Open Journal of Transportation Technologies 交通技术, 2023, 12(5), 377-384
Published Online September 2023 in Hans. https://www.hanspub.org/journal/ojtt
https://doi.org/10.12677/ojtt.2023.125041

Hans汉斯

基于TF-IDF和jieba分词的交通运输综合执法语音文件和文本文件关联匹配技术

刘文平[1]，李艳春[1]，张　贺[1]，张宇驰[1]，丁　鼎[1]，于　泉[2]，王传炀[2]

[1]北京市交通运输综合执法总队，北京
[2]北方工业大学电气与控制工程学院，北京

收稿日期：2023年6月18日；录用日期：2023年8月22日；发布日期：2023年8月31日

摘　要

在交通运输综合行政执法听证环节中，传统听证环节均是线下举行的，听证记录员需要对整个听证环节的笔录进行详细记录。由于会后需要与整个案件的证据材料进行归档整理，对于执法人员的工作强度要求很高。因此，针对交通运输综合执法办案流程中的听证业务环节提供一定的技术支撑，利用TF-IDF算法对听证内容进行关键词提取，和jieba分词进行优化开发语音文件和文本文件关联匹配技术，实现听证语音文本与案件关键要素信息的精确关联匹配，构建完整证据链确保行政处罚有据可依，整体提升交通运输综合行政执法针对听证案件的处罚判决的充分与准确，助力政府治理系统和治理能力现代化建设。

关键词

TF-IDF，jieba分词，交通运输综合执法，关联匹配，听证会

TF-IDF-Based Transportation Integrated Law Enforcement Voice File and Text File Association Matching Technology

Wenping Liu[1], Yanchun Li[1], He Zhang[1], Yuchi Zhang[1], Ding Ding[1], Quan Yu[2], Chuanyang Wang[2]

[1]Beijing Transportation Comprehensive Law Enforcement Corps, Beijing
[2]School of Electrical and Control Engineering, North China University of Technology, Beijing

Received: Jun. 18th, 2023; accepted: Aug. 22nd, 2023; published: Aug. 31st, 2023

Abstract

In the comprehensive administrative law enforcement hearing process of transportation, the tra-

文章引用: 刘文平, 李艳春, 张贺, 张宇驰, 丁鼎, 于泉, 王传炀. 基于TF-IDF和jieba分词的交通运输综合执法语音文件和文本文件关联匹配技术[J]. 交通技术, 2023, 12(5): 377-384. DOI: 10.12677/ojtt.2023.125041

ditional hearing process is held offline, and the hearing recorder needs to keep detailed records of the entire hearing process. Due to the need to archive and organize the evidence materials of the entire case after the meeting, there is a high demand for the workload of law enforcement personnel. To provide certain technical support for the hearing business process in the comprehensive law enforcement process of transportation, TF-IDF algorithm is used to extract key words from the hearing content, and jieba segmentation is used to optimize the development of voice evidence files and text file association matching technology, Realize accurate correlation and matching between hearing voice text and key element information of the case, construct a complete evidence chain to ensure that administrative penalties are based on evidence, comprehensively improve the adequacy and accuracy of punishment judgments for hearing cases in transportation comprehensive administrative law enforcement, and assist in the modernization of government governance system and governance capacity.

Keywords

TF-IDF, Jieba Participle, Comprehensive Law Enforcement of Transportation, Association Matching, Hearing

1. 引言

语音文件与文本文件关联匹配技术的研究，主要是用于针对次场景下减少执法人员工作量的同时又能够将听证程序产生的证据材料能够准确地与案件文书材料进行关联匹配，避免人为的误操作导致执法部门的公信力下降。通过实现听证语音文本与案件信息的关联匹配，支撑执法人员的事后回溯处理、证据链完整性，确保在线听证业务与上下游业务节点的高效衔接，助力政府治理系统和治理能力现代化建设。

TF-IDF 算法作为有效的关键词提取算法已经广泛应用于诸多行业中，像传媒行业的新闻语音文本、新闻标题字幕内容进行关键词提取[1]，医药行业的症状、方剂等多种类型对象聚类发掘均有一定的技术研发利用[2]。但是在交通运输综合执法行业中，此技术的相关研发还相对欠缺，目前执法行业仍然采用人为记录的方式处理执法业务，研究相关技术应用到交通运输执法行业是具有一定意义的。

语音文件和文本文件关联匹配技术主要研究关键词提取算法和文本匹配技术，基于 TF-IDF 关键词提取算法并借助 jieba 开源库进行分词处理来实现听证语音文本与案件关键要素信息的精确关联匹配，以此来适配案件办理需求。

2. 文件关联匹配整体思路

本文研究的语音文件和文本文件关联匹配技术包括 3 个部分：一是语音文件采集，在听证程序过程中，在听证室的现场会有全程录音采集并会将语音转换为文本信息，主要能够对语音文件进行结构化存储，存储的信息包括语音文本、文件流、文件类型、文件描述等信息，主要用于后续操作使用。

二是语音文件存储与识别，首先，针对语音文件的结构化数据的关键信息进行抽取，包括文件流、文件存储地址、文件创建日期、文件格式、文件名、文件大小等，主要用于日后查询语音文件使用。其次，针对语音文本需要进行关键信息的提取。听证过程中产生的文本信息量很大，但是其中需要采集的

关键要素主要是案件的关键信息，针对此环节先期针对交通运输行政执法的案件信息进行了研究，利用TF-IDF 算法对近一个月的案件文书信息关键字进行了分析，几乎所有的案件中都包括当事人姓名、检查时间、车牌号信息，并且在听证的过程中也会对当事人询问这些信息，因此只需要提取这些信息就能够满足后续关联匹配的业务需要。然后，采用开源分词组件对语音文本进行分词，在根据关键字提取出所对应的关键信息。

三是语音文件建模匹配，在未来业务使用的角度进行建模，将语音文本信息与案件信息进行建模，形成语音文本与案件信息关联模型，其中包括案件编号、文书编号、当事人姓名、检查时间、车牌号、语音文本、文件存储地址。根据分词提取到的关键信息检索案件信息，并将模型所对应的信息存入数据库中，语音文件与文本文件的匹配关系就成立了，未来业务需要可以直接通过关联关系得到语音证件文件和案件的关键信息。具体流程如图 1 所示。

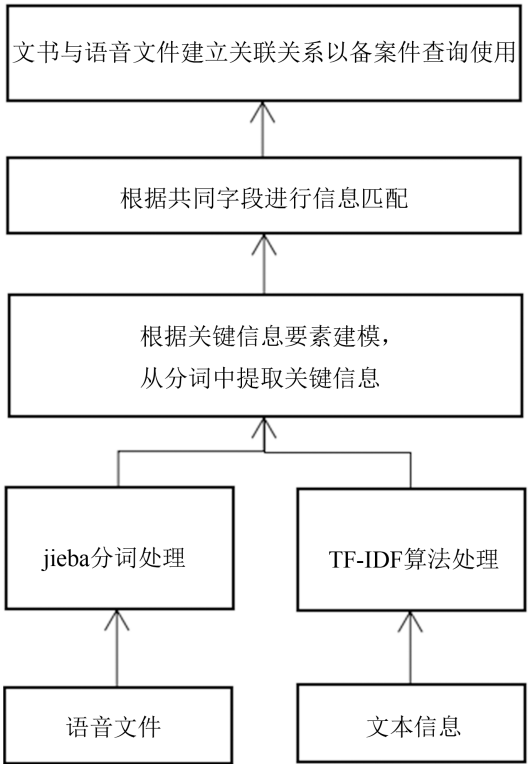

Figure 1. Text association matching overall flowchart
图 1. 文本关联匹配整体流程图

3. 基于 TF-IDF 和 jieba 分词的文本匹配处理

3.1. 概述

语音文件和文本文件关联匹配技术在交通运输行政执法领域的应用，首先要打通听证程序，执法业务系统在听证环节增加语音文件生成的流程。语音文件生成的同时，系统接收音视频相对应的文本文件，建立文本文件存储表，包括：文本内容，音视频文件存储地址，音视频文件大小(KB)，文件类型(音频、视频)，文件格式，文件描述等信息。并且调用研究的算法将语音文本进行分词处理，提取出关键信息包括当事人姓名、检查时间、车牌号。通过关键信息检索案件信息，建立关联匹配存储表，包括案件编号、文书编号、当事人姓名、检查时间、车牌号、语音文本、文件存储地址等。并且系统还应该支持结合业

务更新分词库关键字的功能。

其具体步骤如下：

第一步：定时更新词库信息，每隔五分钟更新一次词库，词库中需要加入的信息包括案件信息库中未办结的车牌号、当事人姓名、当事人证件号码、检查时间信息。

第二步：利用 jieba 分词的精准模式，使用 jieba 分词对收集到的数据进行切词处理，同时加载自定义词库对文本文件进行分词。

第三步：利用 TF-IDF 算法对关键词进行提取，利用 TF-IDF 的计算公式，先计算出每一个关键词的特征权值 TF (即每个关键词在所有文本中出现的概率)和相应的 IDF 值。然后根据 TF-IDF 公式将每一个关键词的 TF 值与 IDF 值相乘，计算出每个关键词的特征权值 TF-IDF，并按照特征权值的大小进行排序，选取排名靠前的词作为关键词。

第四步：通过提取的关键词与数据库中的案件基本信息进行文本匹配。

通过以上四步可以依据关键词中的当事人姓名，检查时间、车牌号精准的匹配到案件的基本信息。具体流程如图 2 所示。

Figure 2. Technical implementation process
图 2. 技术实现流程

3.2. Jieba 分词处理

分词是自然语言处理的第一步，也是比较重要的一步。分词是将由字符序列构成的句子按照一定的规则重新组合成词的集合。

本技术利用 jieba 分词对文本文档进行分词处理。jieba 分词主要用于对文本文档进行分词处理，jieba 分词库是一款流行的、广泛使用、分词效果较好的中文开源分词库，具有高性能、高准确率、可扩展等特点，主要功能包括分词、词性标注和关键词抽取等。

jieba 分词主要有以下几种模式：

1) 精准模式：该模式是默认模式，句子可以被精确地分离开，每个字符只会出现在一个词中，适用于文本分析；

2) 全模式：该模式把句子中的所有词都扫描出来，速度非常快，有可能一个字同时出现在多个词中；

3) 搜索引擎模式，在精确模式的基础上，该模式对长度大于 2 的词再次分词，召回其中长度大于 2 或者 3 的词，从而提高召回率，常用于搜索引擎。

并且 jieba 分词支持自定义词库，利用 jieba 分词自定义词库的特点，可以通过动态调整词库的方式确保文本文件和案件信息匹配成功。

语音文件和文本文件关联匹配技术研究采用的是精准模式。将文本文件进行分词整理，同时利用关键字提取所对应的关键信息。并且 Jieba 分词支持定制化更新分词库，从而可以根据实际业务调整关键字。

3.3. TF-IDF 模型

TF-IDF 是一个词频统计的算法模型，该模型通过词频统计与反文档频率，对文本的词频进行综合分析[3]。这种加权技术经常被用于信息检索与文本挖掘的。

TF-IDF 的主要思想是：如果某个单词在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，

则认为此词或者短语具有很好的类别区分能力，适合用来分类。

1) TF 是词频(Term Frequency)

词频(TF) = 某关键词出现次数/文章中关键词总数。TF 的计算公式如下：

$$\mathrm{TF}_{ab} = \frac{n_{ab}}{\sum_i n_{ab}} \tag{1}$$

其中，n_{ab} 为该词在文档 d_b 中出现的次数，$\sum_i n_{ab}$ 是文档 d_b 中所有词汇出现总次数。

但是，TF 表示法存在着天生的缺陷，当特征词为常见词的时候，其 TF 值会很高，但是这个特征词的重要程度却很低。例如《中国互联网络发展状况统计报告》，经过词频统计之后，可能会发现“中国”“互联网”这两个特征词出现的词频一样，但这并不意味着两个词对文本的重要性相同，也许“互联网”对文本的重要性更高。因此根据统计学语言表达，在原本 TF 表示法的基础上，为各个词计算逆文档频率(inverse Document Frequency, IDF)，以此来调整重要性系数[4]。IDF 越大，表明一个词越不常见，反之越常见。它能够加强一些作为低词频特征关键词的重要程度，削弱一些常见词的权值，进而减少这些常见词对最终实验结果的影响[5]。

2) IDF 是逆向文件频率(Inverse Document Frequency)

逆向文件频率(IDF)：log(语料库文档总数/(包含该词的文档数+ 1))

$$\mathrm{IDF} = \log \frac{|D|}{\left|\{b : t_a \in d_b\}\right| + 1} \tag{2}$$

其中，$|D|$ 是语料库中的文档总数，$\left|\{b : t_a \in d_b\}\right|$ 为包含词汇 t_a 的文档数，若该词不在语料库中，则分数为 0，故一般情况下分母设置为 $\left|\{b : t_a \in d_b\}\right| + 1$。

3) TF-IDF 实际上是：

$$\mathrm{TF} * \mathrm{IDF} \tag{3}$$

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

因此，在交通运输综合执法的业务领域，主要依赖 TF-IDF 算法分析出交通运输执法案件文书中哪些关键字是高频出现的，并且在听证程序中能够提取到的。

系统接收音视频文件，同步接收音视频相对应的文本文件，建立文本文件存储表，包括文本内容，音视频文件存储地址，音视频文件大小(KB)，文件类型(音频、视频)，文件格式，文件描述等信息。同时系统结合当前音视频使用的场景，将音视频与具体文书进行关联。智能提取文本中包含的命名实体信息，包括当事人姓名、当事人证件号码、当事人手机号、当事人车牌号码、专业术语等，对文本按照内容类型进行自动分类，分类基于已有类目体系，按照文本内容包含的语义信息自动完成文本分类；聚类即根据文本的内在数据分布、语义特征，将海量文本数据自动聚合成多类，并为每一类数据给出描述性关键词。从非结构化数据的属性中抽取有意义的结构化信息，在此基础上进行数据分析，提升数据价值。基于深度语义分析模型，自动抽取语音文本中涵盖内容主旨的关键信息并生成指定长度的文本摘要。可用于最终文书制作过程中。

4. 实验测试分析

4.1. 实验准备

本实验使用研究的语音文件和文本文件关联匹配技术和百度智能云、文章提取关键词标签提取工具以及 TextRank 工具包分别将提取的关键词与文本数据源进行匹配测试分析，匹配的目的通过听证案件当

事人的相关信息与具体文书进行匹配从而进行文书的制作。

实验使用数据为真实的听证案件数据，分别使用不同的方法进行处理和分析，比较它们的效果和性能，使测试结果更加真实可靠。

4.2. 实验结果

取四个听证案件文本文件数据来模拟不同的测试情况，设置百度关键字提取算法默认提取 25 个关键词，根据匹配结果可以得出与四个样本数据听证案件当事人信息是否匹配成功，以及每个样本数据的消耗时间。其中可以看出百度智能云均匹配成功平均耗时在 1335 ms 左右，四个案件的匹配结果如下表 1 所示。

Table 1. Baidu intelligent cloud experiment results
表 1. 百度智能云实验结果

序号	姓名	车牌号	匹配结果	平均耗时
1	刘能	冀 FB97XX	匹配成功	1337 ms
2	赵佳俊	冀 FB98XX	匹配成功	1365 ms
3	张鑫	冀 FB96XX	匹配成功	1348 ms
4	吴某军	冀 FB95XX	匹配成功	1310 ms

由于文章提取关键词标签提取工具是一个应用软件，它可按需设置提取的关键词数量提取，提取结果快速准确，那么在前期进行简单调试的时候可以作为一个独立的模块进行单独测试。但由于该方法无法在后期集成到实际项目中去，也无法完流程闭环的性能和效果测试。

考虑项目实际应用情况，应当使用自动化测试工具或者手动测试的方式来测试该应用软件的功能和性能。同时，还可以使用模拟数据或者模拟场景来模拟不同的测试情况，以评估应用软件在不同情况下的表现和效果。

利用 TextRank 工具包取四个听证案件文本文件数据来模拟不同的测试情况，这四个案件的匹配结果、词典和结果设置如下表 2 所示。

根据匹配结果可以得出与四个样本数据的消耗时间平均在 30 ms 范围以内，但是每个数据的听证案件当事人信息均匹配失败，不符合要求。

Table 2. TextRank toolkit experimental results
表 2. TextRank 工具包实验结果

序号	姓名	车牌号	匹配结果	平均耗时
1	刘能	冀 FB97XX	匹配失败	23 ms
2	赵佳俊	冀 FB98XX	匹配失败	11 ms
3	张鑫	冀 FB96XX	匹配失败	9 ms
4	吴某军	冀 FB95XX	匹配失败	19 ms

结合交通运输综合执法部门提供的听证模板进行分析，采用 TF-IDF 算法对关键词进行提取，匹配关键词进行案件信息管理，实现效果如下表 3 所示。根据匹配结果可以得出与四个样本数据听证案件当事

人信息均匹配成功，而且每个样本数据的消耗时间均在 1 ms 以内。

自研算法在进行完关键词提取之后，得到所有关键词的特征权值组成一个新的集合，其中一些关键词不是特别重要，在文章中出现的次数也比较少，很有可能是一些固定的名词，因此需要将其剔除掉。将其中出现次数较高的筛选出来。按照需求选取前几个出现次数高的作为候选关键词。将筛选出来的关键字保存到本地目录中，之后利用 jieba 筛选前 50 个出现次数最高的词语并且返回其权重值，最后遍历输出筛选到的词语。

对于计算出的新词概率设置一个合理阈值，筛选出大于其阈值的新词，与原有分词词典进行匹配。若该词已被收录在原有词库中就结束；若该词不在原有分词字典当中，就进行添加操作，将关键新词添加到分词字典中去[6] [7]。

Table 3. Experimental results of self-developed algorithms
表 3. 自研算法实验结果

序号	产品	样本数量	匹配结果	平均耗时
1	刘能	冀 FB97XX	匹配成功	1 ms
2	赵佳俊	冀 FB98XX	匹配成功	1 ms
3	张鑫	冀 FB96XX	匹配成功	<1 ms
4	吴某军	冀 FB95XX	匹配成功	1 ms

通过四种方法的对比测试可以得到最终的对比实验总结，如下表 4 所示。首先可以排除关键词标签提取工具和 TextRank 算法工具两种；运用 TF-IDF 算法和 jieba 分词的语音文件和文本文件关联匹配技术的性能是最好的，其适用的范围和业务范围关联度比较高，比如涉车的违法行为用车牌号匹配、涉及到驾驶员的违法行为用从业资格证号匹配。

因此采用本文研究的基于 TF-IDF 和 jieba 分词的交通运输综合执法语音文件和文本文件关联匹配技术可以满足当前的交通运输执法业务需求。

Table 4. Summary of experiments
表 4. 实验总结

序号	姓名	车牌号	匹配成功数量	平均耗时
1	百度智能云	4	4	1340 ms
2	关键词标签提取工具	4	C/S 架构，未来无法与综合执法系统相结合	-
3	TextRank 算法工具	4	0	15.5 ms
4	自研算法	4	4	100 ms

5. 结束语

本文提出的基于 TF-IDF 和 jieba 分词的语音文件和文本文件关联匹配技术，结合文本文件的特点和内容格式，通过 TF-IDF 算法对交通运输综合执法案件的文书按照业务需要进行重新分析整理，形成新的关键字，并且对 jieba 分词的分词库进行更新，以保证分词结果能够满足后续的关键信息提取的需要，提升匹配的效率和准确度。

通过和其它相关技术的测试分析对比得出本文提出的自研算法更优，自研算法在数据采集和关键词

提取层面相较于目前相关技术得到了进一步优化，有效缩短了文本匹配时间，准确率大大提高。更能满足当前交通运输综合执法部门的需求，为交通执法人员对于听证后的文书与文件的管理匹配提供了一定的工程价值。

基金项目

北京市交通行业科技项目(0686-2241B1251413Z)。

参考文献

[1] 王婧. 基于 TF-IDF 与 Word2vec 的新闻热点分析[J]. 中国有线电视, 2023, 451(2): 59-63.

[2] 梁尘逸, 姚远哲. 基于异构信息网络与 TF-IDF 的核心药物发现算法[J]. 计算机时代, 2023, 371(5): 31-35. https://doi.org/10.16644/j.cnki.cn33-1094/tp.2023.05.007

[3] Yang, Z., Dai, Z., Yang, Y., *et al*. (2019) XLNet: Generalized Autore-Gressive Pretraining for Language Understanding. *Proceedings of the* 31*th Conference on Advances in Neural Information Processing Systems*, Long Beach, 8 September 2019, 5754-5764.

[4] 陈铭. 面向微博的文本质量评估与分类技术研究与实现[D]: [硕士学位论文]. 长沙: 国防科学技术大学, 2015.

[5] 金宇杰, 袁明. 基于 TF-IDF 算法的新词发现系统原理与实现[J]. 信息化研究, 2020, 46(5): 39-44.

[6] 柳文婷. 基于改进互信息的微博新情感词提取[J]. 延边大学学报(自然科学版), 2019, 45(4): 349-355.

[7] 王欣. 一种基于多字互信息与邻接熵的改进新词合成算法[J]. 现代计算机(专业版), 2018(11): 7-11.