**作者简介**

林鑫，男，1987年生，博士，副教授，研究方向：信息组织与检索，E-mail：xinlin@mail.ccnu.edu.cn。
余华娟，女，1997年生，硕士研究生，研究方向：信息组织与检索。
闫奕臻，女，2001年生，本科生，研究方向：信息组织与检索。

Research on Cell Semantic Relation Recognition in Complex Table Digitization

LIN Xin[1,2]　YU HuaJuan[1]　YAN YiZhen[1]
( 1. School of Information Management, Central China Normal University, Wuhan 430079, P. R. China;
2. Research Center for Data Governance and Intelligent Decision Making of Hubei Province, Wuhan 430079, P. R. China )

Abstract: Complex tables can describe data in a simple and intuitive way, and are widely used in all walks of life. However, complex tables have problems such as complex structures, diverse cell types, and different forms of table documents. They need to be data processed before they can be shared and reused. Therefore, this paper constructs a cell semantic relationship recognition model based on unsupervised learning to realize the digitization of complex tables. First, it uses machine vision technology to realize the segmentation of complex tables, and then recognizes the same template table based on the similarity of table structure and content. On this basis, heuristic rules are set to identify the semantic relationship of cells in combination with the value and location characteristics of header cells, illustrative cells and table body cells. Finally, the empirical research verifies that the method in this paper can achieve high accuracy and recall rate in complex table digitization, which is feasible.

Keywords: Complex Table; Semantic Relationship; Form Digitization; Machine Vision

（收稿日期：2022-08-27）

**书 讯**

# 《汉语主题词表》

《汉语主题词表》自1980年问世以后，经1991年进行自然科学版修订，在我国图书情报界发挥了应有作用，曾经获得国家科学技术进步二等奖。为适应网络环境下知识组织与数据处理的需要，由中国科学技术信息研究所主持，并联合全国图书情报界相关机构，自2009年开始进行重新编制工作，拟分为工程技术卷、自然科学卷、生命科学卷、社会科学卷四大部分逐步完成。目前工程技术卷和自然科学卷已出版。

《汉语主题词表（工程技术卷）》共收录优选词19.6万条，非优选词16.4万条，等同率0.84，在体系结构、词汇术语、词间关系等方面进行了改进创新。《汉语主题词表（自然科学卷）》共收录专业术语12.4万条，包含数学、物理学、化学、天文学、测绘学、地球物理学、大气科学、地质学、海洋学、自然地理学等学科领域，收词系统、完整，语义关系丰富、严谨，每条词汇都有相应的学科分类号表现其专业属性，并与同义英文术语对应。同时，建立《汉语主题词表》网络服务系统，提供术语查询、文本主题分析、知识树辅助构建等服务。《汉语主题词表》可用于汉语文本分词、主题标引、语义关联、学科分类、知识导航和数据挖掘，是文本信息处理及检索系统开发人员不可或缺的工具。

《汉语主题词表（工程技术卷）》已于2014年由科学技术文献出版社出版，分为13个分册，总定价3 880元。《汉语主题词表（自然科学卷）》已于2018年5月由科学技术文献出版社出版，分为5个分册，总定价1 247元。两卷均可分册购买。