引文格式：黄梓航, 陈令羽, 蒋秉川. 基于文本解析的栅格类图表知识抽取方法[J]. 地理空间信息, 2023, 21(10): 23-27. doi:10. 3969/j.issn. 1672-4623.2023.10.006

基于文本解析的栅格类图表知识抽取方法

黄梓航[1]，陈令羽[2*]，蒋秉川[2]
（1. 战略支援部队信息工程大学 研究生院，河南 郑州 450001；
2. 战略支援部队信息工程大学 地理空间信息学院，河南 郑州 450001）

摘　要：栅格类图表包含了丰富的时空含义，文本信息是其知识特征的主要体现。提出一种基于文本解析的栅格类图表知识抽取方法。首先，通过CRAFT+CRNN的方法，进行图表文本信息的定位与矢量化；其次，在分析大规模样本数据的基础上，解析了常见图表数据的结构特征，并提出采用模板和机器学习2种方法进行文本块的结构类别判断；最后，利用自然语言处理方法，构建栅格类图表的知识标签体系。实验表明，该方法在栅格类图表的文本定位矢量化、文本块的结构判断等方面具有较高的准确率，能够进行栅格图表的知识标签体系构建，抽取图表中包含的知识，为海量数据的快速检索和理解提供支撑。

关键词：栅格图表；文本解析；知识抽取；矢量化；结构判断；自然语言处理
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2023）10-0023-05

Knowledge Extraction Method for Raster Chart Based on Text Parsing

HUANG Zihang[1], CHEN Lingyu[2], JIANG Bingchuan[2]
(1. Graduate School, Strategic Support Force Information Engineering University, Zhengzhou 450001, China;
2. Institute of Geospatial Information, Strategic Support Force Information Engineering University, Zhengzhou 450001, China)

Abstract: Raster charts contain rich spatio-temporal meanings while text information is the main embodiment of their knowledge characteristics. We proposed knowledge extraction method for raster chart based on text parsing. Firstly, we used CRAFT+CRNN to locate and vectorize the chart text information. Then, we figured out structural characteristics of common chart data, and put forward template and machine learning to judge the structural category of text blocks after analyzing numerous sample data. Finally, we constructed the knowledge label system of raster charts by using natural language processing. The results show that this method can improve accuracy in the vectorization of text location and the structure judgment of text blocks in raster charts, extract the knowledge contained in the chart, construct the knowledge label system of raster charts, and provide support for the rapid retrieval and understanding of massive data.

Key words: raster chart, text parsing, knowledge extraction, vectorization, structure judgment, natural language processing

图表将繁冗的数据以直观形象的形式展现出来，在地理信息表达和传输等领域应用广泛[1-2]。随着互联网技术的不断发展，泛在图表成为大众接收各类信息的主要渠道之一，包含了大量具有重要价值的各类信息，通过爬取解译可为具体应用提供数据支撑。图表数据和时空数据相结合，可以极大程度提升人类的空间感知和认知能力。特别是人工智能技术的不断进步，为计算机快速理解和应用这类图表数据提供了新的技术支持，丰富了图表数据的应用领域和层次。

网络数据中的泛在图表按照数据类型可以分为2种：一种是矢量型图表，数据可以通过访问对应的数据库直接进行提取；另一种是栅格型图表，本质上就是栅格型数据图像，文件中记录的是不同位置上的像素值，依靠计算机视觉技术进行信息提取。矢量数据的解析主要停留在数据层面的处理上，方法相对成熟，研究栅格类图表数据的智能解析成为热点和难点问题[3-7]。

图表数据在形成过程中经过了一定程度的综合，实现栅格图表的完全解译难度很大。针对海量泛在图表，能够快速根据特征标签检索出对应的信息，再进行针对性的数据解译更符合时空大数据的研究规律。本文通过分析图表结构，解析关键文本信息，抽取图表中的关键知识，构建图表特征标签体系，具体方法

收稿日期：2022-05-25。
项目来源：国家重点研发计划资助项目（2017YFB0503500）；国家自然科学基金重点资助项目（4213000495）。
第一作者简介：黄梓航（1990—），博士，讲师，研究方向为战场环境仿真与知识图谱，E-mail：zhengzhouhzh@126.com。
通信作者：陈令羽（1987—），博士，讲师，研究方向为空间数据处理与安全，E-mail：ccllyy123456@126.com。

如图1所示。

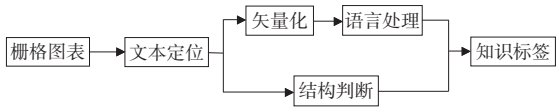


图1 栅格类图表数据知识标签体系构建流程

1 图表文本定位与矢量化

图表文本定位与矢量化是从栅格图表中提取出文字信息的位置和内容，将图表中的所有可能的知识识别出来，为知识标签判断提供基础，包括离线识别和在线识别2种方法。在线识别主要是借助主流的云工具，例如腾讯云、百度云、阿里云等，提供了相应的文字信息位置判断和识别的在线服务，能够为图像数据中文字信息的快速识别和分析提供相应接口，且具有很高的准确性和效率，但仅支持互联网情况下的识别，对应的离线工具价格昂贵。离线识别不需要连接互联网，主要通过文本检测和内容识别2个环节抽取文本信息[8-11]。综合分析现有方法和图表数据的特点，采用CRAFT+CRNN的方法进行文本定位与矢量化[12-14]。

1.1 基于CRAFT的文本检测与定位

基于CRAFT检测文字区域的基本原理是设计一个深度神经网络，通过预测单个字符的高斯热图以及字符间的连接性来检测文本。其主要思想是：

1）图像分割，采用u-net结构，先下采样再上采样，聚合深层和浅层特征。

2）非像素级分割，将一个字符视为一个检测目标对象，而不是一个词组（所有的词组都由若干个字符构成），即不把文本框当作目标。

3）标注时采用文本框级别的标注，而非字符级别的标注，并使用了一种弱监督学习思路（先利用合成样本进行预训练，再将预训练模型对真实数据集进行检测，得到预测结果，经过处理后得到高斯热度图作为真实数据集的字符级标签）。

由于CRAFT方法具有先进的性能和良好的泛化性，特别适合于栅格类图表中各类文本的检测和定位。

1.2 基于CRNN的文字内容矢量化

文字识别是对序列的预测方法，CRNN采用了对序列预测的RNN网络。通过CNN将图片的特征提取出来后采用RNN对序列进行预测，最后通过一个CTC的翻译层得到最终结果。主要采用的是CNN+RNN+CTC三层网络结构，从下到上，依次为：

1）卷积层，使用CNN，从输入图像中提取特征序列。

2）循环层，使用RNN，预测从卷积层获取的特征序列的标签（真实值）分布。

3）转录层，使用CTC，把从循环层获取的标签分布通过重新整合等操作转换成最终的识别结果。

泛在栅格类图表的质量难以得到保证，噪声和干扰信息多、图像不清楚的情况普遍存在。在进行文本矢量化之前，可先进行灰度变换、滤波操作、图像二值化等图像预处理，减少无关信息，在一定程度上恢复包含的文本特征信息从而增强特征可检测性和矢量化效果，提高提取的效率与准确率。

2 文本块结构判断

图表隶属于视觉传达信息的范畴，将知识、信息的概念和统计数据等进行视觉整理，用形状、颜色和文字、数字来表达，视觉化某种事物的现象或某种思维的抽象观念，以生动的表达形式传递信息，蕴含丰富的知识体系和内容，具有直观性和易懂性的特征。通过对图表的结构进行拆分，从各个文本块对应的结构特征上对图表的信息表达进行研究，是进行知识标签构建的关键环节[15]。

2.1 图表数据的结构解析

根据表现形式，可以将图表分为柱状类图表、线状类图表、点状类图表、多边形类图表和其他类图表等6种类型。图表蕴含的丰富信息，主要通过图表结构形式表现出来，不同的结构中包含的文字、数字内

表1 图表结构（元素）解析

序号	图表类型	图表结构元素
1	柱状类图表	标题（主标题、副标题）、图例、X轴、Y轴、绘图区（数据序列、数据标志、网格线）、版权信息、其他
2	线状类图表	标题（主标题、副标题）、图例、X轴、Y轴、绘图区（数据序列、数据标志、网格线）、版权信息、其他
3	点状类图表	标题（主标题、副标题）、图例、X轴、Y轴、绘图区（数据序列、数据标志、网格线）、版权信息、其他
4	饼状类图表	标题（主标题、副标题）、图例、绘图区（数据序列、数据标志）、版权信息、其他
5	多边形类图表	标题（主标题、副标题）、图例、绘图区（数据序列、数据标志）、版权信息、其他
6	其他类图表	标题（主标题、副标题）、图例、绘图区、版权信息、其他

容，能够解译出不同的信息内容。

知识标签的内容主要通过图表中的文字信息体现，不同位置的文字信息对应知识标签的价值含量也不相同。泛在图表的知识抽取是在图表结构解析的基础上，通过自动化的方式提取图表中的文字信息，并自动判别其对应的图表结构，判断信息的价值含量。基本思路是首先将图表中的所有文本信息识别出来并记录其对应的位置信息，然后根据统计规则或者机器学习的方式进行结构识别，提取对应的底层特征信息，并进行文本矢量化，抽取核心关键信息。

2.2　文本块的类别判断

文本块类别判断是在文本识别的基础上，根据文本块特征，基于相应类型图表的结构判断其对应的结构特征。主要有基于机器学习和基于模板匹配2种方法。

2.2.1　基于模板进行类别的快速判断

基于模板的类别判断主要是在文字信息识别和图表结构解析的基础上，进行文本块的类别判断。相较于普通图像，图表数据的设计和生成具有明显的规律性，特别是基于常用的软件工具生成时，提供了固定的模板，可以快速进行底层特征的类别判断。在总结分析大量图表数据的基础上，得到了主要图表数据的结构分布模板，表2为主要图表类型中图名和图例的分布情况统计。

表2　常用图表的结构模版

图表类型	图名	图例
柱状图	上左32.5%、上中65%、下中2.5%	左上25%、上中10.7%、右上17.8% 右中32%、右下3.5%、下中10.7%
饼状图	上左14%、上中79%、左下3.5% 下中3.5%	左上6%、上中6%、右上18% 右中52%、右下6%、下中12%
折线图	上左11%、上中83%、下中6%	左上14%、上中14%、右上9% 右中50%、下中14%
散点图	上左18%、上中68%、下中2%	左上12.5%、上中25%、右上12.5% 右中29%、下中21%
雷达图	左上10%、上中72.5%、右上7.5% 右下2.5%、下中7.5%	左上3.5%、上中10%、右上14% 右中3.5%、右下14%、下中24%

基于表2，结合文本块的位置信息，可以快速进行文本块的类别识别，并进行知识的抽取和利用。

2.2.2　基于机器学习的方法进行类别判断

基于模板进行文本块的类别判断便捷迅速，针对特定软件生成的图表数据准确率高，但是存在不同类别数据之间的模板位置有交叉，面向泛在图表数据的准确率会降低等问题。根据文本块的几何信息，基于统计学和机器学习的方法，判断其对应的图表结构要素，可以有效解决这一问题。主要步骤如下：

1）通过手工标注的表示，建立了特定类别图表的结构样本集。因为分布特征差异明显，不同类别的图表建立不同的结构样本集。

2）将标注的样本转化为特征数据，并计算特征值，选取文本块的起始点位置和长宽作为特征。记录图表数据中对应的位置信息、矢量化结果和对应结构类型。

3）构建CNN神经网络，对样本数据进行训练和测试，并进行新数据的结构预测。

根据2种方法的特点，可以发现基于模板进行类别判断的方法主要适用于规则图表，特别适用于特定软件自动生成的图表数据，能够快速准确地进行类别判断，不同软件生成的图表在结构模板上会有部分不同；基于机器学习的方法进行类别判断适用于各类泛在图表，具有较强的普适性，但是准确性和训练模型强相关。

3　知识标签体系的构建

图表的知识标签主要对应时间、地点、人物、事件、事物、现象、场景等社会七要素[16]，是表示图表某类关键知识特征的词语。知识标签抽取是结合文本块类型，对文本块的矢量化结果进行分词，并判断其对应的知识标签属性，具体过程如图2所示。

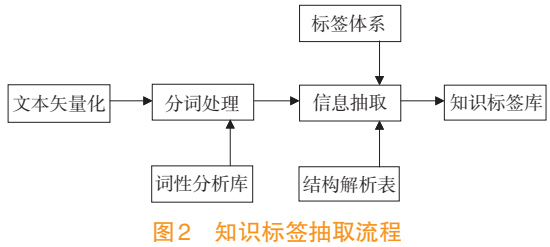


图2　知识标签抽取流程

将矢量化进行分词，并根据分词结果进行词性和特征标签判断是知识标签抽取的难点和重点，综合比较目前的分词工具，选用哈工大LTP分词工具，可进行词语属性的识别和判断。根据图表结构解析结果，设定两级标签体系：一级标签对应图名、图例两个图表结构中的知识标签，是图表的主要知识标签；二级标签对应图表区和坐标轴等其他内容，是图表的次要知识标签。

4 实验与分析

4.1 图表文本定位与识别

利用CRAFT方法，对栅格类图表进行文本定位，得到的结果如图3所示。

采用CRNN方法进行文字识别，为提高识别效率，首先将本文块进行二值化，然后再进行矢量化，得到的结果如图4所示。

分析图3、4可以发现：利用CRAFT+CRNN的方法对栅格类图表中的文本进行定位和识别，具有较高的准确度。分别对100张栅格类图表进行文本定位和识别测试，定位准确率和识别准确率均达到90%以上的图表数量为93张，说明了该方法的有效性。

4.2 结构类别判断

以柱状图和饼状图为例，通过手工标注，建立了图表结构样本集。其中，柱状图按照标题、X轴、Y轴、图例、图表区5类结构进行构建，训练集包括250个样本，测试集包括30个样本；饼状图按照标题、



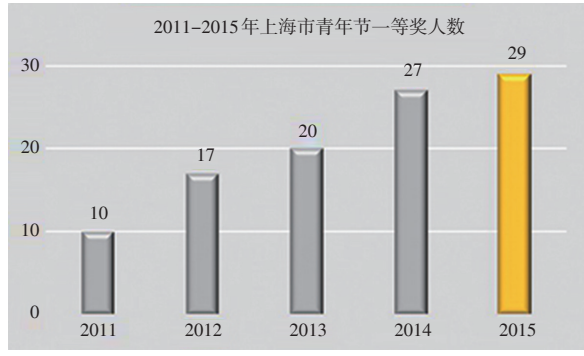


图3 CRAFT算法的图表文本定位结果



图4 基于CRNN的栅格文本块矢量化结果

图例、图表区3类结构进行构建，训练集包括130个样本，测试集包括30个样本。测试集的30个样本中，包含Office、Echart和泛在图表各10个。分别利用模板和机器学习方法进行测试，得到结果如表3所示。

表3 结构类别判断结果/%

类型及方法	Office		Echart		泛在图表	
	柱状图	饼状图	柱状图	饼状图	柱状图	饼状图
模板	95.2	95.4	95.5	96.1	89.2	90.7
机器学习	94.5	96.3	94.9	96.4	92.3	92.5

分析表3可以发现针对常见软件生成的规范图表，基于模板和机器学习方法进行图表结构解析得到的准确率基本相同，但是基于模板的方法所消耗的时间和空间资源远小于机器学习方法；针对随意性更强的泛在图表，机器学习方法准确性更高，要明显优于模板方法。在实际应用中可以根据具体需求选择不同的判断方法。

4.3 知识标签体系构建

根据图2，构建图表的两级知识标签体系，以图5为例，得到结果如表4所示。

分析表4，可以发现图5中的主要信息均被抽取出来，

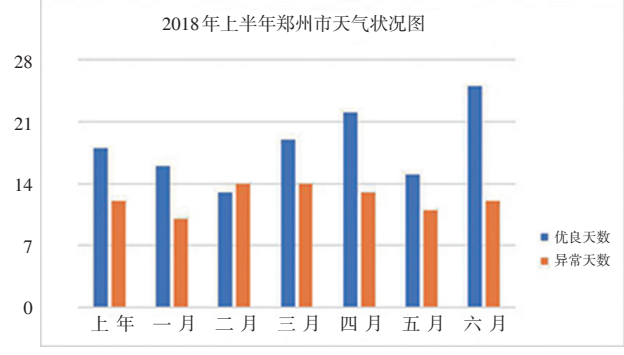


图5 原始栅格类图表

形成了能够表达图表核心内容的两级知识标签体系。

表4　知识标签构建体系

矢量化结果	结构类型	知识标签	标签等级
2018年上半年郑州市天气状况图	标题	时间：2018年、时间：上半年地点：郑州市、事件：天气状况	一级
优良时间段	图例	/	/
异常时间段	图例	/	/
上年	X轴	时间：上年	二级
一月	X轴	时间：一月	二级
二月	X轴	时间：二月	二级
三月	X轴	时间：三月	二级
四月	X轴	时间：四月	二级
五月	X轴	时间：五月	二级
六月	X轴	时间：六月	二级
0	Y轴	/	/
7	Y轴	/	/
14	Y轴	/	/
21	Y轴	/	/
28	Y轴	/	/

5　结　语

作为泛在信息的一种主要类型，栅格类图表包含了丰富的时空信息。本文设计并实现了一种栅格图表知识抽取方法，将文本信息作为理解图表数据的主要内容，结合栅格类图表的特点，采用CRAFT和CRNN方法分别进行文本定位和内容矢量化，解决了栅格图表中关键信息的矢量化问题。通过对图表结构进行解析，总结图表数据的结构特点，根据不同情况选择模板匹配和机器学习2种方法能够进行文本块的类别判断。利用自然语言处理技术，实现了知识的自动理解和抽取，建立栅格类图表的知识标签，有助于海量数据的快速检索和理解，可以为泛在信息特别是图像类数据的智能解译提供一种新的思路。

参考文献

[1] Qin X, Luo Y N, Tang, et al. Making Data Visualization More Efficient and Effective: A Wurvey[J]. The Vldb Journal,2020,29(1):93-117

[2] 孙恺.统计图表的视觉形式研究[D].南京:南京师范大学,2014

[3] 韩冰,王光霞,陈令羽,等.一种基于迁移学习的栅格型图表自动分类方法[J].测绘科学技术学报,2021,38(1):75-82

[4] Zhang L M, Wang G X , Chen L Y. Chart Master: An End-to-end Method to Recognize, Redraw, and Redesign Chart[J]. Discrete Dynamics in Nature and Society,2021(10): 1 155-1 159

[5] Yan P Z, Tan C L. Hough Technique for Bar Charts Detection and Recognition in Document Images[C] //International Conference on Image Processing. IEEE,2000:605-608

[6] Yan P Z, Tan C L.Learning-based Scientific Chart Recognition[C]//4th IAPR International Workshop on Graphics Recognition,2001

[7] 何君.栅格影像向矢量要素转换的关键技术研究[J].地理空间信息,2019,17(8):80-83

[8] 林晓蕊.自然场景下的文本检测算法研究[D].成都:电子科技大学,2021

[9] 姜毅.基于百度智能云的权源信息录入方法[J].地理空间信息,2022,20(5):145-149

[10] 王依宁.自然场景下的文本检测与识别研究[D].哈尔滨:哈尔滨理工大学,2021

[11] 王德青,吾守尔·斯拉木,许苗苗.场景文字识别技术研究综述[J].计算机工程与应用,2020,56(18):1-15

[12] 汪洪涛,李魁,潘昊,等.基于STN-CRNN的自然场景英文文本识别研究[J].武汉理工大学学报(信息与管理工程版),2021,43(6):574-579

[13] 曾诚,温超东,孙瑜敏,等.基于ALBERT-CRNN的弹幕文本情感分析[J].郑州大学学报(理学版),2021,53(3):1-8

[14] 冯海.基于深度学习的中文OCR算法与系统实现[D].北京:中国科学院大学,2019

[15] Yokokura N, Watanabe T. Layout-based Approach for Extracting Constructive Elements of Bar-Charts [C]//1997 Graphics Recognition, Algorithms and Systems (GREC'97), 1997:163-174

[16] 闾国年,袁林旺,俞肇元.信息地理学:地理三元世界的新支点[J].中国科学(地球科学),2022,52(2):374-376