软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(9):2816−2836 [doi: 10.13328/j.cnki.jos.006010]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

碎片化家谱数据的融合技术*

吴信东[1,2,3,4], 李 娇[1,2,3], 周 鹏[5], 卜晨阳[1,2,3]

[1](大数据知识工程教育部重点实验室(合肥工业大学),安徽 合肥 230009)

[2](合肥工业大学 计算机与信息学院,安徽 合肥 230601)

[3](合肥工业大学 大知识科学研究院,安徽 合肥 230009)

[4](明略科技集团,北京 100102)

[5](安徽大学 计算机科学与技术学院,安徽 合肥 230601)

通讯作者: 吴信东, E-mail: xwu@hfut.edu.cn

摘 要: 家谱数据是典型的碎片化数据,具有海量、多源、异构、自治的特点.通过数据融合技术将互联网中零散分布的家谱数据融合成一个全面、准确的家谱数据库,有利于针对家谱数据进行知识挖掘和推理,从而为用户提供姓氏起源、姓氏变迁和姓氏间关联等隐含信息.在大数据知识工程 BigKE 模型的基础上,提出了一个结合 HAO 智能模型的碎片化数据融合框架 FDF-HAO(fragmented data fusion with human intelligence, artificial intelligence and organizational intelligence),阐述了架构中每层的作用、关键技术和需要解决的问题,并以家谱数据为例,验证了该数据融合框架的有效性.最后,对碎片化数据融合的前景进行展望.

关键词: 碎片化数据;数据融合;家谱数据;多源异构;HAO 智能模型

中图法分类号: TP311

中文引用格式: 吴信东,李娇,周鹏,卜晨阳.碎片化家谱数据的融合技术.软件学报,2021,32(9):2816−2836. http://www.jos.org.cn/1000-9825/6010.htm

英文引用格式: Wu XD, Li J, Zhou P, Bu CY. Fusion technique for fragmented genealogy data. Ruan Jian Xue Bao/Journal of Software, 2021,32(9):2816−2836 (in Chinese). http://www.jos.org.cn/1000-9825/6010.htm

Fusion Technique for Fragmented Genealogy Data

WU Xin-Dong[1,2,3,4], LI Jiao[1,2,3], ZHOU Peng[5], BU Chen-Yang[1,2,3]

[1](Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Hefei 230009, China)

[2](School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China)

[3](Research Institute of Big Knowledge, Hefei University of Technology, Hefei 230009, China)

[4](Mininglamp Technology, Beijing 100102, China)

[5](School of Computer Science and Technology, Anhui University, Hefei 230601, China)

Abstract: Genealogy data is a typical example for data fragmentation with massive, multiple, heterogeneous, and autonomous sources. Merging scattered genealogy data on the Internet into a comprehensive and accurate genealogy database through data fusion technologies, can be beneficial to knowledge mining and reasoning from genealogy data, and can provide users with implicit information such as surname origins, surname changes, and surname associations. Based on BigKE, a big data knowledge engineering model for big knowledge, this study proposes an FDF-HAO framework (fragmented data fusion with human intelligence, artificial intelligence, and organizational intelligence), describes the functionalities, key technologies, and problems to be solved of each layer in the framework, and

* 基金项目: 国家重点研发计划(2016YFB1000901); 国家自然科学基金(91746209); 教育部创新团队项目(IRT17R3)

Foundation item: National Key Researh and Development Program of China (2016YFB1000901); National Natural Science Foundation of China (91746209); Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education (IRT17R3)

收稿时间: 2019-06-22; 修改时间: 2019-09-20, 2019-11-19; 采用时间: 2020-01-02; jos 在线出版时间: 2020-04-21

© 中国科学院软件研究所 http://www.jos.org.cn

verifies the validity of the data fusion framework by using genealogy data as an example. Finally, the challenges and opportunities of fragmented data fusion are also discussed.

Key words: fragmented data; data fusion; genealogy data; multiple heterogeneous sources; HAO intelligence model

随着互联网、云计算等技术的迅猛发展,网络空间中的数据以不可预计的速度增长,信息社会进入了大数据时代[1].除了大数据的“5V”特征外,数据碎片化已成为大数据处理中不可忽视的问题.如何有效地融合这些碎片化数据,从多源异构的碎片化数据中获取整个大数据集合的全局数据特征,继而从海量碎片化数据中提取出有价值的信息,已成为学术界的研究重点和热点.

在大数据环境下,由于数据的多源异构性,来自不同数据源的碎片化数据往往具有不同的数据结构和形式.碎片化数据融合的首要挑战就是:如何从这些多源异构的数据中抽取出真正有价值的信息,使用合适的处理机制对碎片化数据进行提取和分析.另外,碎片化数据融合并不只是简单地将数据“拼凑”在一起,而是通过分析碎片化数据之间的内在联系,得到新的、完整的数据.此外,经过融合后的数据通常具有复杂的语义关系,为此,我们需要寻找一种标准化的数据表示方式对其进行存储与表示.故而,碎片化数据融合极具挑战性[2].本文以碎片化家谱数据融合为例,详细阐述了数据融合过程中存在的问题和解决方案.

家谱资料的数字化使得网络中的家谱数据资源不断增多,是典型的碎片化数据.家谱与正史、地方志并列为我国历史研究的三大基石之一[3],它不仅记录族人最基本的世系状况,还记录族人的姓氏源流、族规家训等内容,涉及历史、人口、经济等多门学科[4],具有重要的学术价值和史料价值[5].从这些多源分散的家谱数据开始,使用大数据技术及手段对其进行碎片化重组及融合,有利于对家谱中历史、经济等复杂信息的研究与分析,深度揭示家谱大数据中尚未发现的或难以处理的问题,增强人民群众对寻根问祖的需求,增加海外华侨对祖国的认同感,实现大数据技术与人文社会科学研究的“双赢”[6,7].

现有的专门针对家谱数据的研究较少,且大多数都侧重于对家谱数据的存储研究[8−11],缺少对家谱数据融合及知识挖掘与推理方面的研究.其主要原因在于:一方面,大量的家谱数据属于传统数据资源,在大数据时代,这些数据必须与其他数据进行有效整合才能更易于被用户使用,比如电子化、网络化等,因此往往需要面对着非常巨大的处理成本和转换成本[12];另一方面,家谱大数据真正难以对付的挑战来自于数据类型多样、数据多源异构的特征和数据的不确定性[13].

针对碎片化大数据的分析和应用,吴信东等人提出了一种大数据知识工程模型 BigKE[2],该模型采用一种三层次的知识建模方法:首先,对多源异构数据中的碎片化知识进行建模;然后,使用知识图谱对碎片化知识进行非线性融合;最后,以用户需求为导向,提供具有个性化和实时使用价值的知识服务[14].BigKE 考虑到大数据的异构和自治特征,对大数据挖掘形成的知识图谱提出了个性化服务的导航,更有利于和具体的应用实例结合.

在大数据知识工程 BigKE 的技术框架下,吴信东团队推出了面向所有华人姓氏的家谱系统——华谱系统(华谱系统网址:http://zhonghuapu.com).华谱系统通过对家谱数据进行碎片化知识融合,旨在为用户提供姓氏的起源、姓氏的变迁、姓氏间关联等信息.目前,华谱系统中人物数量已超过 1587 万,姓氏数目已超过 720,数据源超过 500 个.系统数据量还在与日俱增.

在华谱系统中碎片化家谱数据融合过程的基础上,结合 HAO 模型[15],本文提出一个针对碎片化数据的融合框架 FDF-HAO(fragmented data fusion framework with human intelligence, artificial intelligence and organizational intelligence).该框架从碎片化数据开始,通过 HI(人类智能)、AI(人工智能)和 OI(组织智能)三者的交互和协同,实现多源异构的碎片化数据的融合,最后形成一个由实体和各种关系链接而成的网状知识库,即家谱人物知识图谱.人类智能指领域专家们所提供的专家知识.人工智能指机器完成的智能工作,如自然语言处理技术、机器学习算法等.组织智能涵盖了一个组织的全部知识能力[15],在本文中指家谱领域内的领域规范或行业标准.

本文提出的 FDF-HAO 融合框架具有以下几个特点.

(1) 通过 HI、AI 和 OI 三者的交互和协作,为大规模、异构、多源的碎片化数据融合提供智能支持;

(2) 通过对家谱数据语义和语法特征的观察与分析,基于 HI 和 OI 提供的专家知识和数据标准,在框架内

© 中国科学院软件研究所 http://www.jos.org.cn

提出了新的针对家谱数据的数据抽取方法;

(3) 在 HI 的协作下,采用了一个面向家谱人物的无监督实体对齐算法,能够准确高效地从海量家谱数据中识别出相同人物;

(4) 结合 OI 制定一套多源异构家谱人物属性的冲突解决机制,能够简单高效地从多个冲突值中选择真值;

(5) 根据家谱数据的特点,在 HI 和 OI 的智能支持下,设计了一个面向家谱领域的属性融合算法,能够从多源、碎片化的数据中凝练出实体的统一的、准确的、有用的描述.

本文第 1 节对相关工作进行阐述.第 2 节对本文提出的碎片化数据融合框架 FDF-HAO 进行详细描述.第 3 节举例验证本文提出框架的有效性,并对框架中采用的关键技术与同类技术进行性能对比分析.第 4 节对碎片化数据融合过程中仍存在的挑战进行阐述,并对其应用前景进行展望.第 5 节对全文做总结.

1 相关工作

1.1 数据抽取

数据抽取的主要任务是从大量结构化或非结构化的数据中准确、快速地抽取实体、关系以及实体属性等结构化信息[16].根据所需抽取信息的种类,数据抽取可分为 3 个模块:实体抽取、关系抽取、属性抽取.

1.1.1 实体抽取

实体抽取,也称为命名实体识别(name entity recognition,简称 NER),指识别文本中具有特定意义的实体,主要包括人名、组织机构名、地名等[17].早期对实体抽取的方法主要是基于规则的方法,即人工构建规则,再从文本中寻找匹配这些规则的字符串.例如,Rau[18]采用启发式算法与人工编写规则相结合的方法,从财经新闻中自动抽取公司名称,实现了不错的效果.但是,人工制定这些规则需要耗费大量时间和精力,而且规则对领域知识的依赖性较高,当领域差别很大时,制定的规则无法重用,可扩展性较差.

后来,随着机器学习在 NLP 领域的兴起,人们开始尝试使用机器学习方法解决实体抽取问题.机器学习方法是指从样本数据集合中统计出相关特征和参数,以此建立识别模型[19].Lai 等人[20]结合统计原理和条件随机场模型,对专利中的化学名称进行识别,在不同数据集上的 F 值均高于 70%.Hwang 等人[21]通过分析学术期刊摘要中同时出现在特定词语周围的特定词语之间的搭配关系,建立了一个实体识别模型.Akkasi 等人[22]利用条件随机场模型为命名实体识别创建各种基线分类器,然后结合粒子群优化算法和贝叶斯方法对分类器进行选择和有效组合.实验表明,该方法选择的分类器集成性能优于单一的最优分类器,也优于采用其他常用选择/组合策略形成的两个语料库的集成性能.

近年来,基于神经网络的深度学习技术成为机器学习领域新的热潮,一些学者开始将深度学习技术应用在 NER 问题上,以求进一步提高 NER 的效果[23].Peng 等人[24]借鉴 LSTM 在自动分词上得到较好的结果,提出一种 LSTM 与 CRF 相结合的模型.结果显示,该方法的 F 值比之前的方法高了将近 5%.Qiu 等人[25]提出了一种基于条件随机域的残差扩张卷积神经网络(RD-CNN-CRF),使模型在计算上具有异步性,大大加快了训练周期,实现了中文临床命名实体识别.

1.1.2 关系抽取

实体和实体之间存在着语义关系,当两个实体出现在同一个句子或同一段落里时,上下文环境就决定了两个实体间的语义关系,通过关系将实体联系起来,才能够形成网状的知识结构[26].

经典的实体关系抽取方法主要分为有监督、半监督、弱监督和无监督这 4 类.有监督的实体关系抽取主要分为基于特征和基于核函数的方法[27].甘丽新等人[28]通过将 2 个实体各自的依存句法关系进行组合,获取依存句法关系组合特征,利用依存句法分析和词性标注选择最近句法依赖动词特征,使用支持向量机实现了实体关系的抽取.但是有监督方法需要大量的标注数据,浪费时间和精力.因此,人们继而提出了基于半监督、弱监督和无监督的关系抽取方法.陈立玮等人[29]针对弱监督学习中标注数据不完全可靠的情况,提出基于 booststrapping 思想的协同训练方法来对弱监督关系抽取模型进行强化,并且对预测关系时的协同策略进行了详细分析.

© 中国科学院软件研究所 http://www.jos.org.cn

Hasegawa 等人[30]提出了一个无监督的关系抽取方法,其核心思想是,根据命名实体之间的上下文词的相似性对命名实体进行聚类.

随着近年来深度学习的崛起,学者们逐渐将深度学习应用到关系抽取任务中,主要基础方法有 CNN,RNN,LSTM 等.Leng 等人[31]提出了一种改进的叠加去噪自动编码器的深度学习模型,用于提取不同命名实体之间的关系.Ji 等人[32]充分利用知识库的有监督信息,在 PCNN 和注意力机制的基础上实现了关系的抽取.

1.1.3 属性抽取

属性抽取是指在无序信息文本中将关注实体的属性特征进行集中的提取,可以观察和总结出此实体关于此属性的价值信息.目前,针对人物属性的抽取研究逐渐增多,并通过不断改进研究方法,抽取工作已取得不错的成果.

属性抽取当前的研究热点是对半结构化数据的信息抽取.然而,有大量的实体属性信息隐藏在非结构化数据中,如何从海量非结构化数据中抽取实体属性是值得关注的问题.对于非结构化数据的属性抽取,目前有两种解决方案:一种是通过自动抽取半结构化数据中的实体属性,生成训练语料库,用于实体属性标注模型,然后将其应用在非结构化数据的实体属性抽取中[33];另一种方案是采用数据挖掘的方法直接从文本中挖掘实体属性与属性值之间的关系模式,实现对非结构化数据的属性抽取.实际上,实体属性值附近一般都存在一些用于限制和界定该属性值含义的关键词,因此可以利用这些关键词来定位实体属性值,进行属性抽取[34].

1.2 数据融合

数据融合主要是指整合表示同一个现实世界对象的多个数据源和知识描述,形成统一的、准确的、有用的描述的过程[35],其过程可分为实体对齐、冲突消解、属性融合.

1.2.1 实体对齐

在真实语言环境中,经常会遇到同一实体指称项对应着多个不同实体的情况.例如,“李娜”这个姓名可以对应于作为歌手的李娜,也可以对应于作为网球运动员的李娜.另一种情况同样存在,即不同实体指称项对应于同一实体.例如,“孔子”“孔丘”“孔仲尼”等姓名都代表同一个人物“孔子”.因此,实体对齐问题应运而生.实体对齐[36]是判断相同或不同数据集中的两个实体是否指向真实世界同一对象的过程.

最初,实体对齐方法主要基于文本相似性函数对实体进行特征匹配.但这种方法仅考虑实体的上下文语义信息,忽略了实体之间存在的“共现”关系.1969 年,Fellegi 和 Sunter[37]提出一种基于传统概率模型的实体对齐方法,通过将基于属性相似性评分的实体匹配问题转化为分类问题,建立了这个问题的概率模型.这种模型是实体对齐领域的重要方法,迄今为止,仍然有大量的实体对齐方面的工作建立在这种方法之上.

随着机器学习的兴起,很多机器学习方法也逐渐应用到实体对齐领域,并取得了巨大的进展.机器学习方法主要将实体对齐问题看作是二元分类问题,根据是否使用标注数据,可以分为有监督学习和无监督学习两类.Chen 等人[38]结合两种监督学习的方法,将多种基础实体对齐系统和上下文特征映射起来,形成统一的聚类决策模型.

但是在大规模数据的情况下,实体对齐过程中的训练数据是较难获取的,往往需要耗费大量的时间和精力去对数据进行标注.Guan 等人[39]提出了一种自学习的实体对齐方法,充分利用了实体属性中包含的语义信息,迭代查找语义对齐的实体对.

在实体对齐过程中,候选实体对的生成对结果的正确性起着十分重要的作用.通常来说,为了发现所有的候选实体对,需要将一个知识库中的所有实体与另一个知识库中的所有实体进行比较,这将导致算法的计算复杂度随着数据规模二次增长.

1.2.2 冲突消解

检测出碎片化数据中的相同实体后,我们需要对相同的实体的信息进行融合,将同一实体的所有属性信息合并成一条完整的实体描述信息.但在融合过程中,不同数据源中同一实体的信息可能会因为错误、丢失、数据过期等原因出现冲突的情况[40].因此,我们需要在各数据源提供的值中,选择与真实世界相一致的值,即数据的真值.这个过程我们称之为数据冲突消解[41].

© 中国科学院软件研究所 http://www.jos.org.cn

数据冲突消解方法层出不穷,现有的数据冲突消解方法大都通过关系扩展的方式实现,并定义了若干冲突消解策略和冲突消解函数[36].但这类方法在适应性和准确性方面分别存在着一定的不足,难以适应大规模数据的冲突消解任务.另外,还有一些冲突消解策略是从多个冲突值中选择真值.Yin 等人[42]基于一些启发式规则提出了一个解决数据冲突问题迭代计算的准则,设计出了 TruthFinder 算法.但这种方法仅考虑数据源和数据值之间的关系,没有考虑到数据源之间的依赖关系,这在一定程度上会对最终结果造成不利影响.Lyu 等人[43]提出一种无监督的冲突消解模型,利用数据源-数据源和数据源-数据值之间的关系构造一个异构网络,并将其嵌入至一个低维空间中,自动地发现数据的真值.

另外,现有冲突消解方法主要是对所有属性的数据冲突问题采取同等对待的方式.但这些方法并没有考虑不同属性的冲突程度可能不同,也没有考虑不同属性间的相互影响,这在一定程度上也会导致冲突消解的准确率降低.

1.2.3 属性融合

在对不同数据源的实体信息进行融合时,我们发现这些数据源的信息中,存在名称不同含义相同或名称相同含义不同的属性.因此,我们需要对实体的属性进行判断,把名称不同但含义相同的属性进行合并,或者把名称相同却含义不同的属性进行拆分,从而获得更准确、更丰富的属性信息.这个过程我们称之为属性融合.

现有的属性融合的方法包括基于相似距离计算的方法、基于统计语言模型的方法和基于词典匹配的方法等,主要通过建立模型等方式对实体属性进行相似度计算.2014 年,Jakub 等人[44]通过比较数据集的特征和聚集属性信息来计算两个属性的最小距离,再通过 KNN 算法实现属性对齐.该方法能够在没有丢失重要信息的前提下实现属性对齐,能够预测个人属性和对齐属性的距离.

2 碎片化数据融合框架

本节先阐述碎片化数据融合框架的主要结构,然后以华谱系统中碎片化家谱数据融合为例,详细介绍碎片化数据融合框架中家谱数据在每层的处理过程和解决方案,以验证本文提出的碎片化数据融合框架的有效性.

2.1 概述

本文提出了一个碎片化数据融合框架 FDF-HAO,通过 HI、AI 和 OI 三者的交互和协同,为多源异构碎片化数据的融合过程提供智能支持.该框架在 HAO 智能的技术背景下,以碎片化数据为起点,通过数据获取、数据抽取、数据规范和数据融合这 4 个模块的处理,最后形成一个由实体和各种关系链接而成的网状知识库,即知识图谱.框架图如图 1 所示.碎片化数据融合过程可分为以下 4 个部分.

(1) 数据获取层.数据获取层的主要功能是使用爬虫技术(AI)从互联网中获取不同来源和形式的数据.不同数据源所涉及的数据类型有很多种,如文本文件、表格文件、网页数据等.因此,数据获取层中获取的碎片化数据具有多源、异构的特点;

(2) 数据抽取层.为了实现数据的统一存储,数据抽取层从底层多源异构的碎片化数据中提取出有价值的信息,其关键在于结合 HI 和 OI,采用自然语言处理技术(AI),通过对自然语言的词法、句法的分析,实现实体、关系、属性的抽取;

(3) 数据规范层.数据规范层的主要功能是:在 OI 提供的数据规范标准下,将从数据抽取层中提取的信息标准化、规范化,以避免因语义异构性引起的数据冲突等问题;

(4) 数据融合层.数据融合层是碎片化数据融合框架的核心,在 HI 和 OI 的智能支持下,使用机器学习技术(AI)将数据规范层中标准化后的数据进行实体对齐、冲突解决和属性融合,形成以关系为有向边的数据网络,为后期的高级知识应用和服务提供数据基础.

© 中国科学院软件研究所 http://www.jos.org.cn

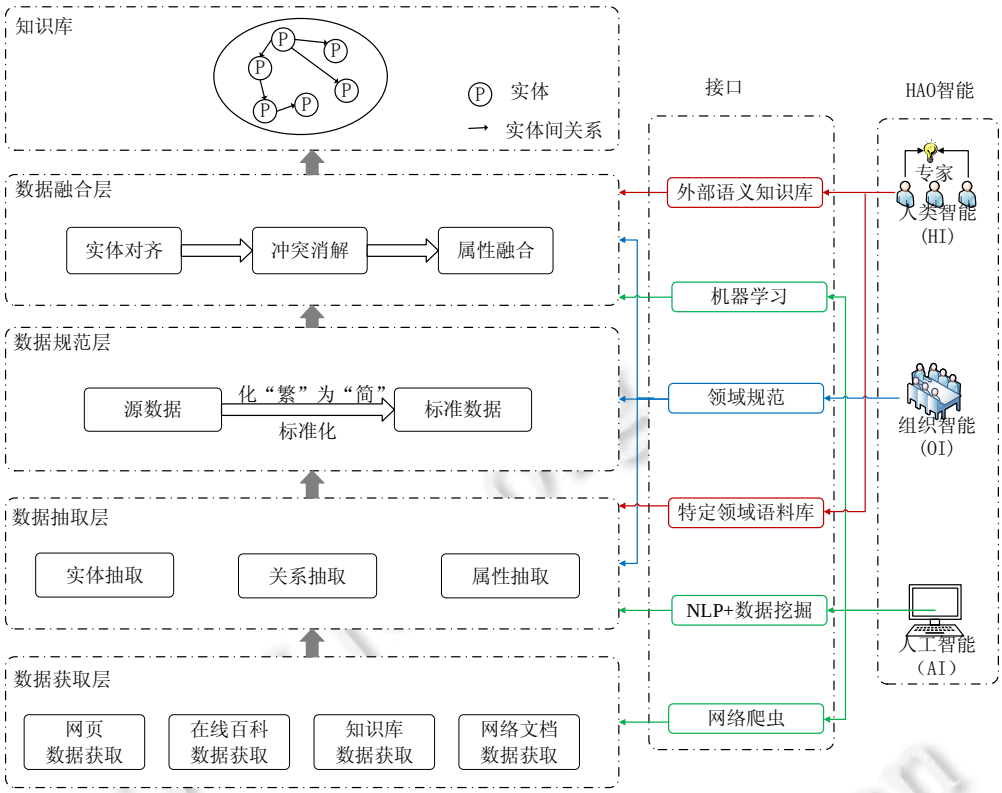

Fig.1 FDF-HAO framework

图 1 FDF-HAO 框架结构图

2.2 数据获取层

首先,在数据获取层中,主要是从互联网中采集多源、异构、碎片化的家谱数据.本文使用爬虫技术(AI),在利用 webcollector(https://github.com/CrawlScript/WebCollector)爬虫框架的基础上,实现对网络中家谱数据的获取.家谱数据源主要为上海图书馆、百度文库、豆丁网等网站.数据获取层主要包括以下 3 个过程.

(1) 确定网页地址(URL)

通常情况下,传入爬虫的是网站的主页,即用户最先浏览的主页,而后根据不同的需求在主页相关的网页之间进行切换.确定网页地址有两种方式:a) 通过获取网站主页中的超链接,确定需要爬取数据的网页地址;b) 寻找网站中各网页地址的规律,通过代码自动生成网页地址.

(2) 解析网页文件

观察爬取后的数据及其格式,通过程序对数据进行解析,过滤无用数据,提取所需要的信息.

(3) 存储数据

根据爬取数据的格式,为其选择合适的存储方式,一般可以存储为 TXT,WORD 等格式的文件.

2.3 数据抽取层

本节首先对家谱数据的文本特点进行总结与分析,然后介绍本文在家谱数据特点的基础上所设计的家谱数据抽取方法.

2.3.1 家谱数据文本特点

家谱数据的形式主要有表格、文本、网页这 3 种,其中,文本是家谱中较常见的数据类型.而与传统的文本相比,家谱数据中的文本不管是结构还是语言,都具有其独特性.

© 中国科学院软件研究所 http://www.jos.org.cn

(1) 碎片化特征

随着家谱文献的数字化,互联网中的家谱数据逐渐增多,海量的家谱数据堪称人文社科领域的大数据.由于大数据的特征——海量、多源、异构、碎片化,针对家谱数据的信息抽取任务存在着巨大的挑战.

(2) 结构特点

家谱数据中的文本通常以“世系图+人物描述”形式记录人物信息:“世系图”为树形结构,记载男性人物亲属关系,该部分可用于人物姓名及关系抽取;“人物描述”为一段记叙性文字,记载该人物属性信息及其人物关系,该部分是人物信息抽取的关键部分.家谱数据的这两部分结构中的内容可用于交叉验证人物姓名及关系抽取的正确性.

(3) 语言特点

与传统的文本数据相比,家谱数据使用的语言有其独有的特点:a) 包含大量繁体字及生僻字;b) 经常使用一些偏文言文的词汇和语法,如“妣”“适”等;c) 同一份家谱中,人物的描述性信息通常具有相同的模式.

家谱中的人物信息隐藏在文本数据中,计算机很难自动对其进行处理.因此,自然语言处理、数据抽取等技术在家谱数据的挖掘和推理中将发挥重要的作用.同时,家谱数据的碎片化特征、结构特点和语言特点给这些技术在家谱领域内的应用带来新的机遇和挑战,其挑战主要在于家谱的用词语法和行文风格与开放领域文本或其他领域文本截然不同.因此,已有的自然语言处理工具如分词、依存句法分析等在家谱数据上都有可能失效.同时,通过充分利用家谱数据的特点,如家谱数据中较强的模式化表达习惯、语言精练准确无歧义等,可以使得处理难度大为降低,并利用结构之间的联系进行信息归纳和推理.

2.3.2 家谱数据抽取方法

家谱数据多是以自然语言描述的非结构化文本,并且包含大量领域内特有词汇及语法,与机器语言之间存在巨大鸿沟,导致用计算机直接处理和分析家谱数据的效率较低,也影响了分析结果的质量.通过对家谱数据的观察,我们发现:家谱数据在行文和布局上具有一定的结构和规律,但不同家谱的行文方式和布局结构又不尽相同.对于具有一定结构的家谱数据来说,采用基于启发式规则的方法进行数据抽取最为简单高效.但面对大量不同种类不同结构的家谱数据,如果对每一份家谱均编写一套相应的规则,将耗费大量的人力物力,在实际应用中,实用性非常低,不具有通用性.因此,本文提出一种基于HAO模型的通用家谱信息抽取方法,在OI提供的家谱领域规范的标准下,利用HI和AI的协同作用,对家谱数据进行信息抽取.由上一小节中对家谱数据的分析可知,家谱数据中的文本通常以“世系图+人物描述”形式记录人物信息.因此,我们分别对“世系图”及“人物描述”中的信息进行抽取,在经过专家(HI)确认后的“世系图”数据抽取结果的协助下,对“人物描述”中所包含的人物属性信息和人物间关系进行抽取.

(1) “世系图”数据抽取

“世系图”以树形结构记载家谱中男性人物的亲属关系.我们可以通过计算机读取家谱中的“世系图”部分,从中抽取家谱人物姓名.但是计算机无法自动区分家谱中的“世系图”和“人物描述”,因此,本文通过与 HI 的交互,为计算机提供少量信息,确定家谱中“世系图”所处范围.之后,计算机自动抽取“世系图”中的人物姓名.另外,我们将抽取出的家谱人物姓名作为有监督数据,构建家谱领域人名词典,以便提高 NLP 工具对家谱进行分析处理的精确性.

(2) “人物描述”数据抽取

“人物描述”中蕴含着丰富的人物信息,包括人物姓名、属性及关系.通过对家谱数据的观察,我们发现:在“人物描述”中,每个家谱人物的描述信息独立成一段或多段;并且每份家谱以固定的模式化语句对人物属性信息和人物间关系进行介绍.

首先,HI即领域专家们通过对家谱数据的观察与分析,根据家谱文本描述的前后语义关系,结合对语境的理解,对家谱数据的语言模式进行总结,构建家谱领域全局知识库,为计算机提供家谱领域外部语义知识.由于家谱语法结构复杂,信息不一,表 1 为简化后的家谱语言模式.其中,N 表示人物姓名,FN 表示父亲姓名, PN 表示配偶姓名,SN 表示儿子姓名,XX 为属性信息.

© 中国科学院软件研究所 http://www.jos.org.cn

Table 1 Language schema
表 1 语言模式

	语言模式
模式 1	N,FN 之子,字 XX,号 XX,生于 XX,卒于 XX,…,配 PN,一子:SN.
模式 2	FN 之子,N,字 XX,号 XX,生于 XX,卒于 XX,…,配 PN,一子:SN.
模式 3	N,PN,PN,…一子:SN
模式 4	N,字 XX,号 XX,生于 XX,卒于 XX,…,…配 PN,生于 XX,卒于 XX,…,…一子:SN

由于不同家谱具有不同的语言模式,本文使用 HanLP 汉语语言处理工具[45]提取家谱数据的浅层词法特征,对家谱数据进行分词、命名实体识别和词性标注.但由于家谱的用词语法和行文风格与开放领域文本不同,为了保证 Hanlp 分析结果的准确性,我们需要对家谱中特殊的用词进行总结,构建家谱领域词典,为 Hanlp 提供家谱领域语义支持.

从表 1 中可以看出:人物之间关系和属性的描述具有一定的规律性,不同种类信息附近通常有固定的、具有语义信息的关键词.因此,在对家谱进行分析处理后,根据分析后的词语词性及词语附近关键词,在全局知识库提供的语义知识的协助下,判断当前待处理家谱所对应的语言模式,对家谱进行初步的信息抽取,并自动构建适用于当前家谱的局部规则库.例如,在关键词“配”“妣”“娶”后的词性为“人物姓名”的词语一般为人物配偶姓名,关键词“字”后的词性为“名词”的词语一般为人物属性“字”的属性值.

之后,我们根据上一步中构建的局部规则库,对少部分家谱数据进行抽取,将结果反馈给用户:若用户确定当前抽取结果正确,则采用现有局部规则库;若用户对抽取结果不满意,则用户对数据进行标注,并将标注信息返回至计算机.计算机根据用户标注信息与原抽取结果的差异,对局部规则库中的规则进行修正.

另外,由于家谱领域的特殊性,家谱文本中通常蕴含着一些隐性的人物信息.为了确保数据抽取的全面性和准确性,我们根据 OI 提供的领域知识,抽取家谱中的隐性信息,对抽取结果进行扩充.例如,利用人物属性“辈份”添加隐性父子关系:若存在人物集合 $P=\{p_1,p_2,\ldots,p_m\}$的辈份为 n,且辈份为 $n-1$ 的人物仅有一个,记为人物 B,则我们可以认为,人物 B 为人物集合 P 的父亲.

2.4 数据规范层

数据规范层的主要功能是将数据抽取层中抽取到的信息用一个统一的标准规范化,将这些数据转换为一种统一的描述,则有利于消除信息的语义异构性.但不同领域通常具有不同的数据规范标准.本文提出在 OI 提供的领域数据规范标准的基础上,制定数据规范原则,具体原则如下.

(1) 化“繁”为“简”原则

对于中文数据,经常会出现繁体字信息.为了数据的统一性,我们需要将所有的繁体字转换为简体字进行存储.目前有许多开源工具类可以进行中文繁简体的转换,如 ZHConverter(https://github.com/program-in-chinese/zhconverter),OpenCC(https://github.com/BYVoid/OpenCC),OpenCC4j(https://github.com/houbb/opencc4j)等.

(2) 标准化原则

对于数据源中存在的表述不一致性问题,我们为不同的数据类型制定一个统一的标准,将数据标准化.人物属性信息值的数据类型主要有三类:字符串类型、数值类型和时间类型数据.我们分别为这三类数据制定一个标准.

- 对于数值类型数据,以阿拉伯数字为统一标准.如果同一人物属性的两个事实内容等价,仅是数值表示方式不同,则我们统一选择以阿拉伯数字表示的数据.假设有两条来自不同数据源的同一人物的信息:“张三享年七十二岁”和“张三享年 72 岁”.采用该条融合规则后,以“72 岁”作为人物“张三”的享年;
- 对于字符串类型数据,会存在缩写词、简称等表示方式,以名称的全称为统一标准.名称的缩写、简称等形式可能与另一名称的缩写或简称重合,造成歧义.例如,“南大”一词,可能指代“南京大学”,也可能指代“南昌大学”.因此,为了保证融合结果的清晰明确,在进行融合时,我们选择名称的全称或较为完整的数据;
- 对于时间数据,在家谱中时间大多数以字符串型数据存在,如“开皇十八年二月二十二日”“顺治乙酉年

© 中国科学院软件研究所 http://www.jos.org.cn

八月十二日”等.我们需要将这类数据转换成常见的数据类型,即“yyyy 年 mm 月 dd 日”.鉴于家谱领域数据的特殊性,我们根据 OI 提供的家谱领域数据规范标准,人工构建外部语义知识库,对特殊属性值进行规范化.例如:对于上述提及的时间属性,人工构建古代皇帝年号时间表(见表 2)和中国古代纪年表(见表 3),计算标准化后的日期.

Table 2 Years of ancient emperors

表 2 古代皇帝年号时间表

年号	开皇	仁寿	…	顺治	康熙	雍正	乾隆	嘉庆	…
起始年份(年)	581	601	…	1644	1662	1736	1736	1796	…

Table 3 Chronology of ancient China

表 3 中国古代纪年表

古代纪年	甲子	乙丑	…	甲申	乙酉	丙戌	丁亥	戊子	…
年份(年)	1804	1805	…	1824	1825	1826	1827	1828	…

2.5 数据融合层

数据融合层是数据融合框架中的核心层,为数据应用层的接口和服务提供数据基础.数据融合层通过使用机器学习技术(AI),将上一层规范后的数据融合成一个统一、互联的数据网络,具体过程可分为以下 3 个部分.

(1) 实体对齐,找出不同数据源中的相同实体,以便进行下一步的融合;

(2) 冲突消解,解决不同数据源中对于同一实体的数据冲突问题;

(3) 属性融合,通过对不同数据源中实体属性的融合,从多源、碎片化的数据中凝练出实体的统一的准确的描述.

2.5.1 家谱人物对齐

对于海量多源的家谱数据,如果我们对其中的人物做人工标注,将耗费大量的精力.因此,本文使用无监督的方法对家谱中存在的相同人物进行识别.无监督的实体对齐方法的主要思想是,利用相似性等特征将实体聚类到同一类别中[36].

在对大量家谱数据进行研究与分析的基础上,HI 构建家谱领域内候选实体对生成规则库,组成候选实体对,之后,利用基于相似性的无监督实体对齐算法(AI)判断候选实体对中的人物是否相同.实体对齐算法可以分成两个子模块:候选实体对生成和候选实体对相似度计算.

(1) 候选实体对生成

在该模块,检测出两份家谱中所有可能相同的人物,组成候选实体对.为了提高召回率并且尽可能全面地检测出人物的候选实体,我们在对家谱数据分析后,总结出在家谱数据中存在相同人物的情况,如下所示.

- 两个人物姓和名完全相同.同名人物为相同人物是家谱数据中最常见的情况;
- 两个人物姓相同,名部分相同.家谱中的人物姓名通常由“姓+辈份+名”组成,但有时人物姓名仅为“姓+名”.例如,“吴自忠”的辈份为“自”,则“吴忠”可能也指代“吴自忠”;
- 两个人物姓相同,一人的名与另一人的字或号完全相同.在一些家谱中,会存在以人物的字或号表示人物的情况.例如,唐朝诗人“李白”字“太白”,因此“李太白”也指代“李白”;
- 两个人物姓不同,名完全相同.随着时间的推移,“姓氏改易”现象经常发生.皇室赐姓、家族迁徙、人物过继等情况均会导致姓氏的变化.因此,同一人物在不同时期可能具有不同的姓氏,出现同名不同姓的情况;
- 两个人物相同,则其后代极有可能相同.如果我们已经确定两份家谱中的人物相同,那么其后代也有很大可能为相同人物.

HI 将上述总结出的家谱数据相同人物的特点转换成计算机能够读取的语言,设计一组简单有效的候选实体生成规则,使用基于规则的候选实体生成方法为每个待判断的人物生成一系列候选实体,组成候选实体对.该

© 中国科学院软件研究所 http://www.jos.org.cn

方法不仅能充分考虑到所有可能性的发生,提供较为全面的候选实体对,保证了结果的召回率,还大大降低了实体对齐的计算复杂度.

(2) 候选实体对相似度计算

在这一模块中,我们采用无监督的基于相似性的实体对齐方法(AI),通过计算候选实体对之间的相似度,判断候选实体对中的两个实体是否指代真实世界中的同一个实体.根据对家谱人物数据的语义信息与特征的观察与分析,我们将候选实体对之间的人物相似度分为语义相似度和关系相似度两部分.

a) 语义相似度

人物的语义信息是判断人物是否相同的重要依据.语义相似度用来测量人物之间语义信息的相似度.给定两个待判断的人物 e_i 和 e_j,$p=\{p_1,p_2,\ldots,p_n\}$为人物相同属性的集合.我们通过两个人物之间相同属性的属性值相似度来计算两个人物之间的语义相似度,语义相似度计算公式如下:

$$\phi(e_i,e_j)=\sum_{i=1}^{n}\omega L_{p_i}(e_i,e_j) \tag{1}$$

其中,$p=\{p_1,p_2,\ldots,p_n\}$表示人物相同属性的集合;$L_{pi}(e_i,e_j)$表示第 i 个属性 p_i 的字符串相似度,计算方法选择较为常用的 Levenshtein 编辑距离[46];ω为每个属性相似度的权重.这里.我们认为每个属性的重要程度相同,即每个属性的权重相同,若属性的个数为 n,则属性权重为 $1/n$.

b) 关系相似度

家谱数据中,每个人物除了具有语义信息以外,人物与人物之间还拥有大量的亲属关系.人物之间关系的相似度,也是判断人物是否相同的一个重要依据.本文采用基于 Jaccard 相关系数[47]的关系相似度计算方法.给定两个待判断的人物 e_i 和 e_j,其关系相似度计算公式如下:

$$R(e_i,e_j)=\frac{|R(e_i)\cap R(e_j)|}{|R(e_i)\cup R(e_j)|} \tag{2}$$

其中,$R(e_i)$代表人物 e_i 的亲属关系,$|R(e_i)\cap R(e_j)|$表示人物 e_i 和 e_j 相同的人物关系数量,$|R(e_i)\cup R(e_j)|$表示人物 e_i 和 e_j 所拥有的人物关系数量总和.判断人物 e_i 和 e_j 的关系是否相同时,为了便于比较,如果人物的对应关系人物的姓名相同,则我们认为人物 e_i 和 e_j 的关系相同.

综上所述,候选实体对相似度的计算公式如下:

$$Sim(e_i,e_j)=\gamma\phi(e_i,e_j)+\delta R(e_i,e_j) \tag{3}$$

其中,γ和δ分别为语义相似度和关系相似度的权重,用来平衡二者在人物相似度测量中的重要程度.通过对家谱数据的观察发现:在家谱人物的对齐中,人物的关系相似度比语义相似度更重要,更能反映两个人物是否为同一个人.举例来说,如果两个人物的父亲和儿子的姓名均相同,无需考虑人物的属性,就基本可以判断这两个人物为同一个人.并且,家谱数据中人物的属性信息会存在稀疏性的情况,此时人物的语义相似度对家谱人物的对齐贡献度较小.因此,考虑家谱数据的实际情况,我们适当增加关系相似度的权重δ.具体的权重设置如下:a) 如果家谱数据的属性稀疏,即属性的个数小于 5,则γ=0.2,δ=0.8;b) 如果家谱数据的属性充足,即属性的个数大于等于 5,则γ=0.4,δ=0.6.本文设置一个阈值 S,若相似度分数 $Sim(e_i,e_j)$大于阈值 S,则说明两个人物相同.

2.5.2 家谱数据冲突消解

针对家谱数据冲突问题,充分考虑到家谱领域特性、分布数据源中的表述不完整性、数据本身可能存在的不一致等,对这些问题进行分析、处理,在 OI 对家谱数据真值进行审核和确认后,本文将家谱人物属性分为两类——单真值属性和多真值属性,并对不同类别的属性采用不同的冲突消解机制.

(1) 单真值属性

对于单真值属性,如人物的性别、出生日期、过世日期等,有且仅有一个真值.多数投票规则是指:若某一个值是多数信息源都投票赞成的,则认为这个值有更大的代表性[48].通常来说,对同一实体属性,出现次数最多的事实往往是准确的:

$$\mathrm{Max}\mathit{Frequence}(ea,f)\Rightarrow \mathit{IsAccurate}(f) \tag{4}$$

© 中国科学院软件研究所 http://www.jos.org.cn

(2) 多真值属性

对于多真值属性,如人物的描述信息,如人物简介、成就等,没有标准的正确描述,人物的职业、官职等信息由于时间的推移,会存在多个不同的真值.因此我们认为:如果同一实体属性 *ea* 的事实的内容是相互补充的,则它们合并后具有更高的准确性.为了保证最终融合结果的全面性,采用合并原则,将多数据源的不同描述信息整合后生成一个更为完整的信息:

$$\bigcup_{i=1}^{L}(ea, f_i) \Rightarrow IsAccurate(f) \tag{5}$$

2.5.3 家谱数据属性融合

通过对大量家谱数据的研究与分析,我们发现,家谱人物属性中主要存在以下两种特殊情况.

a) 属性名称不同、含义相同.随着时间的推移,古代人物的一些属性可能逐渐演变为具有现代特色的属性,存在“属性演变”的情况.例如,古代人物的“官职”属性与现代人物的“职务”属性名称不同却具有相同的含义;

b) 属性名称相同、含义不同.例如时间属性,时间属性有农历和公历之分:我国古代传统历法为农历,1912年后开始渐渐使用公历.因此,家谱记载此年之前的时间通常为农历,而后的时间通常为公历.

对于上述情况,现有的单纯依靠计算机的属性融合方法均难以解决.考虑到家谱数据的特殊性,为了保证融合结果的正确性,根据OI提供的家谱数据格式标准,HI即领域专家们针对家谱领域内对数据的特性及家谱人物属性的需求,人工构建属性语义知识库,使用一种基于启发式的方法进行家谱人物属性融合,具体过程见算法1.

算法 1. 属性融合算法.

输入:属性集合 *PRO*,属性拆分规则库 *split_rules*,等价属性知识库 *equal_rules*,人物属性集合 *PER_PRO*;

输出:融合后的属性集合 *PRO*.

1: 初始化属性集合 *PRO*={*name*,*gender*,…}
2: **for each** *per_pro*∈*PER_PRO* **do**
3: **if** *per_pro* 满足属性拆分规则库 *split_rules* 中的规则 **then**
4: 拆分属性 *per_pro*
5: **end if**
6: **if** *per_pro* 满足等价属性知识库 *equal_rules* 中的规则 **then**
7: 标准化属性 *per_pro*
8: **end if**
9: **if** *per_pro*∉*PER_PRO* **then**
10: *PRO*=*PRO*∪{*per_pro*}
11: **end if**
12: **end for**

本文构建的家谱属性语义知识库包括:

(1) 属性拆分规则库:主要针对属性名称相同却含义不同的属性.例如:时间属性有公历和农历之分,若属性值中包含表2和表3中“年号”或“古代纪年”中的值,则该时间属性为农历时间,在属性名字前添加“农历”二字后进行存储;反之,则为公历时间并添加“公历”二字.在数据规范层中,我们已经对农历时间进行星号标记,因此可以直接为带有(*)标记属性值的属性名称添加“农历”二字;

(2) 等价属性知识库:主要针对属性名称不同却含义相同的属性.根据专家(HI)提供的领域知识,考虑到“属性演变”情况,对等价的属性进行整理并记录,并为其规定一个标准属性名称.例如“官职”⇔“职务(·)”,“职务”为标准属性名称,其等价属性最终均映射为“职务”属性.

© 中国科学院软件研究所 http://www.jos.org.cn

3 结果展示及对比分析

3.1 结果展示

(1) 数据获取层结果

本文选取 4 份家谱数据为例,展示其运行结果.文本是家谱数据中较为常见的数据类型,因此本文选取的家谱示例均为文本格式.在家谱文本数据中,每个人物的描述信息独立成段,如图 2(a)~图 2(d)所示.由于家谱数据篇幅较长,本文仅截取家谱部分内容以供展示.

黄帝(公元前 2733 年~公元前 2599 年):少典之子,古华夏部落联盟首领,中国远古时代华夏民族的共主.五帝之首.被尊为中华"人文初祖".史载黄帝因有土德之瑞,故号黄帝.黄帝以统一华夏部落与征服东夷、九黎族而统一中华的伟绩载入史册.黄帝在位期间,播百谷草木,大力发展生产,始制衣冠、建舟车、制音律、创医学等.

玄嚣,黄帝之子,上古部落首领.

(a)

人文始祖:黄帝(公元前 2733 年~公元前 2598 年).姬姓,少典之子.黄帝娶有 4 妃:西陵氏、方雷氏、彤鱼氏和嫘母.4 室有 25 个儿子.西陵氏嫘祖,是黄帝的元妃,生有 2 子,长子叫玄嚣,次子叫昌意.

黄帝二世:玄嚣,黄帝长子.

黄帝三世:蟜极,玄嚣之子.

黄帝四世:帝喾,蟜极之子,又称高辛氏.

黄帝五世:后稷,帝喾之子,名弃,周朝始祖.初仕尧,官司农,教民稼穑;继佐舜,亦官大司农,播种五谷,封国于邰.

(b)

始祖:黄帝(前 2733~前 2598),姓公孙,少典之子,母曰附宝.生于轩辕寿坵,故号轩辕氏;居于姬水,因改姓姬,国于有熊,也称有熊氏.黄帝生性灵活,能说会道,道德情操高尚,聪明敏捷,多智善谋,被拥为西北方游牧部族之首领,后统一各部落,代神农而成为部落联盟之首领,称为"黄帝".公元前 2697 年登基,时 37 岁,活 111 岁,卒葬陕西黄陵县的桥山上.娶西陵氏之女嫘祖为正妻.据说黄帝共娶 4 妃,生子 25 人,其中有昌意、玄嚣等.

(c)

张士绾,字素卿,号澹岩.生明万历庚子年十二月二十九日午时,卒顺治乙酉年八月十二日子时.福建福州府卫经历.配龙氏,文学、讳承锦女,生万历己亥年八月二十日巳时,卒康熙癸卯年二月二十一日亥时.合葬龙旺山四甲蟠龙地,亥山巳向.

五子:秉彡、晓、秉骥、秉豫、曙.

三女:长适余能及,次适庐江王承宠,三适龙璋.

(d)

Fig.2 Genealogy data

图 2 家谱数据示例

(2) 数据抽取层结果

获取家谱数据后,将数据送入数据抽取层,进行信息抽取.为了方便查看,将数据抽取结果以表格形式展示,如表 4(a)~表 4(d)所示.每一行为一条人物信息,每一列分别为人物的属性.由表 4(a)~表 4(d)中可以看出:在数据抽取层中,除一些人物基础属性如"姓名""性别"外,不同家谱中能抽取出的人物属性不尽相同.例如,表 4(c)中人物具有"出生地"属性,表 4(d)中人物具有"字""号"和"官职"等属性.

另外,由表 4(a)~表 4(d)可见:本文提出的基于 HAO 模型的通用家谱信息抽取方法,在 HI 和 OI 的协助下,实现对语义的理解,从而较为有效地对家谱数据中的人物属性和关系进行抽取,能保证数据抽取结果的正确性.

Table 4(a) Results of data extraction in Fig.2(a)

表 4(a) 图 2(a)展示内容的数据抽取结果

编号	姓名	性别	出生日期	过世日期	配偶编号	父亲编号	母亲编号	简介
1	少典	男						
2	黄帝	男	公元前 2733 年	公元前 2599 年		1		古华夏部落联盟首领,中国远古时代华夏民族的共主…
3	玄嚣	男				2		上古部落首领

© 中国科学院软件研究所 http://www.jos.org.cn

Table 4(b) Results of data extraction in Fig.2(b)

表 4(b) 图 2(b)展示内容的数据抽取结果

编号	姓名	性别	出生日期	过世日期	配偶编号	父亲编号	母亲编号	简介
1	少典	男						
2	黄帝	男	公元前2733年	公元前2598年	3/4/5/6	1		
3	西陵氏	女			2			
4	方雷氏	女			2			
5	彤鱼氏	女			2			
6	嫫母	女			2			
7	玄嚣	男				2	3	
8	昌意	男				2	3	
9	蟜极	男				7		
10	帝喾	男				9		
11	后稷	男				10		名弃,周朝始祖.初仕尧,官司农,教民稼穑;继佐舜,亦官大司农,播种五谷,封国于邰.

Table 4(c) Results of data extraction in Fig.2(c)

表 4(c) 图 2(c)展示内容的数据抽取结果

编号	姓名	性别	出生日期	过世日期	配偶编号	父亲编号	母亲编号	出生地	简介
1	少典	男							
2	黄帝	男	公元前2733年	公元前2598年	3	1		轩辕寿坵	黄帝生性灵活,能说会道,道德情操高尚,聪明敏捷,…
3	嫘祖	女			2				
4	玄嚣	男				2	3		
5	昌意	男				2	3		

Table 4(d) Results of data extraction in Fig.2(d)

表 4(d) 图 2(d)展示内容的数据抽取结果

编号	姓名	性别	字	号	出生日期	过世日期	配偶编号	父亲编号	母亲编号	朝代	官职	葬于
1	张士绾	男	素卿	澹岩	明万历庚子年十二月二十九日	顺治乙酉年八月十二日	2			明	福建福州府卫经历	龙旺山四甲蟠龙地
2	龙氏	女			万历己亥年八月二十日	康熙癸卯年二月二十一日	1			明		
3	张秉豸	男						1	2	明		
4	张晓	男						1	2	明		
5	张秉骥	男						1	2	明		
6	张秉豫	男						1	2	明		
7	张曙	男						1	2	明		
8	张士绾长女	女					9	1	2	明		
9	余能及	男					8			明		
10	张士绾次女	女					11	1	2	明		
11	王承宠	男					10			明		
12	张士绾三女	女					13	1	2	明		
13	龙璋	男					12			明		

(1) 数据规范层结果

以表 4(d)中“顺治乙酉年八月十二日”为例,经过分析可知:“顺治”为中国古代皇帝年号,“乙酉年”为中国古代纪年,一甲子(60 年)为一个循环.为了将其转换为标准日期格式,表 2 为古代皇帝年号表,表 3 为中国古代纪年表.由表 2 可知,“顺治乙酉年”在 1644 年~1661 年之间.由表 3 可知,“顺治乙酉年”与 1825 年的差是 60 的整数倍.

© 中国科学院软件研究所 http://www.jos.org.cn

因此,“顺治乙酉年”为 1645 年,“顺治乙酉年八月十二日”应标准化为“1645 年 8 月 20 日”.对于如“开皇十八年十二月二十二日”这种形式的日期,在年号的基础上加上相应年份数即可.因此,“开皇十八年十二月二十二日”可转换为“598 年 12 月 22 日”.另外,对转换后的日期进行十字星号标记(†),以便于下一层的数据融合.表 4(d)规范化后的结果如表 5 所示.表 4(a)~表 4(c)在数据规范层中的输出结果不变.

由表 5 可见,本文提出的数据规范方法能够简单有效地将家谱中的人物属性值转换为统一的描述,特别是家谱中较难处理的时间类型数据,为下一步家谱数据的融合提供了便利.

Table 5 Results of data specification

表 5 数据规范结果

编号	姓名	性别	字	号	出生日期	过世日期	配偶编号	父亲编号	母亲编号	朝代	官职	葬于
1	张士绾	男	素卿	澹岩	1600 年 12 月 29 日(†)	1645 年 8 月 12 日(†)	2			明	福建福州府卫经历	龙旺山四甲蟠龙地
2	龙氏	女			1599 年 8 月 20 日(†)	1663 年 2 月 21 日(†)	1			明		
3	张秉豸	男						1	2	明		
4	张晓	男						1	2	明		
5	张秉骥	男						1	2	明		
6	张秉豫	男						1	2	明		
7	张曙	男						1	2	明		
8	张士绾长女	女					9	1	2	明		
9	余能及	男					8			明		
10	张士绾次女	女					11	1	2	明		
11	王承宠	男					10			明		
12	张士绾三女	女					13	1	2	明		
13	龙璋	男					12			明		

(2) 数据融合层结果

a) 实体对齐

以表 4(a)中编号为 2 的人物“黄帝”(记为“(a)2”)为例,根据上述提到的候选实体对生成的情况,为该人物在表 4(b)~表 4(d)中选取候选实体.生成的候选实体对为〈“(a)2”,“(b)2”〉,〈“(a)2”,“(c)2”〉,〈“(b)2”,“(c)2”〉.然后,对每个候选实体对使用第 3.4.1 节中的公式(3)进行相似度计算.当γ和δ分别取 0.4 和 0.6、阈值设为 0.5 时,结果如表 6 所示.最终结果表明,表 4(a)中编号为 2 的人物“黄帝”与表 4(b)、表 4(c)中的人物“黄帝”为同一人.

根据家谱内容,我们可以看出:本文所提的实体对齐算法最终识别结果,即表 4(a)~表 4(c)中的人物“黄帝”均为同一人,是与现实世界一致的.这一结果表明,本文所提的实体对齐算法在实体为家谱人物时的对齐结果是准确有效的.

Table 6 Results of entity alignment

表 6 实体对齐相似度结果

候选实体对	语义相似度	关系相似度	实体对相似度
〈“(a)2”,“(b)2”〉	0.775	0.44	0.574
〈“(a)2”,“(c)2”〉	0.789	0.667	0.7158
〈“(b)2”,“(c)2”〉	0.8	0.545	0.647

b) 冲突消解

观察我们识别出的相同人物“黄帝”的属性信息,发现表(a)中人物“黄帝”的过世日期与表 4(b)和表 4(c)不同.根据我们制定的冲突消解机制,过世日期为单真值属性,利用公式(4)得出,“黄帝”的属性过世日期的真值为“公元前 2598 年”.由此看出,我们可以根据本文提出的数据冲突机制,简单高效地解决不同来源的数据中出现的数据冲突问题.

© 中国科学院软件研究所 http://www.jos.org.cn

c) 属性融合

根据第 3.4.3 节中描述的属性融合过程,读取属性拆分规则库,对时间属性“出生日期” “过世日期”进行拆分,拆分结果为“农历出生日期”“农历过世日期”“公历出生日期”和“公历过世日期”.另外,读取等价属性知识库,我们可知“官职”属性和“职务”属性等价,因此将“官职”映射为“职务”.

表 7 展示了本文选取的 4 份家谱数据的数据融合结果,从结果可以看出:我们能够将不同来源的碎片化家谱数据中的人物进行融合,凝练出一套的关于家谱人物的统一描述,进而表明本文提出的碎片化数据融合框架 FDF-HAO 在技术上的可行性和有效性.最终家谱人物数据的属性集合除了表 7 所展示的属性外,还包括“曾用名”“世”“辈份”“家庭排行”“住址”等属性.

Table 7 Results of data fusion

表 7 数据融合结果

编号	姓名	性别	字	号	公历出生日期	公历过世日期	农历出生日期	农历过世日期
1	少典	男						
2	黄帝	男			公元前 2733 年	公元前 2598 年		
3	西陵氏	女						
4	方雷氏	女						
5	彤鱼氏	女						
6	嫫母	女						
7	玄嚣	男			公元前 2422 年	公元前 2322 年		
8	昌意	男						
9	蟜极	男						
10	帝喾	男			公元前 2480 年	公元前 2345 年		
11	后稷	男			公元前 2300 年			
…	…							
100	张士绾	男	素卿	澹岩			1600 年 12 月 29 日	1645 年 8 月 12 日
101	龙氏	女					1599 年 8 月 20 日	1663 年 2 月 21 日

接下表

配偶编号	父亲编号	母亲编号	职务	朝代	葬于	简介
			部落首领	远古		
3/4/5/6	1			远古		黄帝生性灵活,能说会道,道德情操高尚,聪明敏捷,…古华夏部落联盟首领,中国远古时代华夏民族的共主…
2				远古		西陵氏嫘祖,是黄帝的元妃,生有 2 子,长子叫玄嚣,次子叫昌意.
2				远古		
2				远古		
2				远古		
	2	3	部落首领	远古		黄帝长子.上古部落首领.
	2	3		远古		
	7			远古		玄嚣之子
	9			远古		蟜极之子,又称高辛氏.
	10			远古		名弃,周朝始祖.初仕尧,官司农,教民稼穑;继佐舜,亦官大司农,播种五谷,封国于邰.
100			福建福州府卫经历	明	龙旺山四甲蟠龙地	
101				明		

3.2 结果分析

本小节将碎片化数据融合框架 FDF-HAO 中数据抽取层和数据融合层所采用的技术与同类技术的进行对比和分析.

3.2.1 数据抽取层

我们将本文所采用的信息抽取方法与目前较为成熟的开源信息抽取工具 DSNFs[49]和 Jiagu[50]进行对比.以

© 中国科学院软件研究所 http://www.jos.org.cn

图 2(b)展示的家谱为例,表 8 展示各方法对人物“黄帝”的相关抽取结果.

Table 8 Data extraction results by different methods

表 8 数据抽取对比结果

方法	抽取结果
我们的方法	[“黄帝”,“父亲”,“少典”]、[“黄帝”,“妻子”,“西陵氏”]、[“黄帝”,“妻子”,“方雷氏”]、[“黄帝”,“妻子”,“彤鱼氏”]、[“黄帝”,“儿子”,“玄嚣”]、、[“黄帝”,“儿子”,“昌意”]
DSNFs[49]	[“黄帝”,“有”,“西陵氏”]、[“黄帝”,“有”,“方雷氏”]、[“黄帝”,“有”,“彤鱼氏”]、[“西陵氏嫘祖”,“妃”,“黄帝”]
Jiagu[50]	–

由表 8 可以看出:我们的方法在家谱数据上能够准确全面地抽取出人物间关系和人物属性,DSNFs 仅能抽取部分信息,而 Jiagu 未能抽取到人物信息.其原因在于:DSNFs 和 Jiagu 均是在依存句法分析的基础上对实体和关系进行抽取,这类方法受限于中文分词等 NLP 技术的性能,适用于文本句法结构简单、NLP 技术能对文本进行有效分析和处理的情况下.但家谱数据的用词语法与我们常用的文本不同,行文风格偏向古文,甚至一些家谱不包含完整的语句.由于家谱数据的文本特点,现有主流信息抽取工具通常很难有效地对家谱文本中不同成分的结构关系进行提取.为此,我们的方法针对家谱数据特点进行设计,通过分析家谱中的浅层词法特征,在 OI 提供的领域知识下,结合专家(HI)对家谱数据的分析,能够有效地对家谱信息进行抽取.

3.2.2 数据融合层

数据融合层中最为关键的一步为家谱人物对齐,下文对家谱人物对齐方法进行对比分析.鉴于家谱人物对齐过程分为两部分——候选实体对生成和候选实体对对齐,本文将从这两部分对算法的性能进行对比分析.

(1) 候选实体对生成方法

目前,实体对齐算法中,候选实体对生成的方法通常为基于字符串相似度和基于词典的方法.基于字符串相似度的方法容易产生大量不能对齐的候选实体,导致后续算法的计算复杂度增加.基于词典的方法需要人工构建词典,从词典中寻找所有可能对齐的实体.而构建词典的过程将耗费大量人力物力.本文通过对家谱数据的分析,制定了一套家谱领域内候选实体对生成规则,采用基于规则的方法为待对齐实体生成候选实体.优点在于:一方面能够保证候选实体集合中包含可以对齐的实体,即保证了结果的召回率;另一方面,也避免了不能对齐的候选实体数目过多,降低了后续计算的复杂度.

(2) 候选实体对对齐方法

在缺乏训练数据的情况下,除本文使用的基于相似性的实体对齐方法外,还可以采用基于词嵌入的方法,将实体及其上下文转换为词向量进行相似度计算.但词向量的训练过程通常需要大规模语料库或少量种子数据,生成词向量的好坏依赖于语料库或种子数据的质量[51,52].这类方法适用于语料库或标记数据质量较为成熟、训练出的词向量效果好、能很好地表示实体语义信息的情况下.而家谱数据领域性较强,缺乏适合的语料库.在家谱中,判断两个人物是否相同的依据就是实体之间的属性及关系是否相同.本文采用的基于相似性的实体对齐方法,考虑了实体的属性及实体间关系的相似性,相较于其他实体对齐方法,能够根据家谱领域特点,简单高效地计算家谱领域内实体之间的相似性.

4 碎片化数据融合的挑战和前景

面向多源异构的碎片化家谱数据,本文提出的碎片化数据融合框架能够对其进行有效融合,但仍存在一些挑战.

- 挑战 1:数据的多模态性

在大数据时代,碎片化数据以文本、图片、视频、音频等不同模态存在.我们在处理这些数据时,需要对其中包含的内容进行识别、提取并存储.但由于不同模态数据之间的结构差异巨大,没有统一的数据表示形式和统一的逻辑结构,这使得多模态数据的融合具有一定的挑战性.另外,互联网中的多模态数据如图片、视频等存在着模糊、有噪声等情况,因此,多模态数据的信息抽取精度无法得到保证,从而对多模态数据的融合精度造成

© 中国科学院软件研究所 http://www.jos.org.cn

一定的负面影响.

- 挑战 2:数据的不确定性

数据真伪难辨是数据处理及应用的最大挑战[12].海量多源的碎片化数据,使我们的研究获得了前所未有的大规模样本,但也带来了更多错误的、不完整的数据.数据质量良莠不齐,不同来源的数据值可能存在冲突、缺失、描述模糊等情况.为了从海量多源的碎片化数据中准确地找出真实确定的数据,需要利用数据处理方法对数据、数据源等信息进行建模求解.但对于一些数据,即使最好的数据处理方法也难以消除其固有的不可预测性.例如在家谱领域内,一份家谱中的同一人物在不同版本中存在姓名不同的情况.根据家谱内容,我们无法确定造成不同的原因是人物的姓名更改还是书写时的笔误,因此该人物的姓名具有无法消除的不确定性.

- 挑战 3:数据的单源小体量性

碎片化数据最显著的特征就是单源小体量性.来自单个数据源的碎片化数据通常内容较短,包含的信息不充足,数据具有较高的稀疏性.因此,在对碎片化数据进行信息抽取和融合时,大多需要借助外部语义知识库中的语义信息.这种方法虽然能提高算法的精确度,但对外部知识库依赖度较高.当出现知识库中不存在的信息时,需要对知识库进行及时地更新,否则将无法提取新的信息.

- 挑战 4:数据的语义异构性

不同数据源的碎片化数据在语义表述上存在一定的差异性,相同含义的词汇具有不同的表述,我们将之称为语义异构.数据的语义异构性可能会造成来自不同数据源的碎片化数据无法相互融合,进而导致数据共享、重用无法进行,因此我们必须考虑消除碎片化数据之间的语义异构性.通常来说,我们采用将不同数据源的数据映射到同一套概念体系即本体的方法来解决语义异构.但是本体的构建本身就是一个工作量大的任务.另外,大数据时代中数据的不断更新也会带来一些新的概念,这就需要一个合适的机制对本体进行不断地更新和维护.

碎片化数据融合在多源数据分析和大知识融合领域具有广泛的研究和利用前景,下面我们分析几个应用场景.

- 应用场景 1:同姓家谱的知识扩充以及跨姓家谱的知识挖掘和推理.

碎片化家谱数据融合有利于同姓家谱的合并与扩充.通过对已有的同姓家谱进行关联计算和合并计算,实现家谱的补齐和扩充,扩展知识网络.例如,假设存在两份同姓家谱 *A* 和 *B*,经过计算发现二者之间存在关联:家谱 *A* 记录某家族 *P* 从第 1 世~第 20 世的人物信息,家谱 *B* 记录同一家族 *P* 从第 10 世~第 30 世的人物信息.合并家谱 *A* 和 *B*,我们可以得到一份全新的、更为完整的家谱 *C*,记录家族 *P* 从第 1 世~第 30 世的人物信息.另外,碎片化家谱数据融合也为跨姓家谱的知识挖掘和推理提供了数据支撑.通过对不同姓氏家谱的人物进行对比和分析,寻找跨姓家谱之间的相同人物,以该人物为纽带,建立家谱之间的关联,挖掘其中潜藏的姓氏起源、姓氏演变等信息.从家谱数据库中已有的数据出发,经过计算机推理,建立人物之间的新关联,从而拓展和丰富知识网络,推理人物间的爱恨情仇,为用户解决寻根溯源等问题.

- 应用场景 2:社交网络信息分析.

社交网络用户数量庞大,微博、推特、豆瓣等常见的社交平台上每天产生大量的图片、文字及音频信息.这些碎片化社交数据中隐藏着许多有用的信息,包括用户的日常琐事、兴趣爱好、热点事件的发展过程等等.通过对碎片化社交数据的融合,以用户为中心,构建用户社交知识图谱,预测用户之间潜在的联系,为其提供好友推荐、信息推送等个性化社交服务.

5 总 结

本文在 HI、AI 和 OI 三者的交互和协同下,提出了一个碎片化数据融合框架 FDF-HAO,并论述了碎片化数据融合框架的层次结构,详细介绍了每一层的作用、所需要解决的问题和使用的技术.其中,数据获取层使用爬虫技术(AI),从互联网中各数据源获取碎片化数据,包括文本文件、表格文件、网页文件等;数据抽取层通过自然语言处理技术(AI),在 HI 和 OI 的交互和协作下,从这些多源异构的碎片化数据中提取实体、属性及关系;数据规范层根据 OI 提供的领域数据规范标准,负责将数据抽取层中抽取的信息进行规范化和标准化,消除了数据

© 中国科学院软件研究所 http://www.jos.org.cn

的语义异构性;数据融合层是实现数据融合的核心层,领域专家们(HI)在 OI 的协作下构建外部语义知识库,为数据融合提供智能支持,然后通过实体对齐技术(AI)识别出碎片化数据中的相同实体,再通过冲突消解机制(AI)从冲突数据中寻找数据的真值,最后通过属性融合(AI)凝练出实体的统一的、准确的、有用的描述,进而完成数据的融合,形成知识库.

与已有的特定领域知识图谱构建相似,本文是在现有的知识图谱构建技术的基础上,通过对数据的观察和分析,对技术进行优化和改进.但不同之处在于:本文结合 HAO 智能模型,通过 HI、AI 和 OI 三者的交互和协作,为海量多源异构的碎片化数据融合提供了智能支持,能够解决一些仅依靠计算机无法解决的问题.另外,本文结合家谱领域特征,将家谱领域知识贯穿于碎片化家谱数据融合的过程中,对各阶段结果进行约束和改进,有效地提高了数据融合结果的准确性和全面性.

本文以华谱系统中碎片化家谱数据融合过程为例,详细介绍所提框架在每层中的具体处理思路和方案,为解决碎片化数据融合问题和中文知识图谱构建问题提供了一个新思路,即:在现有成熟模型和方法的基础上,结合 HAO 智能模型,为中文知识图谱构建提供智能支持,以便更好地提高数据的准确性和可用性.另外,本文在框架内各层次中提出的方法也具有一定的通用性,对其他领域的中文知识图谱构建可能具有一定的借鉴意义.

目前,关于碎片化数据融合的研究尚处于初步阶段,仍存在着许多困难和挑战.本文通过对碎片化数据融合过程进行高度抽象和建模,提出了 FDF-HAO 框架,若将该框架迁移到其他领域,需根据领域数据特点调整 FDF-HAO 框架的各部分具体实现,存在一定的难度.例如在复杂的社交网络场景中,包含着以用户为中心的不同维度、不同领域的碎片化社交数据.但是随着互联网的不断发展,网络平台更新换代,网络词汇层出不穷,网络信息多元多样,社交网络数据在自然语言理解和分析方面上具有很大的挑战性,这为社交网络数据的信息抽取和融合增加了一定的难度.同样,在网页数据中也包含着大量涉及以人物为中心的人物生平、经历、传记、新闻等碎片化数据.然而在不同网络平台中,数据的描述方式和内容侧重点不同,并且存在着大量的数据不确定性、语义异构性等问题,因此给现有的数据融合研究带来了很大的挑战.在后续的研究中,我们将首先继续优化本文提出的 FDF-HAO 框架;接着,研究将该框架分别应用于融合碎片化的社交网络数据和互联网中碎片化的网页数据;最后,以构建整合的人物知识图谱为目标,将家谱、社交网络、网页这 3 个维度的碎片化数据进行融合,从亲属关系、社交关系、人物生平等多个维度构建更加完善的人物知识图谱,从而为用户提供更好的大知识服务.

References:

[1] Wang YZ, Jin XL, Cheng XQ. Network big data: Present and future. Chinese Journal of Computers, 2013,36(6):1125−1138 (in Chinese with English abstract). [doi: 0.3724/SP.J.1016.2013.01125]

[2] Wu XD, Chen HH, Wu GQ, Liu J, Zheng QH, He XF, Zhou AY, Zhao ZQ, Wei BF, Gao M, Li Y, Zhang QP, Zhang SC, Lu RQ, Zheng NN. Knowledge engineering with big data. IEEE Intelligent Systems, 2015,30(5):46−55. [doi: 10.1109/MIS.2015.56]

[3] Zhan L. Literature questions in genealogy. Journal of Peking University (Philosophy and Social Sciences), 2007,53(1):150−151 (in Chinese with English abstract). [doi: 10.16113/j.cnki.daxtx.2007.01.010]

[4] Huang XY. Analysis and enlightenment on the causes of the boom of utilization of foreign genealogy. Archives Science Bulletin, 2007,29(1):30−33 (in Chinese with English abstract).

[5] Wu XL. Chinese genealogy and its academic value. Historical Research, 1988,35(6):20−34 (in Chinese with English abstract).

[6] Ouyang K. The reform and innovation of big data and humanities and social science research. Guangming Daily, 2016-11-10(016) (in Chinese).

[7] Sun JJ. How to develop humanities and social sciences in the age of big data. Guangming Daily, 2014-07-07(011) (in Chinese).

[8] Xia CJ, Liu W, Chen T, Zhang L. A genealogy data service platform implemented with linked data technology. Journal of Library Science in China, 2016,42(3):27−38 (in Chinese with English abstract). [doi: 10.13530/j.cnki.jlis.160014]

[9] Chen T, Xia CJ, Liu W, Zhang L. Research and implementation of visualization technology of linked data. Library and Information Service, 2015,59(17):113−119 (in Chinese with English abstract).

[10] Mao JJ. Development and construction of digital genealogy resources in China. Archives and Construction, 2007,24(1):22−24 (in Chinese with English abstract).

© 中国科学院软件研究所 http://www.jos.org.cn

[11] Hu D, Wen YN, Lv GN, Shen JW. GIS-based family tree resources integration. Human Geography, 2012,27(1):50−53 (in Chinese with English abstract). [doi: 10.13959/j.issn.1003-2398.2012.01.010]

[12] Cheng XQ, Jin XL, Wang YZ, Guo JF, Zhang TY, Li GJ. Survey on big data system and analytic technology. Ruan Jian Xue Bao/Journal of Software, 2014,25(9):1240−1252 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4674.htm [doi: 10.13328/j.cnki.jos.004674]

[13] Li SQ, Ding H, Xu X. Considering on user service of digital library in the age of big data. Journal of the China Society for Scientific and Technical Information, 2018,37(6):569−579 (in Chinese with English abstract). [doi: 10.3772/j.issn.1000-0135.2018.06.002]

[14] Wu XD, He J, Lu RQ, Zheng NN. From big data to big knowledge: HACE+BigKE. Acta Automatica Sinica, 2016,42(7):965−982 (in Chinese with English abstract). [doi: 10.16383/j.aas.2016.c160239]

[15] Wu MH, Wu XD. On big wisdom. Knowledge and Information Systems, 2018,58(1):1−8. [doi: 10.1007/s10115-018-1282-y]

[16] Liu Q, Li Y, Duan H, Liu Y, Qin ZG. Knowledge graph construction techniques. Journal of Computer Research and Development, 2016,53(3):582−600 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2016.20148228]

[17] Grishman R, Sundheim B. Message understanding Conference-6: A brief history. In: Proc. of the Int'l Conf. on Computational Linguistics. New York: ACM, 1996. 466−471.

[18] Rau LF. Extracting company names from text. In: Proc. of the IEEE Conf. on Artificial Intelligence Application. Piscataway: IEEE, 1991. 29−32. [doi: 10.1109/CAIA.1991.120841]

[19] Yang JF, Yu QB, Guan Y, Jiang ZP. An overview of research on electronic medical recordoriented named entity recognition and entity relation extraction. Acta Automatica Sinica, 2014,40(8):1537−1562 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2014.01537]

[20] Lai PT, Huang MS, Yang TH, Hsu WL, Tsai RTH. Statistical principle-based approach for gene and protein related object recognition. Journal of Cheminformatics, 2018,10(1):64:1−64:9. [doi: 10.1186/s13321-018-0314-7]

[21] Hwang S, Hong JE, Nam YK, Towards effective entity extraction of scientific documents using discriminative linguistic features. KSII Trans. on Internet and Information Systems, 2019,13(3):1639−1658. [doi: 10.3837/tiis.2019.03.030]

[22] Akkasi A, Varoglu E. Improving biochemical named entity recognition performance using PSO classifier selection and bayesian combination method. IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2016,14(6):1327−1338. [doi: 10.1109/TCBB.2016.2570216]

[23] Liu L, Wang DB. A review on named entity recognition. Journal of the China Society for Scientific and Technical Information, 2018,37(3):329−340 (in Chinese with English abstract). [doi: 10.3772/j.issn.1000-0135.2018.03.010]

[24] Peng N, Dredze M. Improving named entity recognition for Chinese social media with word segmentation representation learning. In: Proc. of the Association for Computational Linguistics. Stroudsburg: ACL, 2016. 149−155.

[25] Qiu JH, Zhou YM, Wang Q, Ruan T, Gao J. Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field. IEEE Trans. on NanoBioscience, 2019,18(3):306−315. [doi: 10.1109/TNB.2019.2908678]

[26] Yang YJ, Xu B, Hu JW, Tong MH, Zhang P, Zheng L. Accurate and efficient method for constructing domain knowledge graph. Ruan Jian Xue Bao/Journal of Software, 2018,29(10):2931−2947 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5552.htm [doi: 10.13328/j.cnki.jos.005552]

[27] E HH, Zhang WJ, Xiao SQ, Cheng R, Hu YX, Zhou XS, Niu PQ. A survey of entity relationship extraction based on deep learning. Ruan Jian Xue Bao/Journal of Software, 2019,30(6):1−28 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5817.htm [doi: 10.13328/j.cnki.jos.005817]

[28] Gan LX, Wan CX, Liu DX, Zhong Q, Jiang TJ. Chinese named entity relation extraction based on syntactic and semantic features. Journal of Computer Research and Development, 2016,53(2):284−302 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2016.20150842]

[29] Chen LW, Feng YS, Zhao DY. Extracting relations from the Web via weakly supervised learning. Journal of Computer Research and Development, 2013,50(9):1825−1835 (in Chinese with English abstract). [doi:10.7544/issn1000-1239.2013.20130491]

[30] Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora. In: Proc. of the Association for Computational Linguistics. Stroudsburg: ACL, 2004. 415−422. [doi: 10.3115/1218955.1219008]

[31] Leng J, Jiang P. A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm. Knowledge-Based Systems, 2016,100:188−199. [doi: 10.1016/j.knosys.2016.03.008]

© 中国科学院软件研究所 http://www.jos.org.cn

[32] Ji G, Liu K, He S, Zhao J. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: Proc. of the Conf. on Artificial Intelligence. Menlo Park: AAAI, 2017. 3060−3066.

[33] Wu F, Weld DS. Autonomously semantifying Wikipedia. In: Proc. of the Conf. on Information and Knowledge Management. New York: ACM, 2007. 41−50. [doi: 10.1145/1321440.1321449]

[34] Zhao JS, Zhu QM, Zhou GD, Zhang L. Review of research in sutomatic keyword extraction. Ruan Jian Xue Bao/Journal of Software, 2017,28(9):2431−2449 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5301.htm [doi: 10.13328/j.cnki.jos.005301]

[35] Bleiholder J, Naumann F. Data fusion. ACM Computing Surveys, 2008,41(1):1−41. [doi: 10.1145/1456650.1456651]

[36] Zhuang Y, Li GL, Feng JH. A survey on entity alignment of knowledge base. Journal of Computer Research and Development, 2016,53(01):165−192 (in Chinese with English abstract). [doi: 10.7554/issn1000-1239.2016.20150661]

[37] Fellegi I, Sunter A. A theory for record linkage. Journal of the American Statistical Association, 1969,64(328):1183−1210. [doi: 10.1080/01621459.1969.10501049]

[38] Chen Z, Kalashnikov DV, Mehrotra S. Exploiting context analysis for combining multiple entity resolution systems. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM, 2009. 207−218. [doi: 10.1145/1559845.1559869]

[39] Guan SP, Jin XL, Wang YZ, Jia YT, Shen HW, Li ZX, Cheng XQ. Self-learning and embedding based entity alignment. Knowledge and Information Systems, 2019,59(2):361−386. [doi: 10.1007/s10115-018-1191-0]

[40] Li YL, Gao J, Meng CS, Li Q, Su L, Zhao B, Fan W, Han JW. A survey on truth discovery. SIGKDD Explorations, 2015,17(2):1−16. [doi: 10.1145/2897350.2897352]

[41] Dong XL, Gabrilovich E, Heitz G, Horn W, Murphy K, Sun S, Zhang W. From data fusion to knowledge fusion. Proc. of the VLDB Endowment, 2015,7(10):881−892.

[42] Yin X, Han J, Yu PS. Truth discovery with multiple conflicting information providers on the Web. IEEE Trans. on Knowledge and Data Engineering, 2008,20(6):796−808. [doi: 10.1109/TKDE.2007.190745]

[43] Lyu S, Ouyang W, Wang YQ, Shen HW, Cheng XQ. Truth discovery by claim and source embedding. In: Proc. of the Conf. on Information and Knowledge Management. New York: ACM, 2017. 2183−2186. [doi: 10.1109/TKDE.2019.2936189]

[44] Smid J, Neruda R. Comparing datasets by attribute alignment. In: Proc. of the IEEE Symp. on Computational Intelligence and Data Mining. Piscataway: IEEE, 2014. 56−62. [doi: 10.1109/CIDM.2014.7008148]

[45] Hankcs. HanLP. https://github.com/hankcs/HanLP/tree/1.x

[46] Navarro G. A guided tour to approximate string matching. ACM Computing Surveys, 2001,33(1):31−88. [doi: 10.1145/375360.375365]

[47] Monge AE, Elkan CP. The field matching problem: Algorithms and applications. In: Proc. of the Conf. on Knowledge Discovery and Data Mining. Menlo Park: AAAI, 1996. 267−270.

[48] Dong XL, Berti-Equille L, Srivastava D. Integrating conflicting data: The role of source dependence. Proc. of the VLDB Endowment, 2009,2(1):550−561. [doi: 10.14778/1687627.1687690]

[49] Jia S, Li M, Xiang Y. Chinese open relation extraction and knowledge base establishment. ACM Trans. on Asian and Low-Resource Language Information Processing (TALLIP), 2018,17(3):15−22. [doi: 10.1145/3162077]

[50] Ownthink. Jiagu. https://github.com/ownthink/Jiagu

[51] Zhu H, Xie RB, Liu ZY, Sun MS. Iterative entity alignment via joint knowledge embeddings. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 2017. 4258−4264. [doi: 10.24963/ijcai.2017/595]

[52] Trisedya BD, Qi J, Zhang R. Entity alignment between knowledge graphs using attribute embeddings. In: Proc. of the Conf. on Artificial Intelligence. Menlo Park: AAAI, 2019. 297−304. [doi: 10.1609/aaai.v33i01.3301297]

附中文参考文献:

[1] 王元卓,靳小龙,程学旗.网络大数据:现状与展望.计算机学报,2013,36(6):1125−1138. [doi: 0.3724/SP.J.1016.2013.01125]

[3] 湛庐.家谱中的文献问题.北京大学学报(哲学社会科学版),2007,53(1):150−151. [doi: 10.16113/j.cnki.daxtx.2007.01.010]

[4] 黄霄羽.国外家谱档案利用热潮之成因探析及启示.档案学通讯,2007,29(1):30−33.

[5] 武新立.中国的家谱及其学术价值.历史研究,1988,35(6):20−34.

[6] 欧阳康.大数据与人文社会科学研究的变革与创新.光明日报,2016-11-10(016).

[7] 孙建军.大数据时代人文社会科学如何发展.光明日报,2014-07-07(011).

© 中国科学院软件研究所 http://www.jos.org.cn

[8] 夏翠娟,刘炜,陈涛,张磊.家谱关联数据服务平台的开发实践.中国图书馆学报,2016,42(3):27–38. [doi: 10.13530/j.cnki.jlis.160014]

[9] 陈涛,夏翠娟,刘炜,张磊.关联数据的可视化技术研究与实现.图书情报工作,2015,59(17):113–119.

[10] 毛建军.中国家谱数字化资源的开发与建设.档案与建设,2007,24(1):22–24.

[11] 胡迪,温永宁,闾国年,沈敬伟.基于 GIS 的家谱资源整合集成研究.人文地理,2012,27(1):50–53. [doi: 10.13959/j.issn.1003-2398.2012.01.010]

[12] 程学旗,靳小龙,王元卓,郭嘉丰,张铁赢,李国杰.大数据系统和分析技术综述.软件学报,2014,25(9):1889–1908. http://www.jos.org.cn/1000-9825/4674.htm [doi: 10.13328/j.cnki.jos.004674]

[13] 李树青,丁浩,徐侠.大数据时代数字图书馆用户服务思考.情报学报,2018,37(6):569–579. [doi: 10.3772/j.issn.1000-0135.2018.06.002]

[14] 吴信东,何进,陆汝钤,郑南宁.从大数据到大知识:HACE+BigKE.自动化学报,2016,42(7):965–982. [doi: 10.16383/j.aas.2016.c160239]

[16] 刘峤,李杨,段宏,刘瑶,秦志光.知识图谱构建技术综述.计算机研究与发展,2016,53(3):582–600. [doi: 10.7544/issn1000-1239.2016.20148228]

[19] 杨锦锋,于秋滨,关毅,蒋志鹏.电子病历命名实体识别和实体关系抽取研究综述.自动化学报,2014,40(8):1537–1562. [doi: 10.3724/SP.J.1004.2014.01537]

[23] 刘浏,王东波.命名实体识别研究综述.情报学报,2018,37(3):329–340. [doi: 10.3772/j.issn.1000-0135.2018.03.010]

[26] 杨玉基,许斌,胡家威,仝美涵,张鹏,郑莉.一种准确而高效的领域知识图谱构建方法.软件学报,2018,29(10):2931–2947. http://www.jos.org.cn/1000-9825/5552.htm [doi: 10.13328/j.cnki.jos.005552]

[27] 鄂海红,张文静,肖思琪,程瑞,胡莺夕,周筱松.深度学习实体关系抽取研究综述.软件学报,2019,30(6):1–28. http://www.jos.org.cn/1000-9825/5817.htm [doi: 10.13328/j.cnki.jos.005817]

[28] 甘丽新,万常选,刘德喜,钟青,江腾蛟.基于句法语义特征的中文实体关系抽取.计算机研究与发展,2016,53(2):284–302. [doi: 10.7544/issn1000-1239.2016.20150842]

[29] 陈立玮,冯岩松,赵东岩.基于弱监督学习的海量网络数据关系抽取.计算机研究与发展,2013,50(9):1825–1835. [doi:10.7544/issn1000-1239.2013.20130491]

[34] 赵京胜,朱巧明,周国栋,张丽.自动关键词抽取研究综述.软件学报,2017,28(9):2431–2449. http://www.jos.org.cn/1000-9825/5301.htm [doi: 10.13328/j.cnki.jos.005301]

[36] 庄严,李国良,冯建华.知识库实体对齐技术综述.计算机研究与发展,2016,53(1):165–192. [doi: 10.7554/issn1000-1239.2016.20150661]

吴信东(1963－),男,博士,教授,博士生导师,主要研究领域为数据挖掘,大数据分析,知识工程.

周鹏(1987－),男,博士,讲师,主要研究领域为数据挖掘,粗糙集,特征选择,知识工程.

李娇(1996－),女,博士生,主要研究领域为数据挖掘和知识图谱.

卜晨阳(1992－),男,博士,讲师,主要研究领域为演化动态优化和知识图谱.

© 中国科学院软件研究所 http://www.jos.org.cn