计算机科学
COMPUTER SCIENCE

基于本体驱动的航空情报表格信息结构化研究

赖欣, 李思宁, 梁昌盛, 张恒嫣

引用本文

赖欣, 李思宁, 梁昌盛, 张恒嫣. 基于本体驱动的航空情报表格信息结构化研究[J]. 计算机科学, 2024, 51(6A): 230800150-7.

LAI Xin, LI Sining, LIANG Changsheng, ZHANG Hengyan. Ontology-driven Study on Information Structuring of Aeronautical Information Tables [J]. Computer Science, 2024, 51(6A): 230800150-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230800150

基于本体驱动的航空情报表格信息结构化研究

赖 欣 李思宁 梁昌盛 张恒嫣
中国民用航空飞行学院 四川 广汉 618307
(lxrzg@163.com)

摘 要 航空资料汇编是国际民航组织推荐的呈现各国航空信息的主要载体,其中以表格数据形式汇总了大量航空数据与航空运行限制信息。为实现航空汇编资料的智能查询,以及对航空资料汇编中静态数据的挖掘与利用,需要对航空汇编资料中的表格信息予以特征提取与结构化处理。将航空资料汇编中表格信息作为研究对象,提出了一种基于本体驱动的航空情报表格信息结构化抽取方法。首先构建航空情报领域信息的本体框架,实现对领域知识统一规范的描述;其次,利用 Document AI 对表格文档的布局结构进行研究与预处理,并利用随机森林算法与条件随机场模型进行特征实体提取验证与分析。实验结果表明,所提方法能够有效提取航空情报表格中的特征实体,为航空情报领域静态数据深入挖掘提供参考。

关键词:航空情报;本体;命名实体识别;条件随机场;随机森林;Document AI
中图分类号 TP391;V355

Ontology-driven Study on Information Structuring of Aeronautical Information Tables

LAI Xin,LI Sining,LIANG Changsheng and ZHANG Hengyan
Civil Aviation Flight University of China,Guanghan,Sichuan 618307,China

Abstract The aeronautical information publication(AIP) is the main carrier recommended by ICAO to present aeronautical information of all countries,in which a large amount of aeronautical data and aeronautical operation restriction information exists in the form of table information. In order to achieve intelligent querying of AIP and to facilitate the extraction and utilization of static data within it,it is necessary to perform feature extraction and structural processing on the tabular information within AIP. In this paper,an ontology-driven structured extraction method for aeronautical information tabular data is proposed,taking tabular data in AIP as the research object. Firstly,the ontology framework of aeronautical information is constructed to realize a unified and standardized description of domain knowledge. Secondly,the layout structure of form documents is studied and preprocessed using Document AI,and the feature entity extraction is verified and analyzed using random forest algorithm and conditional random field model(CRF). Experimental results show that the proposed method can effectively extract the feature entities in AIP,and provide reference for the in-depth mining of static data in the field of aeronautical information.

Keywords Aeronautical Information,Ontology,Named entity recognition,Conditional random field model,Random forest,Document AI

航空情报是支持航空运行的重要信息源,蕴含了与航空运行安全相关的海量、实时的数据,主要包括实时更新的航空动态信息——航行通告(Notice to Airmen,NOTAM),以及全球各国以 28 天为周期更新的航空资料汇编(AIP),为航空运行的规划与实时决策奠定了基础。“十四五”期间,中国民用航空局提出了通过“数字化强基、智能化应用、智慧化融合”三个阶段,在“智慧民航”的建设主线下,构建数字化航空情报产品数据集已经成为必然趋势。未来航空情报专业将完成从航空情报服务(AIS)向航空情报管理(AIM)过渡,强调以数字形式驱动信息管理的所有进程。目前世界各国海量航空情报信息的判读均以人工为主,实现数字化技术的突破是航空情报专业的重点任务,而构建数字化航空情报产品需要从传统的航空情报产品中进行数据获取、处理和结构化。目前航空情报主要以传统文本或图形图像的形式呈现,因此,对文档、图像信息的结构化分析和内容提取已经成为传统航空情报信息数字化转型成功的关键[1]。而航空资料汇编作为主要的航空情报产品,包含结构复杂、多样的表格和文本信息,采用统一标准对其进行结构化信息提取具有必要性。

本体作为一种形式化的知识表示模型,能够对领域内知识进行统一化、规范化的描述[2]。航空情报领域本体模型的建立将有助于实现其领域知识的规范化描述,为后续的信息提取与结构化打下坚实的基础。目前国内外学者多采用基于本体、树结构等方式进行表格信息抽取,而航空情报领域存在大量的文档信息,用户对相关航空数据的查找、判读费时费力。目前航空情报领域相关研究主要集中在航行通告的结构化信息抽取上,尚未有关于航空情报表格结构化的研究。因

基金项目:四川省自然科学基金(2023NSFSC0903);中央高校校级重点项目(ZJ2023-003)
This work was supported by the Natural Science Foundation of Sichuan Province,China(2023NSFSC0903) and Key Program of the Central Universities at the School Level(ZJ2023-003).
通信作者:李思宁(1825614449@qq.com)

此采用基于本体驱动的方法并结合文档人工智能技术、命名实体识别方法对 AIP 表格文档进行结构化信息抽取，可以缩短大量的人工处理时间并支持不同的下游业务场景[3-4]。

本文将以航空资料汇编中的表格信息为研究对象，首先利用全球标准化结构的航空信息交换模型(Aeronautical Information Exchange Model，AIXM)构建进行结构化特征信息抽取的本体框架模型；其次采用文档布局分析与命名实体识别相结合的方法，利用正则表达式进行符合 AIXM 标准的结构化的表格数据提取，对不符合 AIXM 标准的表格数据部分采用基于序列标注的条件随机场模型(CRF)与随机森林算法进行特征实体抽取。

1 相关工作

1.1 本体建模

本体(Ontology)的概念源自哲学领域，在哲学中的定义为“对世界上客观事物的系统描述，即存在论”[5]。当前学界对本体还没有统一的定义，但是达成了一些基本共识，例如本体包括概念化、明确性、形式化、共享、描述领域知识等特征[6]。近年来，随着知识图谱、语义网等研究热度的攀升，本体在知识图谱、推荐系统、信息检索、生物医学等领域得到了广泛关注[7]。

目前本体构建的方法主要有本体论工程法和基于叙词表的构建方法两大类，针对不同应用项目，国内外学者提出了 IDEF5 法、骨架法、TOVE 法、METHONTOLOGY 法、KACTUS 工程法、SENSUS 法、七步法等各类本体构建方法[8-9]。Zhou 等[10]通过总结归纳当前军事领域的本体并结合七步法提出了一种军事领域的本体构建流程。Ding 等[11]基于叙词表向领域本体转化的一系列问题，提出了从叙词表向本体转化的理论实践方法。AIXM 作为国际标准的航空数据交换模型，以数字化方式将航空信息领域的各种概念描述成要素、属性和关系的集合，其中包括空域、机场、飞行程序、航路、服务等航空元素。面对复杂的航空情报信息，构建基于 AIXM 的本体框架作为知识表示和数据处理的中心，将有助于提高数据的可理解性、一致性，实现更丰富的语义化表示和信息结构化，从而支持更智能化的航空情报处理和应用。

1.2 Document AI

Document AI 主要包括文档布局分析、文档信息抽取、文档视觉问答、文档图像分类 4 个主要的任务。目前，Document AI 技术已经从早期的基于规则的启发式方法发展到统计机器学习、深度学习方法，大大提高了分析性能和准确性[12]，其基于深度学习的基本框架如图 1 所示。

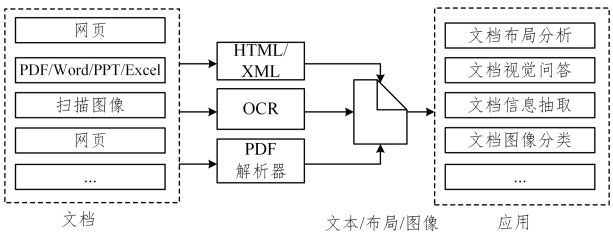

图 1 Document AI 概述

Fig. 1 Overview of document AI

Yang 等[13]将文档语义结构分析任务视为一个逐像素的分类问题，并提出了一个多模态神经网络来同时考虑视觉和文本信息。Prasad 等[14]使用 Cascade R-CNN 模型同时进行表检测和表结构识别。虽然 Document AI 技术已经存在一定的探索，但是仍旧存在多页/跨页、训练数据质量不均、多任务相关性不强等问题。因此将 Document AI 技术与其他技术相结合将能更加有效地实现自动化、结构化的提取文档信息。

1.3 命名实体识别

命名实体识别旨在从非结构化文本中自动获取预定类别的实体知识，可应用于许多 NLP 下游任务中，如信息检索、知识图谱、问答系统、舆情分析、推荐系统等[15]。

目前，国内外学者在命名实体识别的研究中主要采用了基于模板和规则、基于机器学习、基于深度学习的方法[16-17]，主要包括 CRF，bert 和 bilistm-CRF 等。Zhu 等[18]提出了一种规则与统计相结合的中文微博命名实体识别方法，通过选取合适的特征模板，并利用条件随机场模型来进行实体识别。Liu 等[19]提出了基于持续学习的命名实体识别技术框架，并获得了良好的提取效果。

航空情报领域的命名实体识别的主要目的在于识别航空情报相关的实体信息。例如空域名称、空域坐标、航线等实体，存在专有名词较多、实体全称与缩写混合表示的困难，以及缺乏航空情报领域大规模训练数据集的问题。因此本文将采用基于监督学习的条件随机场模型与随机森林模型，通过人工标注数据集来进行命名实体识别，抽取关键特征信息。

1.4 基于监督学习的实体识别模型

监督学习(Supervised Learning)是机器学习中的一种常见方法，其目标是从已标注的训练数据中学习一个模型，用于对未知数据进行预测或分类。本文选择了监督学习中的条件随机场和随机森林模型原理进行实体识别。

1)条件随机场(Conditional Random Field，CRF)是给定一组输入随机变量条件下另一组输出变量的条件概率分布模型，其结构如图 2 所示。

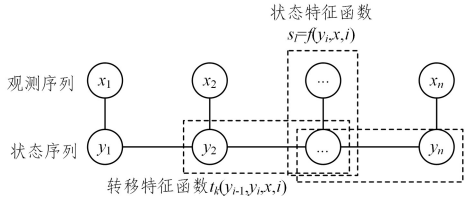

图 2 CRF 模型结构图

Fig. 2 Structure of CRF model

其基本原理如下：

设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$，$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)$ 均为线性表示的随机变量序列，在随机变量 X 取值为 x 的条件下，随机变量 Y 取值为 y 的条件概率形式如下[14]：

$$P(y|x)=\frac{1}{Z(x)}\exp\{\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\} \tag{1}$$

$$Z(x)=\sum_y\{\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\} \tag{2}$$

其中，t_k 和 s_l 为特征函数；λ_k 和 μ_l 为对应权值；$Z(X)$ 为归一化因子。

2)随机森林(Random Forest，RF)是基于决策树模型设计的一种 Bagging 集成模型。随机森林被定义为一个由一系列决策树组成的分类器，基于这些树的投票，随机森林将给出最终的集成结果[20]。随机森林的示意图如图 3

所示，具体步骤如下：

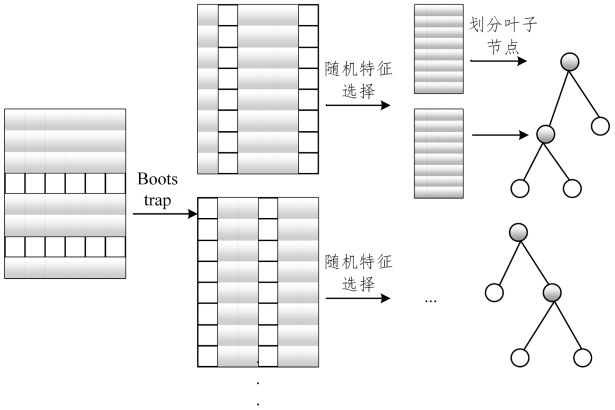

图 3　随机森林算法示意图

Fig. 3　Schematic diagram of random forest algorithm

基于给定的数据集：

$D=\{(\boldsymbol{X}_1,\boldsymbol{Y}_1),(\boldsymbol{X}_2,\boldsymbol{Y}_2),\cdots,(\boldsymbol{X}_m,\boldsymbol{Y}_m)\}$

1)对每个 $t=1,2,\cdots,T$(T 为决策树的个数)

步骤 1　对原训练集进行 Bootstrap 采样，得到新训练集 D_b，并基于 D_b 来训练一棵决策树。

步骤 2　从训练集 D_b 包含的所有特征中，随机选取 K 维特征。

步骤 3　基于 K 维特征，寻找决策树最优划分点，将样本划分到两个子节点中。

步骤 4　重复执行步骤 2、3，直到所有节点都无法被继续划分。

2)基于 T 棵决策树的预测结果进行投票，最终的投票结果即作为最终的集成结果。

2　航空资料汇编结构分析

航空资料汇编(AIP)是由国家或国家授权发行、载有空中航行必需的具有持久性质的航空资料出版物，是一体化航空情报系列资料的一个组成部分，由总则(GEN)、航路(ENR)、机场(AD)3 个部分组成。每一部分根据需要分为若干章节，分别包含同的航空情报资料。

空中交通服务空域部分包含情报区、管制区、终端管制区和进近管制区等信息，以 PDF 表格、图、文本等形式存在。航空资料汇编表格数据虽然以特定的分类进行存储，其中仍然包含自由文本，存在结构化、半结构化、非结构化数据。管制区表格数据如图 4 所示，主要包含名称、水平范围、垂直范围、提供服务的单位、呼号、语言、频率(备用频率)、服务时间和备注。其中名称、水平范围、垂直范围部分包含部分自由文本，以半结构化的形式存储在同一单元格。航路相关表格数据如图 4 所示，主要包括航路代号、重要点名称、类型、识别和重要点坐标两大类，其中部分运行限制数据以非结构化的中、英文自由文本存储。这两类文本结构特征包含词级特征、句子级特征、上下文特征，其中空域类数据由大小写字母、数字组成，并且包含固定长度的坐标数据词级特征较为明显。

因此本文设计了一种基于本体驱动的航空情报表格信息结构化抽取方法，将依据本文搭建的基于 AIXM 的本体模型，分别采用正则表达式与命名实体识别进行特征实体抽取，对航空资料汇编表格数据实现自动化信息抽取，形成符合 AIXM 结构化的航空情报数据。

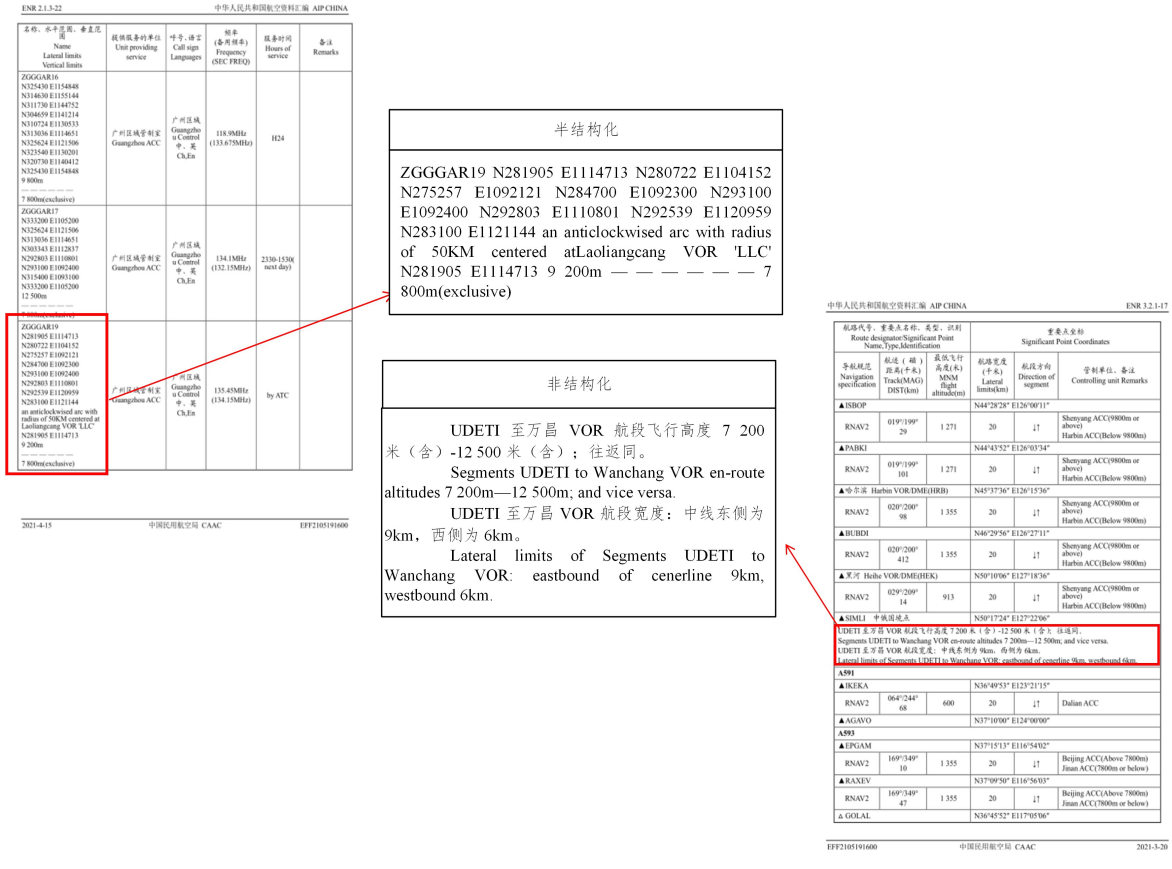

图 4　航空资料汇编部分表格示例

Fig. 4　Examples of selected tables from aeronautical information publication

3 航空资料汇编表格信息抽取方法

根据航空资料汇编中表格数据的结构特点，结合 Document AI 技术与命名实体识别技术，本文设计了一套航空资料汇编表格信息抽取的流程，如图 5 所示。本文提出了一种基于本体驱动的航空情报表格信息结构化抽取方法，搭建了基于 AIXM 的本体框架，并分别对 AIXM 标准下的结构化数据、半结构化和非结构化等限制数据进行特征实体提取。

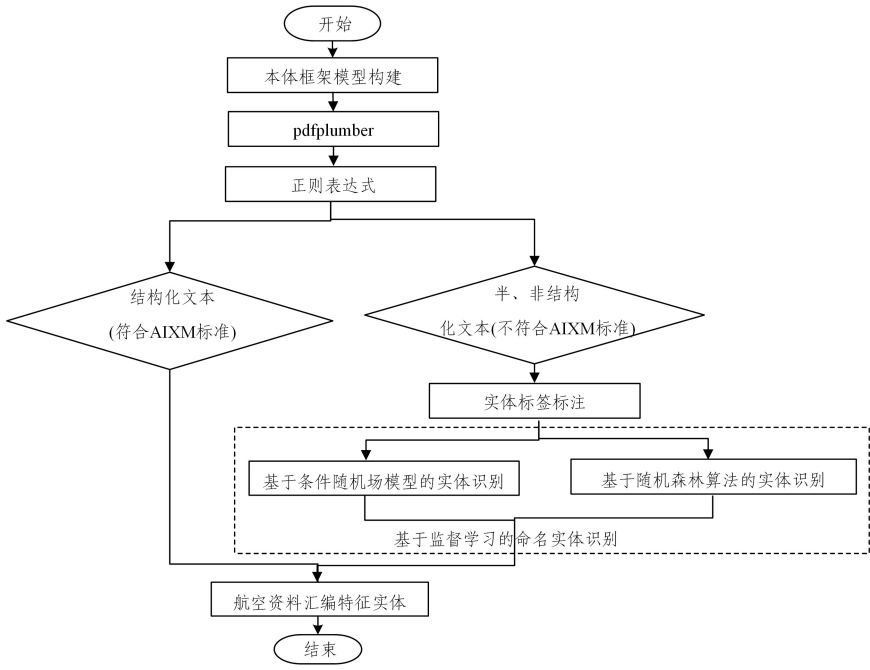

图 5 航空资料汇编表格信息抽取流程图

Fig. 5 Flowchart of information extraction from aeronautical information publication form

3.1 基于 AIXM 的本体框架模型

AIXM 是为了便于计算机系统自动化处理、交换航空数据而定义的数据标准，本文利用全球标准化结构的航空信息交换模型（AIXM）构建结构化特征信息抽取模型，通过引入本体模型并结合 AIXM 模型对航空情报领域内的知识进行规范化描述，由于航空情报信息量巨大，因此以空域类航空情报数据为例搭建了本体框架，如图 6 所示。

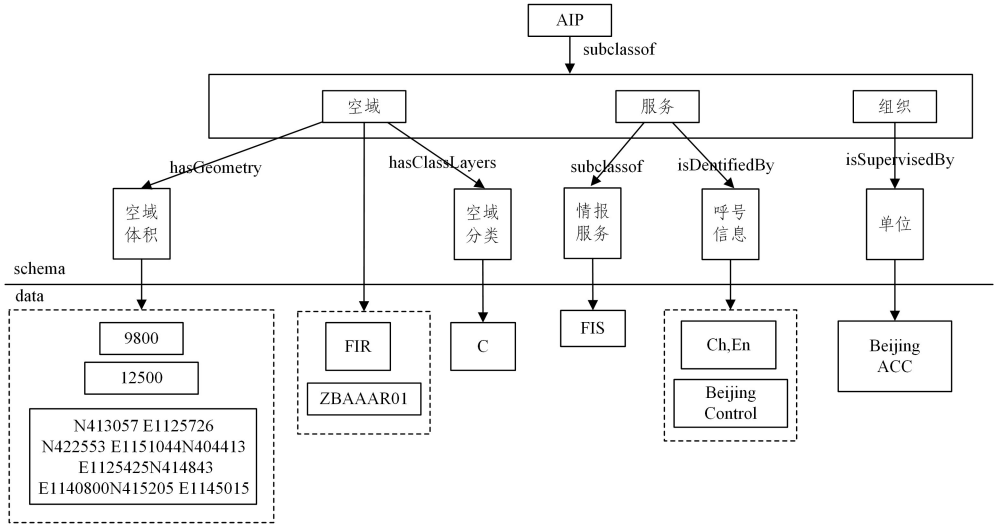

图 6 基于 AIXM 的本体框架

Fig. 6 AIXM-based ontology framework

3.2 正则表达式

本文依据基于 AIXM 构建的本体框架模型进行表格数据的信息抽取，首先采用 pdfplumber 解析 PDF 文件进行文档布局分析，检测并提取表格数据结构及其文本内容，其解析流程如图 7 所示。

对于处理后的表格数据，本文采用正则表达式提取符合 AIXM 标准的结构化数据。正则表达式是由一系列字符和特殊字符组成的模式，用于匹配和查找文本中的字符组合，广泛应用于文本搜索、文本替换、数据提取等领域。结合航空资料汇编的表格数据结构，本文采用正则表达式的方法分析结构化英文表格数据，以管制区、航路部分表格数据为例，设计了以下正则表达式匹配模式。

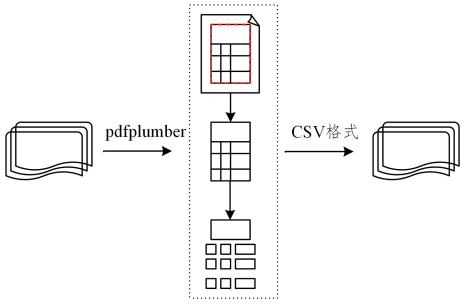

图 7　pdfplumber 流程图

Fig. 7　pdfplumber flowchart

表 1　部分正则表达式模板示例

Table 1　Examples of some regular expression templates

匹配内容	正则表达式
呼号 例：Guangzhou Control	\b[A-Za-z]+\sControl\\b
语言 例：Ch，En	Ch，En
管制单位 例：Guangzhou ACC	\b[A-Za-z]+\sACC\b

3.3　基于条件随机场模型的命名实体识别

由于航空情报表格数据中存在半结构化、非结构化文本数据，因此本文采用了条件随机场模型对无法通过正则表达式提取的半结构化、非结构化数据进行命名实体识别并提取特征实体。

CRF 模型在 NER 任务中能够灵活地捕捉上下文信息及序列内部的相关性，并且能够考虑序列的全局标注，不受人工标注可能导致的局部标注的错误累积影响[21]。因此本文选择使用 CRF 模型来进行航空情报表格数据的命名实体识别，主要分为以下 3 个步骤：

1）实体标注

本文选择使用 BIO 标签标注方法，分别对经过数据处理的航空资料汇编中管制区、终端区和进近区表格数据进行序列标注。

2）确定特征函数

在式(1)中，$t_k(y_{i-1},y_i,x,i)$是定义在 i 和 $i-1$ 之间边上的转移特征函数，$s_l(y_i,x,i)$是定义在结点上的状态特征函数。定义转移特征函数和状态特征函数取值为 1 或 0，当满足特征条件时取值为 1，否则为 0[22]。在 CRF 中特征模板的选取可以确定特征函数，合适的特征模板可以为特征函数的生成提供一个统一的模式。针对航空资料汇编表格数据结构及内容特性，本文选择并设计了特征模板用于确定特征函数，如表 2 所列。

表 2　CRF 模型特征模板

Table 2　CRF model feature template

特征	描述
word. lower	单词的小写形式
word[−3:]	单词的最后 3 个字符
word[−2:]	单词的最后两个字符
word. isupper	单词是否全为大写
word. istitle	单词是否以大写字母开头
word. isdigit	单词是否为数字
−1:word. lower	前一个单词的小写形式
−1:word. istitle	前一个单词是否以大写字母开头
−1:word. isupper	前一个单词是否全为大写
+1:word. lower	后一个单词的小写形式
+1:word. istitle	后一个单词是否以大写字母开头
+1:word. isupper	后一个单词是否全为大写

3）模型训练

本文将标注后的数据作为训练数据集，每个样本是一个文本序列，每个序列都有相应的标签，用于表示实体的起始位置和实体类型。将此训练集作为模型输入，通过对单词的不同特征进行提取和编码来将其转换成特征向量，并将每个单词的特征向量组织成特征矩阵 $\boldsymbol{X}$，将每个单词的命名实体标签组织成标签矩阵 $\boldsymbol{Y}$。根据给定的输入序列 $\boldsymbol{X}$ 和输出序列 $\boldsymbol{Y}$，通过定义的特征模板来确定特征函数并选择 LBFGS 优化算法学习、更新权重。从而计算在输入文本序列 $\boldsymbol{X}$ 的条件下输出标签序列 $\boldsymbol{Y}$ 的条件概率 $P(\boldsymbol{Y}|\boldsymbol{X})$。

3.4　基于随机森林的实体识别

为验证 CRF 模型进行命名实体识别的效果，本文选择了随机森林算法作为对比，对相同的数据进行命名实体识别。由于在实体识别任务中，随机森林通过组合多个决策树的预测结果来提高整体的准确性和泛化能力，它可以有效地处理高维特征和大规模数据集，并且对于处理文本数据中的实体识别任务具有一定的优势[22]。因此本文采用根据随机森林算法构建的实体标签，将经过处理和标注的数据作为训练数据集来进行命名实体识别。

1）词汇向量化：将标签数据集中的每个单词通过特征映射函数转换为特征数组，将数据集中的标签列转换为标签列表，并将特征数组作为预测模型的输入特征，将标签列表作为预测模型的目标标签列。即数据集：

$$D=\{(\boldsymbol{X}_1,\boldsymbol{Y}_1),(\boldsymbol{X}_2,\boldsymbol{Y}_2),\cdots,(\boldsymbol{X}_m,\boldsymbol{Y}_m)\}$$

其中，$\boldsymbol{X}_i$ 是单词的特征数组。

2）训练随机森林模型：使用标签数据集训练随机森林模型。首先从原始数据集 D 中采取有放回的抽样（bootstrap），选取子数据集，构造并训练 20 棵决策树。其次，根据输入样本的特征数随机选取一定的特征，在随机选取的特征中选择最优的特征，根据多数投票来确定最终的实体标签。

4　实验与结果分析

实验分别采用一期管制区、终端区和进近管制区的表格数据集进行测试，验证所提基于本体驱动的航空情报表格信息结构化抽取方法的特征实体提取效果。本文采取了两种命名实体识别算法，分别对管制区、终端区和进近区的两个表格数据集进行特征实体提取，并通过计算 F1 值评价指标对模型提取效果进行对比。

4.1　数据处理与标注

采用 pdfplumber 对空域类表格数据进行文档布局分析提取表格数据，并对提取的表格数据采用正则表达式的方法来提取符合 AIXM 标准的结构化表格数据。

将数据进行初步处理之后，将采用正则表达式的方法提取的数据以 CSV 格式进行存储。本文选取 BIO 标签标注方法来对数据集进行标注，其中 B 表示实体的开端，I 表示实体的中间位置，O 表示非实体[23]。采用 doccano 软件对不符合 AIXM 标准格式的部分进行人工标注处理。对于数据集的标注，本文参考 AIXM 中空域类实体、属性构成，将数据的标注标签分为 8 个类别，如表 3 所列。

表 3 实体标签标注释义

Table 3 Interpretation of entity labeling

标注类别	说明	标注类别	说明
name	空域名称	radius	圆形空域的半径
geo	空域的水平范围	center	圆形空域的圆心
Upper Limit	空域的上限	runway	终端和进近区涉及的跑道
Lower Limit	空域的下限	Airport Type	终端和进近区涉及的机场

4.2 实验设置

模型基于 python 代码编程实现，实验参数设置：对于条件随机场模型，采用 LBFGS 算法作为优化算法，设置 L1 和 L2 正则化项的系数为 0.1，限制最大迭代次数为 200 次，对两个数据集进行训练和预测。对于随机森林算法，将参数设置为 20 个决策树。两个模型均将数据集使用 5 折交叉验证进行评估。实验采用 F1 值作为模型效果评价指标，NER 任务中常用的评价指标有准确率 P、召回率 R 和 F_1 值[18]，计算式如下[24-25]：

$$P=\frac{T_P}{T_P+F_P}$$

$$R=\frac{T_P}{T_P+F_N}$$

$$F_1=\frac{2PR}{P+R}$$

其中，T_P 指正确匹配的数目，F_N 指将本身正类预测为负类的数量，F_P 指将本身负类预测为正类的数量。

4.3 实验结果分析

实验以准确率 P、召回率 R 和 F_1 值为评价指标，将两种模型分别在两个数据集下的实验效果进行对比，实验结果如表 4 所列，其准确率、召回率、F_1 值对比结果如图 8 所示。

表 4 实验结果

Table 4 Experiment results

(%)

	CRF		随机森林	
	管制区	终端区和进近区	管制区	终端区和进近区
P	86	80	72	54
R	88	76	74	51
$F1$	87	77	73	50

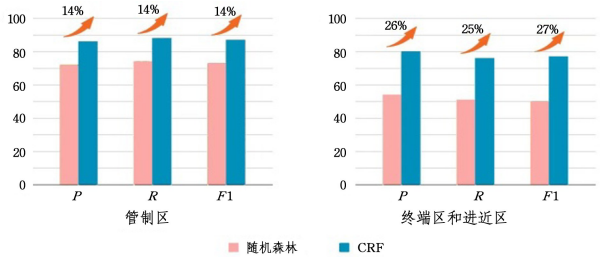

图 8 实验结果对比

Fig. 8 Comparison of experimental results

1)管制区数据集：采用 CRF 和随机森林模型进行管制区的特征提取时，基于 CRF 模型的提取效果优于随机森林算法，其 F_1 值达到了 87%，而基于随机森林算法的 F_1 值为 73%。

2)终端区和进近区数据集：采用 CRF 和随机森林模型进行终端区和进近区的特征提取时，基于 CRF 模型的提取效果也优于随机森林算法，其 F_1 值达到了 77%，基于随机森林算法的 F_1 值为 50%。

CRF 模型在不同数据集上的整体提取效果都优于随机森林算法，对各种标签类型的实体的命名实体识别效果均比较理想，适用于航空情报表格数据的提取。

结束语 航空资料汇编中的表格数据包含大量航空数据与航空运行限制信息，对其表格数据的提取有助于辅助航空情报工作人员进行智能决策。本文采用 Document AI 技术与命名实体识别技术对此类表格信息进行特征提取，提出了一种基于本体驱动的航空情报表格信息结构化抽取方法。本文首先构建了基于 AIXM 的本体框架模型，提取符合 AIXM 结构的数据；其次采用 pdfplumber 进行文档布局，分析提取表格数据，针对结构化数据，采用正则表达式进行提取，针对非结构化、半结构化数据，采用 CRF 模型与随机森林模型对其进行特征实体抽取。实验采用空域类表格数据进行实验分析，结果表明，本文所提表格信息提取方法具有良好的提取效果，其 F_1 值可达到 87%。

本研究为航空信息数字化转换提供了一种新的技术思路。后续研究将丰富表格类型，利用多样数据源进行实验验证，并结合深度学习模型、知识图谱进一步进行航空实体关联关系研究，为实现基于航空信息大数据处理下的多应用领域做铺垫。

参 考 文 献

[1] CUI L, XU Y H, LV T C, et al. Document AI: Benchmarks, Models and Applications[J]. Journal of Chinese Information Processing, 2022, 36(6): 1-19.

[2] SUN S D. Research on semantics knowledge organization of historical newspaper resources in digital humanities Research on Sem[D]. Jilin: Jilin University, 2022.

[3] ZHANG Y T, LI Q Y, LIU S K. Tabular subordination relation extraction based on graph convolutional network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(4): 1308-1315.

[4] TANG R, DENG J X, YE Z X, et al. Survey of Table Extraction in PDF Documents[J]. Computer Applications and Software, 2021, 38(7): 1-7, 22.

[5] SHEN Y F. Construction and Intelligent Application of Public Security Knowledge Graph Model Based on Multi-source Heterogeneous Data[J]. Police Science Research, 2021(5): 79-89.

[6] YU F. Methodothology and empirical research on Domain Ontology—A case of Geomatics[D]. Wuhan: Wuhan University, 2013.

[7] LI A H, XU Y Z, CHI Y X. Review of Ontology Construction and Applications[J/OL]. Information Studies: Theory & Application: 1-9[2023-08-09].

[8] WANG Y L, ZOU J F, WANG K, et al. Injection Molding Knowledge Graph Based on Ontology Guidance and its Application to Quality Diagnosis[J]. Journal of Electronics & Information Technology, 2022, 44(5): 1521-1529.

[9] TANG A M, ZHEN Q, FAN J. Thesaurus-based Approach to Build Domain Ontology[J]. Data Analysis and Knowledge Dis-

covery，2005(4)：1-5.

[10] ZHOU Y W，YANG C H，WANG H Y. Ontology construction of military field[J]. Computer Era，2022(9)：96-99.

[11] DING S C，FU Z. Research on Semi-automatic Construction of Domain Ontology Based on Space Thesaurus[J]. Information Studies：Theory & Application，2011，34(11)：113-116.

[12] SUN X，REN X Y，ZHENG H C，et al. Domain Named Entity Recognition Method Based on Parameter Transfer Learning[J]. Technology Intelligence Engineering，2022，8(3)：13-27.

[13] YANG X W，YUMER E，ASENTE P，et al. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2017：4342-4351.

[14] PRASAD D，GADPAL A，KAPADNI K，et al. Cascadetabnet：An approach for end to end table detection and structure recognition from imagebased documents[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020：572-573.

[15] FENG Y T，ZHANG H J，HAO W N. Named Entity Recognition for Military Text[J]. Computer Science，2015，42(7)：15-18，47.

[16] GAO X，TANG J Q，ZHU J W，et al. Study on Named Entity Recognition Method Based on Knowledge Graph Enhancement [J]. Computer Science，2023，50(S1)：112-117.

[17] KRUENGKRAI C，NGUYENT H，ALJUNIED S M，et al. Improving LowResource Named Entity Recognitionusing Joint Sentenceand Token Labeling[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020：5898-5905.

[18] ZHU R，YANG L C，DING W X，et al. Study on Named Entity Recognition Method Based on Knowledge Graph Enhancement [J]. Journal of Central Journal of Central China Normal University(Natural Sciences)，2018，52(3)：316-321.

[19] LIU P F，QIAN L，ZHAO X W，et al. Continual learning framework of named entity recognition in aviation assembly domain [J]. Journal of Zhejiang University(Engineering Science)，2023，57(6)：1186-1194，1266.

[20] LIN B，WU S B，ZOU Y J，et al. Individual Travel Behavior Prediction of Hong Kong-Zhuhai-Macao Bridge Based on Combinatio on of BLSMOTE Algorithm and Random Forest Model[J]. Traffic & Transportation，2023，39(2)：37-43.

[21] GAO X，WANG S，ZHU J W，et al. Overview of Named Entity Recognition Tasks[J]. Computer Science，2023，50(S1)：26-33.

[22] YANG Z W. Research on Named Entity Recognition Methods for Unstructured Text[D]. Jilin：Jilin University，2023.

[23] XU M X. Application of named entity recognition technology in epidemiological investigation[D]. Guizhou：Guizhou Normal University，2022.

[24] KRUENGKRAI C，NGUYENT H，ALJUNIED S M，et al. Improving LowResource Named Entity Recognitionusing Joint Sentenceand Token Labeling[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020：5898-5905.

[25] WANG J，SHOU L，CHEN K，et al. Pyramid：A layered model for nested named entity recognition[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020：5918-5928.

LAI Xin，born in 1977，Ph.D，associate professor. Her main research interest is aeronautical information services and management.

LI Sining，born in 1998，postgraduate. Her main research interest is traffic and transportation.