DOI:10.13203/j.whugis20190424 文章编号:1671-8860(2022)05-0700-07

人物经历信息模型及其信息提取方法

张三强[1,2] 宋国民[1] 贾奋励[1] 陈令羽[1]

1 信息工程大学地理空间信息学院,河南 郑州,450001
2 69340部队,新疆 伊犁,835000

摘 要:在当前地理信息系统应用中,人物信息的时空解读非常重要,有助于地理研究者生成多种类型的专题地图,实现相关地理内容的表达。在分析现有人物数据模型特点的基础上,结合地理应用需求和信息提取技术的发展现状,提出了一种突出人物时空特征的经历信息模型。以网络百科数据为例,实现了模型中各要素的提取,有效解决了事件描述识别和位置信息提取两个重点问题。测试和分析结果表明,该事件描述的抽取方法具有较强的实用性,而位置信息提取方法在标注语料有限的情况下,也取得了一定的效果,得出了较好的实验结论。

关键词:人物经历信息;信息提取;事件描述识别;位置信息提取;条件随机场

中图分类号:P208 **文献标志码**:A

地理信息与人类活动息息相关,人物信息应用于地理信息系统时有非常重要的作用[1],尤其是在人文、历史、军事、旅游等领域,人物相关的信息可以生成人物群体地域性分布图、古代人物行迹图、政要出访图、人文专题旅游地图等形式多样的地图[2]。结构化的、深层次的人物信息能够有效帮助人物相关地理内容的表达,满足用户对人物相关位置信息的探索需求,增大地图的信息表现力。

目前,与人物信息有关的数据模型主要包含两类,一类是以人物为主体的数据模型;另一类模型中是人物信息作为相关要素而存在。以人物为主体构建的模型主要有哈佛大学地理分析中心主导建成的中国历代人物传记资料库(China biographical database, CBDB)中的古代人物传记模型;在搜韵网项目中构建的以古代文学家为主体、包含人物著作与简要地理轨迹信息的人物信息模型;Filatova等[3]构建的以元事件理论为基础的人物传记摘要模型;Han等[4]利用网络本体语言(ontology web language,OWL)构建的人物事件本体模型;于满泉[5]提出的面向人物追踪和知识挖掘的人物模型等。

人物信息作为非主体要素的数据模型主要有:温永宁等[6]在家谱地理信息系统(geographic information system,GIS)研究中,创建了以人物家庭氏族关系和时空信息为核心的地理数据模型;周丙锋等[7]、胡迪等[8]、李凯等[9]在历史地理信息系统(historical geographic information system,HGIS)研究中,构建的以人物或人物年表为关键要素的历史事件数据模型等。

从研究的应用角度而言,CBDB、搜韵网、家谱GIS和HGIS中人物相关数据模型为满足相关人文研究的需求而建立,模型结构完整和格式规范,可作用于地理分析和地图生成,但数据主要依靠人工收集资料和手工编纂更新,极其耗费时间和人力,难以高效利用互联网上的大量泛在信息。文献[3-5]中的人物数据模型是自然语言处理(natural language processing ,NLP)领域内的研究成果,是为拓展信息提取技术应用而建立的,虽然自动化程度高,但模型结构单一,数据格式不规范,不便于精确检索和分析,更无法满足地理领域的应用需求。

因此,针对现有人物相关地理信息数据模型数据填充自动化程度低,而自然语言处理领域的人物数据模型结构单一的情况,本文提出了一种人物经历信息模型,该模型不但能够有效利用现

收稿日期:2020-05-07

项目资助:国家重点研发计划(2017YFB0503500);国家自然科学基金(41671407,41701457,41801317)。

第一作者:张三强,硕士,主要研究方向为作战环境数据工程。1390724098@qq.com

通讯作者:宋国民,博士,教授。ccllyy123456@163.com

有自然语言处理技术,达到人物信息自动提取的目的,而且突显了人物相关的时空信息,便于地理领域的研究应用。以人物百科数据为例,设计了相应的提取流程,对事件描述和位置信息的提取方法做了重点介绍,并进行了相关实验。

1　人物经历信息模型

1.1　人物经历信息的内涵与特点

所谓经历者,经久历远之意,是人物在时空域内活动的事件记录,蕴含着人物的主要时空信息,其本质是人物相关的事件集合,也有研究者将其称为人物事件[10]。与事件信息的内涵的区分在于经历信息面向的是特殊个人,记录的可能是大事件中人物个人行为,也可能是人物主导的完整事件;而事件表达要求完整性,一个事件可能包含若干子事件,强调起因、经过和结果全过程,事件中的人物是泛指,包括了个人、组织和群体[11]。

和当今技术条件下能够记录的连续轨迹数据相比,以事件为主体的时空轨迹记录存在着断续性、多粒度和交融性的特点。(1)断续性是指事件经历记录的时空轨迹并不连续,它自然地省去了人物在连续时空轨迹中的大量低价值冗余信息;(2)多粒度是指经历记录的时间和空间尺度不一,有的是人物阶段性的事迹,覆盖了一定的时空范围,有的则叙述了人物在时间点上的具体行为,时空信息聚焦于一点;(3)交融性是指人物经历的事件时空变换并非单一线性,多个事件的起止时间可能存在着重合、交汇或包含,空间上也可能不完全相关。

1.2　人物经历信息模型设计基本思路

本文模型的设计尝试满足两个需求:(1)要尽可能丰富信息维度,规范数据格式,突出人物经历中的时空特征;(2)模型中的数据能够自动提取和填充。这两方面存在着一定制约关系,模型越精细,意味着数据可供分析的角度层次越多,越便于应用,同时也意味着信息提取的要求越高,难度越大,现有手段可能无法有效解决。

1)时空要素表征。时间和地名信息的提取包含在命名体识别研究的内容之中,相关方法已极为成熟,但本文关注的空间信息并非简单的地名,而是泛指人物经历事件发生的广义位置信息,包含地名、隐含位置信息的机构设施名称、方位词、地理坐标等,所以需要一定的拓展研究(这里不涉及向具体坐标的映射)。同时根据前文分析的特点可知,人物经历存在阶段性描述,因此时间特征需由发生(起始)时间和结束时间两部分组成。

2)事件要素表述。人物经历的时空交融性特点可从事件类型层面进行有效区分,结合信息提取领域内的研究内容和概念定义,本文模型中的事件信息可由事件类型和事件描述两部分构成。参考 ACE2005(the 2005 automatic content extraction)对事件类型的区分,人物经历事件可标识为“个人生活”“社会行为”“行程游历”和“著作成就”,4 种标签可同时存在,并不完全独立。

3)人物属性表达。人物在语句中存在字号、笔名等多种形式的表达,考虑到同名同姓人物的存在,判断文本信息是否相关,必需有人物的其他特征支持,同时人物多维度的分析也需要相关知识。

4)数据源。现有自然语言处理技术在解决古文的信息提取和翻译上依然有很大难度,本文模型面向的是互联网开放现代汉语文本,其类型应包含新闻、官方简历、人物百科等数据,由于不同来源的文本形式不同,其内容重复或有所出入,有必要标注信息来源。

1.3　人物经历信息模型结构

依据前文所述的基本思路,人物 i 的经历信息模型可以表示为 $M_i=\{A_i,R_i,E_i\}$,其中 $A_i=\{a_{i1},a_{i2}\cdots a_{i33}\}$ 为人物 i 的主要属性,由人物 i 的编号、属性项和属性值 3 部分构成,包含出生地、职业、主要成就等 33 项共有属性;$R_i=\{r_{i1},r_{i2}\cdots r_{ij}\}$ 为该人物的人物关系信息,其中每条记录由主体人物的 ID、关系类型和客体人物姓名 3 段组成;$E_i=\{e_{i1},e_{i2}\cdots e_{ik}\}$ 为该人物的经历事件,是模型中的主体成分,事件信息中每条记录都由“事件起始时间”“结束时间”“事件发生位置”“事件类型”和“事件描述”等构成,如图 1 所示。

周丙锋等[7]在相关研究中,将事件信息的集合视为三维立方体,该立方体可通过时间、地点和人物三轴进行定位和组织。但实际上人物轴是单个独立的人物个体,并未构成完整维度。而本文的模型拓展了人物维度的结构,事件信息可以通过人物轴上多属性、多关系的聚类实现相关事件的聚合分析,同时也为获取特定类型人物的时空分布和轨迹提供数据支持。

2　人物百科网页的经历信息提取

人物经历信息模型中的数据需进一步通过

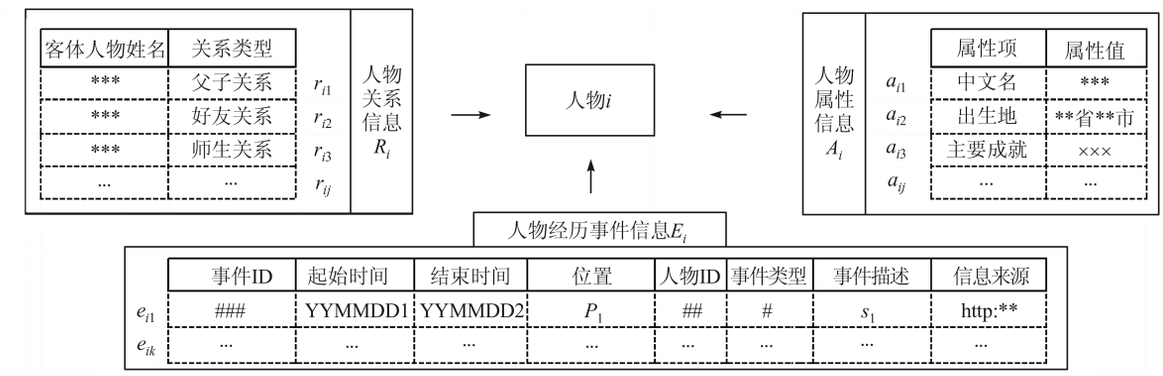

图1 人物经历信息概念模型

Fig.1 Concept Model of Character's Life-Track Information

自动提取的方式从文本中获取。目前,互联网百科信息是知识抽取相关研究中的重要数据来源,人物类的百科网页不仅包含基本信息框这类半结构化的数据,而且包含人物生平、履历和年谱等内容,事件描述类文本样式丰富,利于获取以人物为主体记载的事件信息。基于此,本文以人物百科网页作为研究的基础数据。

整体的提取流程如图2所示,分为人物属性、人物关系和人物经历事件的提取3部分。其中,人物属性和人物关系信息的提取在方法层面一致(为实体对和实体关系的识别),而模型中人物经历事件信息的填充是需要在获取事件描述的基础上逐步完成,涵盖了事件描述、时间信息和位置信息的提取以及事件类型判断。在这些子任务中,人物属性、人物关系和时间信息的提取及事件类型判断都有较成熟的方法,不再赘述。图2中,CRF(conditional random field)为条件随机场模型。

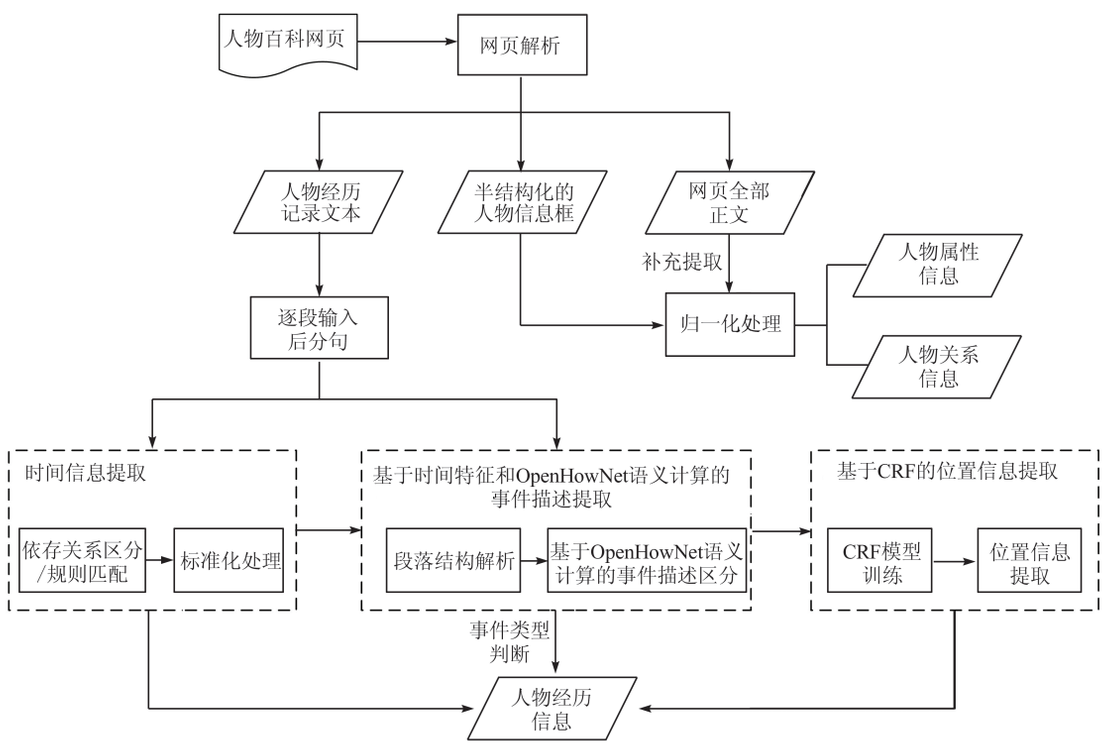

图2 经历信息提取流程

Fig.2 Processing of Life-Track Information Extraction

2.1 事件描述提取

在信息抽取领域,事件描述是事件信息提取的基础,通常为提取任务中给定的内容(以新闻为主),不存在"事件描述提取"的说法。但针对中文的百科类网页,需要将经历记录文本段落从网页正文中抽取出来,在此基础上区分为一条条

独立的事件描述，称之为事件描述提取。

人物经历的描述段落是所有人物词条网页中共有的目录内容（目录标题各不同），获取这些段落可以通过网页解析的方法来处理。通过网页解析，虽然得到的大量段落是只有一句话的简单事件描述，但仍有一部分段落语句构成复杂，包含了多个事件的记述，此时提取事件描述的关键就转化为如何区分这些段落中的语句。针对该问题，本文提出了基于时间特征和 OpenHowNet（通过义原概念来分辨中文词汇语义的开源工具）语义计算的事件描述区分方法。

2.1.1　基于时间特征的段落结构解析

当段落中存在多条人物经历事件记录时，不同事件的发生时间是易获取的，也是最容易区分事件描述的，借鉴抽取式文本摘要研究的做法，可将这些带有时间状语的语句标识为事件描述的关键句，与之相反的为非关键句。由于文本在叙事上要保持连续性，对应到句子结构当中，当段落中的非关键句之前或之后唯一方向存在关键句时，可确定它与邻近的关键句共同构成了一个完整的事件描述。但当非关键句处于两个关键句之间时，则无法判断该句为前一关键句的补充描述还是后一关键句的引导，该情况就需要从前后语句的语义关联度来判断。

2.1.2　基于 OpenHowNet 的语义关联度计算

语句中包含的实词相似度决定了语句的语义相似度，通常用于文本相似性判断或文本聚类[12]，这里可以作为语句前后关联程度的判断依据，若前后语句相关，则两句会出现相同词语或相似的内容表达。因此该步骤下的首要操作是通过分词和词性标注将语句中的实体词分离出来，包括动词、名词、形容词、数词和量词 5 类。记待计算的关键句为 S_1，非关键句为 S_2，对应的实体词集分别为 $W_1=\{w_{11},w_{12}\cdots w_{1m}\}$ 和 $W_2=\{w_{21},w_{22}\cdots w_{2n}\}$，则这两个语句的相似度矩阵为：

$$\boldsymbol{M}=\mathrm{sim}(W_1,W_2)=\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & & c_{2n} \\ \vdots & & & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{pmatrix} \tag{1}$$

式中，矩阵中的任一元素 c_{ij} 的值为词 w_{1i} 和词 w_{2j} 的相似度，由 OpenHowNet 计算求得。遍历该矩阵的列向量，提取出各列中的最大元素，生成 n 列 1 维向量 $\boldsymbol{H}$，意为从 W_1 中得到了与 W_2 语义最为接近的 n 个词及语义近似度。如果考虑到 W_2 中各个实体词在 S_2 语义构成中所充当的重要程度各不相同，则非关键句 S_2 关联于 S_1 的程度为：

$$\varepsilon=\sum_{i=1}^{n}\alpha_{\mathrm{pos}}(w_{2i})\alpha_{\mathrm{tf}}(w_{2i})h_i \tag{2}$$

式中，h_i 为 $\boldsymbol{H}$ 向量的每个元素；α_{pos} 和 α_{tf} 分别代表该词的词性特征权值和词频特征权值。本文认为句中实体词对于语句语义的影响程度是由词性和词频两方面决定的，实体词中动词和名词决定事件的要素信息较多，根据词性按照动词、名词、形容词、数词和量词的顺序，将权值依次定为 0.35、0.35、0.1、0.1 和 0.1。而词频特征权值等于该词出现的次数与语句中实词总数的比值。

综上所述，区分事件描述只需求得非关键句与前后关键句的关联程度 ε 的值，若值越大，则说明关联度越高，关联度较高的两句话即应当标识为同一事件描述。

2.2　位置信息提取

这里的位置信息是指人物经历的事件发生地点，事件发生地包含了古代地名、隐含位置信息的组织机构名、地址形式的表达和隐含式表达等多种复杂情况。常用 NLP 工具对复杂地名识别难有成效，更无法判断某个地名词是否为事件发生地。而现有开放的中文语料中，也没有针对事件发生地进行专门的标注，因此本文自行标注了相关训练语料，采取相应方法实现了提取。

文本中位置信息的出现与其前后文的词语及其性质有着紧密的联系，是 NLP 中典型的序列标注问题。目前，双向长短记忆模型＋CRF 模型是序列标注任务中主流且成熟的方法[13]，但循环神经网络的方法在处理小样本数据时（受人工标注条件所限，本文实验的样本数据量小于 10 万条）效果不明显，且不便于分析。鉴于此，本文选取了经典的 CRF 模型。

2.2.1　语料标注

语义位置信息词组与词组本身、词性等均存在一定关系，因此有词特征、词性特征、实体类型特征、依存关系特征和句中相对位置特征 5 种特征被标注。前 4 种特征均由斯坦福大学开发的自然语言处理工具 StanfordCoreNLP 处理得到，而相对位置特征则是该词组的词序与句子中总词数的比值。

值得注意的是，StanfordCoreNLP 共有 43 种依存关系标签，由于每个词可能同时具有一个或多个依存关系，若利用程序自动将词语的依存关系特征疏化，则会出现多类排列组合特征值选项，无法体现出某种标签的特有支撑关系，因此在标

注文件时,本文将每个词是否具有某种依存关系逐类标出,形成43列取值为0或1的特征行列式。

词组的位置信息标签项通过人工方式采用了表示词语的起始处(begin,B)、中间位置(in,I)、非特定词(out,O)和单独词(single,S)的标签体系来标注,标签值定为PB(位置信息短语的起始)、PI(位置信息短语的中间)、PE(位置信息的结尾)、S(单独的位置信息词语)和O(非位置信息)5种。

2.2.2 窗口及特征值选择

本文以中心词的前后各两个词作为特征选择窗口,除了语料标注的5类特征外,还将该词是否是句首或句尾词的特征,以及每个词组的第一个字和最后一个字提取出来作为特征之一。这是由于在中文的语言习惯中,部分地点名词的词尾会存在一些特殊的字,同时位置相关词组前面往往会出现特定的词组或单字[14]。

3 实验与结果分析

3.1 事件描述提取实验结果

为了验证方法的实用性,本文对1 158个人物的百度百科网页进行了事件描述的提取实验。通过网页解析,共获得记录人物经历的文本段落43 327段,生成经历事件描述47 303条,其中,直接通过时间特征和语句结构区分的段落2 478段,生成事件描述4 962条,正确率100%;用OpenHowNet语义计算区分的段落3 426段,生成经历事件描述7 402条,区分正确的为6 793条,正确率91.8%。而由于时间特征不足未实现事件描述区分的段落共有318段,占所有需要区分处理段落数量的5.1%。

本文方法在实验中虽然取得了较好的效果,但通过分析实验过程记录也发现存在如下问题:(1)OpenHowNet无法处理未登录词汇的语义计算,实验中对未登录词汇逐字进行了语义计算后求和,其结果存在较多问题;(2)句中实词太少时,语义相似度计算差别较小,判断语句相近程度易出错;(3)当段落中的时间表述太含糊识别不出时,无法有效使用该方法进行事件描述的区分。

3.2 位置信息提取实验结果

在事件描述提取的基础上,本文进行了位置信息的提取,实验随机选取并标注了10 640条经历事件描述,利用Scikit-Learn(Python环境下的机器学习工具)进行了样本训练,随机训练的样本数量为总语料数量的1/3,剩余2/3为测试样本。评价指标采用精准率P和召回率R,F_1值为综合评价指标,计算公式为:

$$F_1=\frac{2PR}{P+R} \tag{3}$$

位置信息提取结果如表1所示。从实验结果来看,在标注语料有限的情况下,本文方法表现出了一定的位置信息标注效果,但对比文献[15]和文献[16]地名提取实验结果中的F_1值(0.9左右),还有较大的差距。因此,本文利用ELI5工具包分析了训练后模型的发射矩阵权值,得到以下结论:

1)语料信息中部分特征项对提取位置信息作用不明显,与StanfordCoreNLP标注的误差和部分特征项设置不合理均有关系;

2)位置词组及其边界词本身的特征效果明显,如词组中的一些词素“学院”“村”“堂”等,位置词组前的特殊动词及介词“入”“到”“赴”等;

3)部分位置信息名词在句子中会充当主体或修饰语,同时有些机构名词并不隐含位置信息,这些情况极易导致识别错误,这是引起准确率不高的原因之一,也是导致命名实体类型特征效果不显著的主要原因;

4)训练语料不够,没有纳入重复多例的特殊位置词汇表达,这是位置信息提取整体准确率偏低的最主要原因。

表1 位置信息提取结果

Tab.1 Results of Location Information Extraction

类别	精准率(P)	召回率(R)	F_1值
PB	0.86	0.73	0.79
PI	0.78	0.73	0.75
PE	0.81	0.69	0.75
S	0.85	0.82	0.83

4 结 语

本文通过分析人物信息面向地理信息系统应用的需求,结合当前人物信息提取的研究现状,提出了人物经历信息模型,并以人物百科网页为基础数据,设计了提取流程,实现了人物经历信息中事件描述和时空要素的提取。实验结果表明,事件描述抽取方法虽然具有较强的实用性,但是在位置信息提取方面仍有待提高。

在后续研究中还需注意:(1)需要在该标注体系的基础上加大数据获取和标注的规模,采取更前沿的方法进行实验,如结合转换器模型的双向编码器表示模型等,以提高位置信息提取的准确

性;(2)需要在已获取的数据基础上,面向更广范围的文本类型,探索人物经历信息的提取方法。

参 考 文 献

[1] Lin Hui, Zhang Jie, Yang Ping, et al. Development on Spatially Integrated Humanities and Social Science[J]. *Geo-Information Science*, 2006, 8(2): 30-37 (林珲, 张捷, 杨萍, 等. 空间综合人文学与社会科学研究进展[J]. 地球信息科学, 2006, 8(2): 30-37)

[2] Li Fan. Application and Perspective of GIS in Research on Historical Geography and Cultural Geography[J]. *Geography and Geo-Information Science*, 2008, 24(1): 21-26 (李凡. GIS在历史、文化地理学研究中的应用及展望[J]. 地理与地理信息科学, 2008, 24(1): 21-26)

[3] Filatova E, Prager J. Tell me What You do and I'll Tell You What You Are: Learning Occupation-Related Activities for Biographies[C]// Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, 2005

[4] Han Y J, Park S Y, Park S B, et al. Reconstruction of People Information Based on an Event Ontology [C]// International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, 2007

[5] Yu Manquan. Research on Knowledge Mining in Person Tracking[D]. Beijing: University of Chinese Academy of Sciences, 2006 (于满泉. 面向人物追踪的知识挖掘研究[D]. 北京: 中国科学院研究生院, 2006)

[6] Wen Yongning, Lü Guonian, Chen Min, et al. Data Organization and System Architecture of Sino-Family-Tree GIS[J]. *Journal of Geo-Information Science*, 2010, 12(2): 2235-2241 (温永宁, 闾国年, 陈旻, 等. 华夏家谱GIS的数据组织与系统架构[J]. 地球信息科学学报,2010,12(2):2235-2241)

[7] Zhou Bingfeng, Zhou Wenye, Zhao Wenji. Study on Digital Application Platform of Historical Geography[J]. *Science of Surveying and Mapping*, 2008, 33(4): 199-202 (周丙锋, 周文业, 赵文吉. 中国历史地理数字化应用平台研究[J]. 测绘科学, 2008, 33(4): 199-202)

[8] Hu Di, Lü Guonian, Jiang Nan, et al. Historical GIS Data Model Under Geographic and Historical Perspectives[J]. *Journal of Geo-Information Science*, 2018, 20(6): 713-720 (胡迪, 闾国年, 江南, 等. 地理与历史双重视角下的历史GIS数据模型[J]. 地球信息科学学报, 2018, 20(6): 713-720)

[9] Li Kai, Wang Yanjun. Design and Realization of Historical Human Geographical Information System Based on WebGIS[J]. *Geospatial Information*, 2019, 17(3): 59-61 (李凯, 王艳军. 基于WebGIS的历史人文地理信息系统设计与实现[J]. 地理空间信息, 2019, 17(3): 59-61)

[10] Zhao Rui. Research on Biography Generation Based on Events of Character Roles[D]. Dalian: Dalian University of Technology, 2015 (赵锐. 基于人物角色事件的传记生成方法研究[D]. 大连: 大连理工大学, 2015)

[11] Wang Shuang. Research on Theories and Methods of Spatial-Temporal Narrative Visualization [D]. Zhengzhou: Information Engineering University , 2017 (王双. 时空叙事可视化理论与方法研究[D]. 郑州:信息工程大学, 2017)

[12] Jin Bo, Shi Yanjun, Teng Hongfei. Similarity Algorithm of Text Based on Semantic Understanding[J]. *Journal of Dalian University of Technology*, 2005, 45(2): 291-297 (金博, 史彦军, 滕弘飞. 基于语义理解的文本相似度算法[J]. 大连理工大学学报, 2005, 45(2): 291-297)

[13] Vikas Y, Steven B. A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models[C]// The 26th International Conference on Computational Linguistics, Santa Fe, USA, 2018

[14] Zhang Zhuyu, Ren Feiliang, Zhu Jingbo. A Comparative Study of Features on CRF-Based Chinese Named Entity Recognition[C]// The 4th China National Conference on Information Retrieval and Content Security, Beijing, China, 2008 (张祝玉,任飞亮,朱靖波. 基于条件随机场的中文命名实体识别特征比较研究[C]//第四届全国信息检索与内容安全学术会议,北京,2008)

[15] Wu Lun, Liu Lei, Li Haoran, et al. A Chinese Toponym Recognition Method Based on Conditional Random Field [J]. *Geomatics and Information Science of Wuhan University*, 2017, 42(2): 150-156 (邬伦, 刘磊, 李浩然, 等. 基于条件随机场的中文地名识别方法[J]. 武汉大学学报·信息科学版, 2017, 42(2): 150-156)

[16] Wei Yong, Li Hongfei, Hu Danlu, et al. A Method of Chinese Place Name Recognition Based on Composite Features [J]. *Geomatics and Information Science of Wuhan University*, 2018, 43(1): 17-23 (魏勇, 李鸿飞, 胡丹露, 等. 一种基于复合特征的中文地名识别方法[J]. 武汉大学学报·信息科学版, 2018, 43(1): 17-23)

Character Life-Track Information Model and Information Extraction Method

ZHANG Sanqiang[1,2] *SONG Guomin*[1] *JIA Fenli*[1] *CHEN Lingyu*[1]

1 Institute of Geographical Spatial Information, Information Engineering University, Zhengzhou 450001, China
2 Troops 69340, Yili 835000, China

Abstract: **Objectives**: In the field of human-related geographic information systems (GIS), the spatiotemporal analysis of character information has received increasingly more attention. It is important in that it helps GIS users to generate various thematic maps and achieve the visualization of human geographic content. For adaptation to the development direction of GIS intellectualization, it is of great significance to combine GIS requirements with natural language processing (NLP) methods and build a character information model. **Methods**: Firstly, we expound the research status of character information models in GIS and NLP and put forward the concept of life-track, which is mainly composed of a series of character event mentions. Secondly, considering the feasibility of the existing information extraction methods, a conceptual character life-track information model is determined. This model focuses on event information to highlight character spatiotemporal elements and also includes character attribute and relationship information. Finally, a complete information extraction process is designed for the model with online character encyclopedia pages as the data source. This paper focuses on two sub-tasks in the process: One is to use time features and OpenHowNet semantic calculations to identify event mentions, and the other is to use linguistics features and the conditional random field (CRF) model to extract location information. **Results**: Experiment results show that the method of event mention identification has an accuracy of 91.8%. Although the average F_1 value of location information extraction is only 78% under the condition of a limited labeling corpus, some valuable experimental conclusions have been obtained by analyzing the weight of the transmit matrix of the CRF model: (1) The location phrase and its adjacent words have obvious characteristic effects. (2) The dependency syntactic parsing and the relative position of the word in the sentence have little influence on the extraction. (3) The target of location information extraction is the place where the event occurred, but in a few cases, some location phrases are irrelevant to the location of the event. This is the main reason for the low accuracy. **Conclusions**: Combining GIS with NLP, intelligent GIS development will be promising. The character life-track information model provides an example of the large-scale use of ubiquitous internet information. Improving methods applied in the extraction process and applying those methods to more online text types are the focus of our team's subsequent research.

Key words: character life-track information; information extraction; event mention identification; location information extraction; conditional random field

First author: ZHANG Sanqiang, master, specializes in the operational environment data engineering. E-mail: 1390724098@qq.com

Corresponding author: SONG Guomin, PhD, professor. E-mail: ccllyy123456@163.com

Foundation support: The National Key Research and Development Program of China (2017YFB0503500); the National Natural Science Foundation of China (41671407, 41701457, 41801317).

引文格式: ZHANG Sanqiang, SONG Guomin, JIA Fenli, et al. Character Life-Track Information Model and Information Extraction Method [J]. Geomatics and Information Science of Wuhan University, 2022, 47(5): 700-706. DOI: 10.13203/j.whugis20190424 (张三强, 宋国民, 贾奋励, 等. 人物经历信息模型及其信息提取方法[J]. 武汉大学学报·信息科学版, 2022, 47(5): 700-706. DOI: 10.13203/j.whugis20190424)