基于知识融合的少数民族可视化原型系统的设计与实现

刘 影

江苏旅游职业学院 江苏扬州

【摘要】在快速发展的知识经济时代，知识融合在知识获取和展示上发挥着重要的作用，领域知识库的构建需要从不同的数据源中录入少数民族数据，,这时就会面临可能出现的数据冗余或冲突的问题,为了解决这个问题,本文设计并实现了面向少数民族知识的可视化原型系统，本文对海量异构少数民族文化资源融合进行深入的研究，首先构建少数民族文化资源知识融合模型，依托 Hadoop 平台，MapReduce 框架，开发了融入少数民族文化的可视化原型系统，系统实现了少数民族文化资源爬取，文本资源分词、词性标注、三元组抽取和知识融合等功能。

【关键词】知识融合；少数民族；原型系统

Design and Implementation of Minority Visual Prototype System Based on Knowledge Fusion

Ying Liu

Jiangsu Vocational College of Tourism, Yangzhou, Jiangsu Province

【**Abstract**】Rapid development in the era of knowledge economy, knowledge integration in knowledge acquisition and display, play an important role in the construction of domain knowledge base needs from different data sources in the input data of the ethnic minorities, and then face possible data redundancy and conflict problems, in order to solve this problem, this paper designed and implemented for minority knowledge visualization prototype system, This paper conducts an in-depth study on the integration of massive heterogeneous ethnic cultural resources. Firstly, a knowledge integration model of ethnic cultural resources is constructed. Based on Hadoop platform and MapReduce framework, a visual prototype system integrating ethnic culture is developed, which realizes the crawling of ethnic cultural resources. Word segmentation, part-of-speech tagging, triple extraction and knowledge fusion of text resources.

【**Keywords**】knowledge integration; ethnic minorities; prototype system

前言

在知识融合算法应用研究方面，房立芳提出了一种基于关键属性的知识融合方法，并将该方法应用到数据集成处理系统中，改善对异构数据进行自动合并处理的合理性[1]。王淑营等提出了基于知识图谱的高速列车知识融合方法[2]，郭恒等提出了多域数据融合的高速列车维修性设计知识图谱构建方法[3]。马永军等提出一种基于深度学习模型的卷积神经网络结构实现数据融合的算法 CNNMDA[4]。闫昱姝等提出一种基于本体的多源文本知识融合算法，进而得到粒度小、精度高且完备的文本知识。利用本体概念框架将文本知识结构化，并将概念框架进行融合[5]。沈艳霞等人提出了一种多目标人工蜂群算法，这种算法是基于进化知识融合的[6]。罗安根在融合知识图谱的基础上提出了结构化信息的深层语义匹配的实体链接算法[7]。

1 少数民族文化资源知识融合模型构建

为了更好的保护和传承少数民族文化，方便大众对少数民族文化节日风俗的了解，促进不同民族间的交流，对半结构化尤其是非结构化数据进行抽取并存储，构建少数民族文化资源知识融合模型。知识融合模型由三个部分构成，分别是底层数据、知识抽取和知识融合。底层数据抽取并以 RDF 三元组的形式存储，知识融合部分涉及到实例融合、域集融合、属性

作者简介：刘影（1995-）女，安徽亳州，硕士研究生，助教，知识融合、人工智能。

融合、概念融合。

2 少数民族知识可视化原型系统设计

2.1 少数民族知识可视化原型系统需求分析

随着时间的流逝，少数民族文化越来越不为人所知，信息技术的发展为少数民族文化的保护和传播也提供了解决方案，其中知识融合和可视化技术是重要的技术手段，少数民族文化资源的融合和可视化呈现为少数民族特色文化资源建设和传播提供指导，具有广泛的参考作用。因此，为促进少数民族文化传播，对百度百科、搜狗百科上网络少数民族文化资源进行整合，基于此，将相关词条进行可视化展现，构建少数民族文化资源可视化融合模型。资源整合后应用到原型系统上，将实体间的关系呈现出来，为用户获取结构化知识提供便利。

基于 Hadoop 平台、MapReduce 框架，利用 eclipse Mars.2 Release (4.5.2)开发，构建少数民族可视化原型系统，原型系统的少数民族文化资源库一部分来源于重点实验室现有资源，一部分是利用爬虫工具在互联网上爬取所得，可细分为饮食文化、服饰文化、交通、民俗文化、婚姻家庭等。

该原型系统界面比较简单，菜单栏上有爬虫、分词、词性标注、抽取三元组、知识融合等五个部分，点击相应按钮会有相关功能结果呈现。主界面的左边部分是资源库，中间用于现实相关功能的结果。

2.2 少数民族知识可视化原型系统功能设计

为实现将网络和现实世界少数民族文化资源进行整合，需要采集少数民族文化资源并进行预处理，形成少数民族特色语料库。语料库由结构化资源、半结构化资源和非结构化资源构成，要实现少数民族知识的可视化，重点要对这些资源进行处理，少数民族可视化原型系统功能模块由五个部分构成，涉及从百度百科、搜狗百科、互动百科、少数民族等网站上爬取到的少数民族相关数据，初步预处理后对数据分词、进行词性标注、命名实体识别抽取三元组，最后对来源不同的三元组数据进行融合。

3 少数民族知识可视化原型框架

依托于 Hadoop 平台，MapReduce 框架，互联网搭建原型系统框架结构图，框架的底层是从百度百科、搜狗百科等互联网渠道爬取，还有部分是少数民族重点实验室现有数据，中间层四个功能分别是分词、词性标注、三元组抽取和知识融合，依托 Hadoop 平台、MapReduce 框架实现，最上层是用户，利用用户可访问接口、进行模块化扩展，呈现给用户。

3.1 少数民族资源互联网内容爬取实现

少数民族特色文化资源广泛分布于互联网中，可利用现有搜索引擎，本次来源数据主要是少数民族人民政府网、百度百科、搜狗百科等，通过向百度、谷歌等搜索引擎提交少数民族资源关键词来搜索，获取少数民族文化资源列表，形成原始数据集，抓取到的字段有标题、来源、内容、发布时间等，对原始数据进行预处理，包括去重、信息规范化、无效数据剔除，形成少数民族文化资源库，为少数民族文化资源可视化呈现做准备。少数民族资源互联网内容爬取伪代码算法如下所示：

算法：爬虫算法

```
Step1.首先分析不同网站的标签规则
Step2.设置打开网页的浏览器 System.setProperty("webdriver.chrome.driver", chromePath)
Step3.获取所要爬取网页地址 url ,webDriver.get("url")
Step4.对网页 url 进行解析
Step5.输入少数民族关键字来进行搜索相关资源
    webDriver.findElement(By.cssSelector("input#searchText")).sendKeys("keyword")
Step6.获取网页内容
    for (String wordUrl : urlSet) {
        Document document = Jsoup.connect(wordUrl).get()
    }
```

3.2 少数民族文本资源 HMM 分词实现

从互联网上爬取的数据部分是结构化数据，收集到的数据多为非结构文本数据，无法直接使用，要进行初步预处理后分词，利用自定义词典和分词工具对

网络上收集到的少数民族数据进行分词处理，分词的结果关系到后续三元组抽取的准确度，本文利用隐马尔可夫模型对非结构化文本数据进行分词可以达到较好的分词效果，部分佤族文化文本资源分词结果如图 1 所示。

3.3 少数民族文本资源词性标注实现

本文在隐马尔可夫模型分词的基础上进行词性标注，词性标注使用的是北大词性标注集，用隐马尔可夫模型对词性进行标注，词性标注的准确性对下一阶段三元组提取也有影响，改词性标注效果较好，部分词性标注结果如下图所示。

3.4 少数民族文本资源三元组抽取实现

对词性标注后的数据采用无监督学习的方式，结合上下文特征信息进行命名实体识别并提取关系，。数据关系的建立一般有两个方面，一是描述知识主题，二是通过三元组关系 1 得到其他内容，比如“中国少数民族人口约 1.2 亿”可抽取出如下三元组（中国少数人民，人口，1.2 亿）。将抽取到的实体和关系进行链接，从而得到三元组，提取到的三元组存储在数据库中，此外，在进行相似度计算时引入了密度聚类，提取到的部分三元组如图 3 所示

3.5 少数民族资源知识融合实现

由于来源不同，这些少数民族资源存在语法、语义上的异构，采取一定的融合规则消除这些语法语义上的异构，并将这些资源融合到实验室已有领域知识库，使得知识库更为庞大， 知识库的充裕也为后续各项研究提供高质量数据，比如知识推理、知识推荐等等，在少数民族知识融合平台上清楚直观的展示了少数民族实体之间的关系，如图 4 所示截取了部分可视化结果，呈现出来的结果可以更好的传承少数民族文化，为教育行业在少数民族领域资源上的教学也能起到一定的指导。

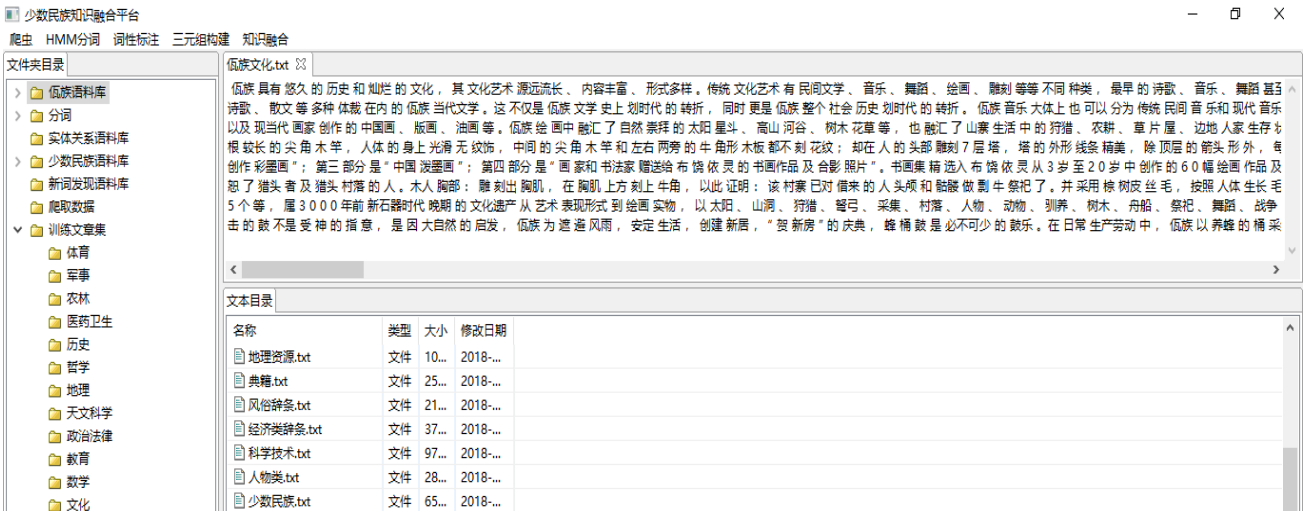

图 1 佤族文化 HMM 分词结果

图 2 佤族文化 HMM 词性标注结果

图 3 提取三元组部分可视化结果

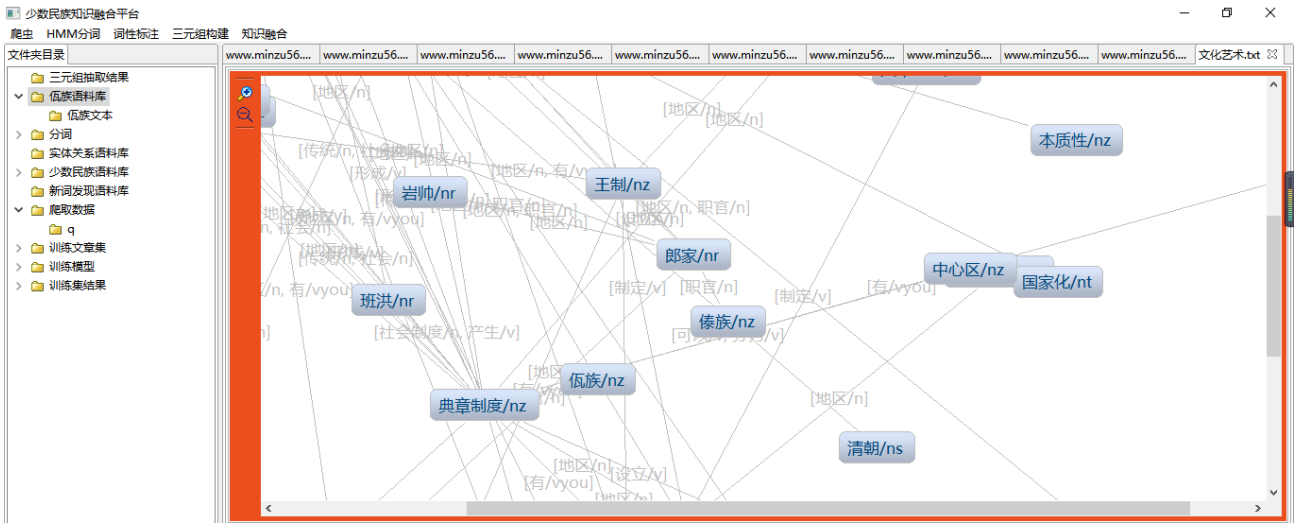

图 4 少数民族资源融合可视化呈现

总结

本文作者在充分研读文献，对知识融合有了一定认识后，以少数民族文化资源为例，构建适合少数民族文化资源的知识融合模型，设计并实现少数民族可视化原型系统，在资源获取上，一部分从搜狗百科、百度百科、少数民族网站等互联网渠道获得，还有部分是少数民族重点实验室现有资源，这些资源当中有结构化数据、半结构化数据和非结构化数据，重点在于对非结构化文本数据的处理，包含预处理、分词、词性标注、三元组数据抽取等环节，并将抽取到的三元组数据进行融合，消除来源不同造成的异构性，本次实验也侧面验证了知识融合技术在少数民族资源的管理上的有效应用，知识融合可视化结果一方面给了学生更为直观的体验，另一方面，外界可以通过我们的平台了解到少数民族文化，使得少数民族文化得以传承。

参考文献

[1] 房立芳. 基于本体的异构数据集成与融合方法研究[D].中国科学技术大学, 模式识别与智能系统， 2010, 5.

[2] 王淑营,李雪,黎荣,张海柱.基于知识图谱的高速列车知识融合方法[J/OL].西南交通大学学报:1-11[2022-09-14].http://kns.cnki.net/kcms/detail/51.1277.U.20220711.1502.002.html

[3] 郭恒,黎荣,张海柱,魏永杰,戴钺滨.多域融合的高速列车维修性设计知识图谱构建[J/OL].中国机械工程:1-10[2022-09-14].http://kns.cnki.net/kcms/detail/42.1294.TH.20220621.1343.002.html

[4] 马永军，薛永浩，刘洋，李亚军.一种基于深度学习模型的数据融合处理算法[J].天津科技大学学报，2017,32(04):71-74+78.

[5] 闫昱姝，雷玉霞.多源文本知识融合算法分析[J].软件导刊，

2018,17(05):62-64.

[6] 沈艳霞，陈杰，吴定会.一种基于进化知识融合的多目标人工蜂群算法[J].控制与决策，2017,32(12):2176-2182.

[7] 罗安根. 融合知识图谱的实体链接的算法研究[D].北京邮电大学，2018.

收稿日期：2022 年 10 月 23 日
出刊日期：2022 年 11 月 27 日
引用本文：刘影，基于知识融合的少数民族可视化原型系统的设计与实现[J],科学发展研究, 2022, 2(6) : 24-28
DOI: 10.12208/j.sdr.20220211

检索信息：RCCSE 权威核心学术期刊数据库、中国知网（CNKI Scholar）、万方数据 （WANFANG DATA）、Google Scholar 等数据库收录期刊