{B}$ 为 $(b_1,b_2,\ldots,b_n)$,则二者之间的欧氏距离 S 可以用式(9)进行计算:

$$S=\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\ldots+(a_n-b_n)^2}. \tag{9}$$

4 算法结果对比分析

本文使用 $F1$ 作为评估指标,其主要计算方法如下所示:

$$F1=\frac{2\times TP}{2\times TP+FP+FN}, \tag{10}$$

$$P=\frac{TP}{TP+FP}, \tag{11}$$

$$R=\frac{TP}{TP+FN}. \tag{12}$$

其中,P 为精准率;R 为召回率;TP 表示预测为正样本,实际为正样本的结果数量;FN 表示预测为负样本,实际为正样本的结果数量;FP 表示预测为正样本,实际为负样本的结果数量。

在本实验中,正样本取监狱教改类案例,负样本取非监狱教改类案例。具体的实验方法为,选择某一案例,分别利用本文所提方法和 TF-IDF 方法(下称传统方法),计算该案例和语料库中所有监狱教改案例的相似程度,并选择与该案例最相似的案例进行分析。若检索出的最相似案例与输入案例之间的欧氏距离大于某一距离参数,则将输入案例分类到非监狱教改案例中;反之,则将输入案例分类到监狱教改案例中。

4.1 测试数据集

本文所选数据集为 12348 中国法网司法行政(法律服务)案例库中的案例。其中语料库中的语料为该案例库中的监狱教改类案例,被试语料为该案例库中的监狱教改类、监狱减刑类、法律援助类、人民调解类、律师工作类案例。其中监狱教改类别共有1 082篇文档,非监狱教改类别共有2 225篇文档。

4.2 算法测试

实验主要针对本文提出的基于 TF-IDF 和法律本体的案例要素提取算法,以及传统的基于 TF-IDF 案例要素提取算法进行测试,从而验证本文算法对相似案例检索结果的准确性及专业性。

(1)本文方法对数据集测试。首先利用人工操作,在哈尔滨工业大学停用词表的基础上,增加法律方面的停用词。之后将语料库中的所有文档进行分词处理,并去除停用词。利用基于法律本体所建立的案件模型,对未经分词处理的原始文档进行第一次案例要素提取,并将每次提取到的案例要素,以词语的形式添加到对应分词后的文档尾部,即可得到补充完案例要素的文档集合,建立词向量 TF-IDF 值的计算模型;再由该模型自动进行第二次案例要素提取,利用 3.4 节所述的方法,计算出各输入案例与语料库中各案例的相似程度,进而进行案例分类。

(2)传统方法对数据集测试。直接对语料库中的所有文档进行分词处理,建立 TF-IDF 词向量,并利用该词向量进行案例要素提取,最后利用 3.4 节所述的方法计算出各输入案例与语料库中各案例的相似程度,进而进行案例分类。

将上述 2 种方法分类结果的 $F1$ 值进行对比,即可对本文所提算法的性能进行分析与评估。

在建立 TF-IDF 词向量的过程中,将参数设定为 $min\_df=0.1$ 的含义是,忽略那些仅在 10%以下的文档数量中出现过的词语。例如,某语料库中有 100 篇文档,某个词语仅在其中的 9 篇文档中出现过,那么则不将其放入语料库的特征词串中,否则,最终得到的 TF-IDF 词向量矩阵将过于稀疏。在利用欧式距离的大小进行分类的过程中,使用的参数值为 1.1,该参数为笔者通过多次实验及经验得出的距离参数。

4.3 结果分析

由图 1 中数据分别可以看出,传统方法将 78.33%的监狱减刑案例错误地分类成监狱教改案例,而本文所提方法的此概率仅为 7.51%。由于监狱减