基于公开履历数据的人物知识图谱构建*

沈科杰　黄焕婷　化柏林
（北京大学信息管理系 北京 100871）

摘要：【**目的**】基于公开履历信息，结合自然语言处理技术与知识图谱构建技术，自动化建立履历知识图谱，为传统研究提供新的视角和工具。【**应用背景**】自动抽取履历数据中的人物背景、职衔信息并构建任职经历和机构同事等关系，通过可视化呈现的方式为企事业单位的人才选拔、人事任免任务提供决策支持。【**方法**】爬虫获取履历数据后，使用BERT-BiLSTM-CRF模型进行实体识别，通过定义规则与融合外部领域知识构建实体间关系，并使用Neo4j图数据库实现实体及关系的存储与图谱可视化。【**结果**】BERT-BiLSTM-CRF模型在实体识别任务测试集上的准确率为84.85%。图谱囊括561位干部履历信息，包含3类共8 174个实体和5类共20 162条关系，能够支持多角度的查询与分析挖掘。【**结论**】构建的知识图谱发掘了履历文本间的内在关联，为基于履历数据的研究应用提供了一种新颖易用的方案，但暂缺乏精细化的实体对齐处理和机构实体之间统属关系的构建。

关键词：履历分析　知识图谱　实体识别　人物图谱
分类号：TP391　G353
DOI：10.11925/infotech.2096-3467.2021.0145
引用本文：沈科杰，黄焕婷，化柏林. 基于公开履历数据的人物知识图谱构建[J]. 数据分析与知识发现，2021，5（7）：81-90. (Shen Kejie, Huang Huanting, Hua Bolin. Constructing Knowledge Graph with Public Resumes[J]. Data Analysis and Knowledge Discovery, 2021, 5(7): 81-90.)

1　引　言

履历是一个人学业、事业经历简明而规范的记录，在我国各行各业都有着广泛的应用。在实践中，职工履历是用人单位进行人事任免决策的重要依据；在学界，也有许多社会科学领域的研究基于公开履历数据探求任职模式、人岗匹配等方面的规律[1-2]。可以说，履历数据是一类具有丰富应用价值的资源。

目前，这类数据资源大多是将履历文本集中存储，使用时检索匹配并逐一阅读提取相关信息，费时费力，也难以发现成长路径、共性特征、人物关系等关键信息，阻碍了进一步的知识发现，对决策的支撑比较有限。而现有履历数据数量大，同时存在格式不统一、表述欠规范等问题，也是履历资源发挥其效用的障碍。

针对上述问题，本文提出一个履历知识图谱构建方案。研究基于公开的履历数据，实现自动化实体抽取、消歧、关联，将分散的、非结构的履历文本以可视化、可交互的知识图谱形式呈现，挖掘蕴藏在履历中的社会网络关系，为基于履历数据的深层知识发现提供一个通用方案。

2　相关研究

履历分析是社会科学领域研究中常见的分析方法。既往研究表明履历与工作绩效[3]、职位晋升[4]等存在正相关关系，可见履历数据中蕴藏着有益的研究价值。但传统履历分析直接面向履历文本，需要

通讯作者（Corresponding author）：化柏林（Hua Bolin），ORCID：0000-0001-9248-6455，E-mail：huabolin@pku.edu.cn。
*本文系国家社会科学基金项目（项目编号：17BTQ066）的研究成果之一。
The work is supported by the National Social Science Fund of China (Grant No. 17BTQ066).

人工提取职位、机构等数据，工作量大但所涉范围有限，制约了履历分析的发展。近年来，得益于自然语言处理技术的进步，准确高效地处理履历文本的技术日益成熟。履历文本语句精炼，但信息复杂且归属歧义大，早期研究采用基于规则及字典的方式进行职衔属性抽取[5]，但规则设计及迁移泛化程度等因素对系统影响因素较大。随后，基于统计学习的实体抽取方法涌现，如谷楠楠等[6]构建了基于规则和隐含马尔可夫模型（Hidden Markov Model，HMM）方法的中文简历信息提取模型与基于支持向量机（Support Vector Machine，SVM）算法的岗位推荐模型，帮助企业快速筛选应聘者并提供适宜岗位。Dong 等[7]提出基于条件随机场（Conditional Random Field，CRF）的高校科研人员简历关键信息抽取方法。近年来，深度学习技术也被引入履历实体抽取任务中。目前，双向长短期记忆网络（Bidirectional Long-Short Term Memory，BiLSTM）及其相关模型的使用最为广泛。例如，祖石诚等[8]应用 BiLSTM 结合卷积神经网络（Convolutional Neural Networks，CNN）和 CRF 模型对简历文本进行实体识别；Gaur 等[9]也采用 CNN 与 BiLSTM 模型实现简历信息中教育机构及学位的自动抽取。深度学习模型在实体识别任务中表现出了优异的性能。

另外，如何组织履历信息以凸显其内在价值也值得重视。科学知识图谱是图情领域的重要研究工具，借助文献的元数据，可构建合著图谱，成为揭示学者合作关系、识别领域专家等深度分析的基础[10]。作为一种有效的知识组织方法，知识图谱逐渐被引入其他各领域的人物关系挖掘研究中。例如，杨海慈等[11]基于中国历代人物传记资料库数据，构建了宋代学术师承知识图谱，有利于史学研究者洞察历史人物之间的复杂关系；王晓萍等[12]基于干部履历表构建了企业成员关系网络图谱，并运用图分析实现人岗关系研判；He 等[13]基于开放政府数据构建贫困家庭的人物关系图谱，以推动精准扶贫。

总体而言，基于结构化数据的知识图谱构建在支撑深层次知识发现方面已取得一定进展，但尚无基于半结构化的履历文本数据构建的人物知识图谱项目，履历蕴藏的价值有待深入挖掘。基于目前的研究现状，本文结合履历信息分析技术与知识图谱技术，构建了一个基于公开履历数据的知识图谱，实现了自动抽取、呈现关联、灵活查询的目标，为相关研究者提供实用的知识发现方案与工具，挖掘履历资源的深层价值。

3　研究思路与系统方案

3.1　系统架构

网页中的履历文本数据常以一种以数据库驱动半结构化的方式来组织，因此，结合正则表达式及实体识别算法，依据 Web 页面结构的模式匹配实现人物的信息采集与抽取。基于该思路，本文提出一个自底向上的知识图谱构建框架，共分为数据获取、信息抽取、关系挖掘、存储和可视化 4 层，如图 1 所示。

在数据获取层，使用网络爬虫采集公开履历数据，通过人名、URL（Uniform Resource Locator）和人工检查的方式去重并保存为 JSON 格式文件。在信息抽取层，一方面，借助正则表达式抽取出履历数据中的人物姓名、籍贯、毕业学校、每一任职经历的起止时间信息；另一方面，通过 BERT-BiLSTM-CRF 深度学习算法抽取每一任职经历中的机构及所任职位。在关系抽取层，对抽取出的信息，确立人、机构、地点三类实体及其对应属性，针对特殊表述情况，对抽取的实体执行共指消歧操作。引入中国地理区划领域知识，在实体间建立表达某人籍贯为某地、某人毕业于某地、某人任职于某机构、某机构位于某地、某地属于某地的关系语义三元组，并存储为实体关联文档。最终在存储与可视化层，将三元组批量导入 Neo4j 图数据库，实现履历知识图谱的关联化存储和可视化展现，支持用户通过点击或 Cypher 查询语言的方式浏览与检索图谱。

3.2　基于 BERT-BiLSTM-CRF 模型的实体识别

由于任职信息中的任职机构、职务的表述方式多样、长短不一，并无统一规律可循，故基于正则表达式匹配的抽取方法在此并不适用，需要借助算法进行实体识别。本文采用 BERT-BiLSTM-CRF 模型识别任职机构、职务实体。BiLSTM 即双向长短期记忆网络，CRF 即条件随机场，两者的结合 BiLSTM-CRF 模型最先由百度提出并在序列标注任务上取得良好的结果[14]。BERT（Bidirectional Encoder Representations from Transformers）模型则是谷歌于

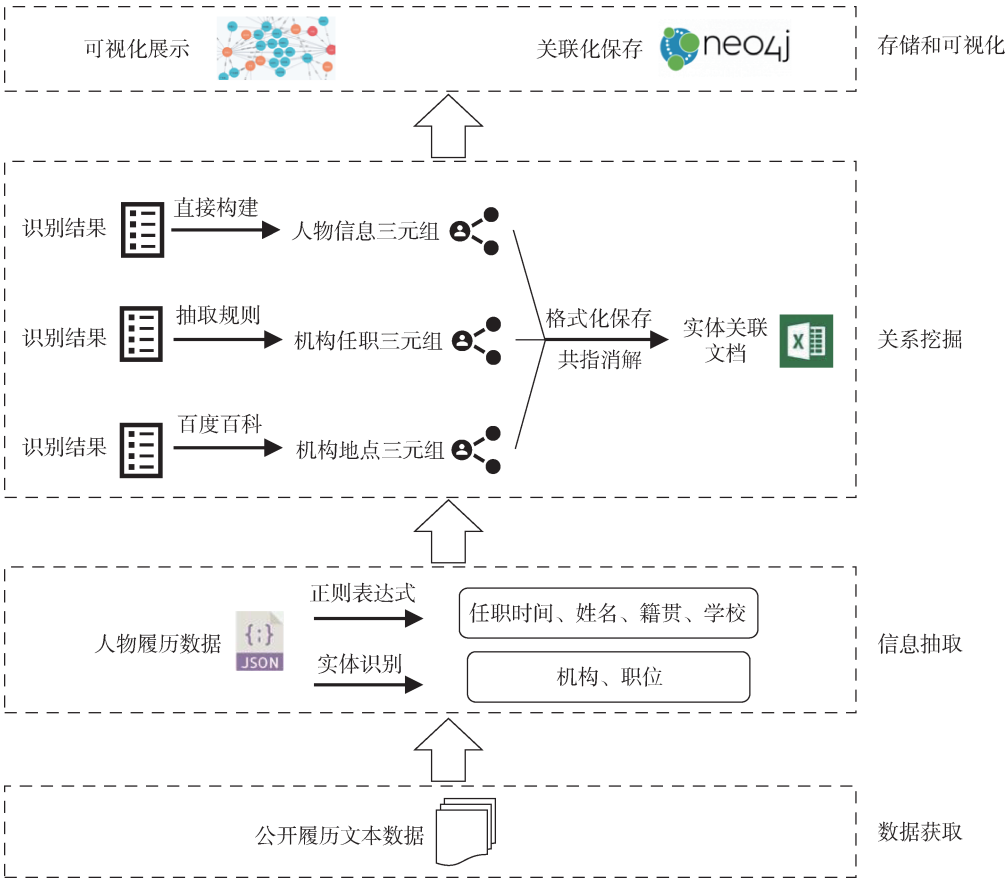

图 1　基于公开履历数据的人物知识图谱构建框架

Fig.1　The Framework of Resumes Knowledge Graph Construction Based on Open Resource Data

2018 年发布的自然语言处理模型，其强大的性能刷新了 11 项自然语言任务的记录[15]。将 BERT 与 BiLSTM-CRF 相结合的模型以提升命名实体识别的性能，成为近期研究的热点[16]。

BERT-BiLSTM-CRF 模型包括三个部分：BERT 预训练语言层、BiLSTM 层和 CRF 层，总体结构如图 2 所示。首先，BERT 预训练语言模型输入语料库，获取句子中每个字的字向量，然后将字向量序列输入 BiLSTM 模型进行特征提取。最后，由 BiLSTM 输出的特征向量由 CRF 模块解码并输出具有最高概率的标记序列。

BERT 预训练语言模型采用双向 Transformer 编码结构，可以得到一个字的上下文相关表示，能够表征字的多义性和句子的句法特征。Transformer 编码结构的核心单元是自注意力机制及前馈神经网络，单元可以连续堆叠。在自注意力机制中，每个输入的词都将转为三个不同的向量，分别是查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$、值向量 $\boldsymbol{V}$。与输入向量维度 $\boldsymbol{d}_k$ 结合，计算每个输入序列中不同单词和其他单词之间的相关程度，并根据相关性调整每个单词的重要性权重，如公式(1)所示。

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{\boldsymbol{d}_k}}\right)\boldsymbol{V} \quad (1)$$

BERT 的输入数据由字向量（Token Embedding）、位置向量（Position Embedding）、段向量（Segment Embedding）三部分构成，通过深层双向编码生成最终的字向量，输入到 BiLSTM 中进一步学习上下文特征。

BiLSTM 是模型的第二部分，能够进一步捕获句子的上下文特征从而获得更全面的语义信息。LSTM 克服了传统循环神经网络（Recurrent Neural

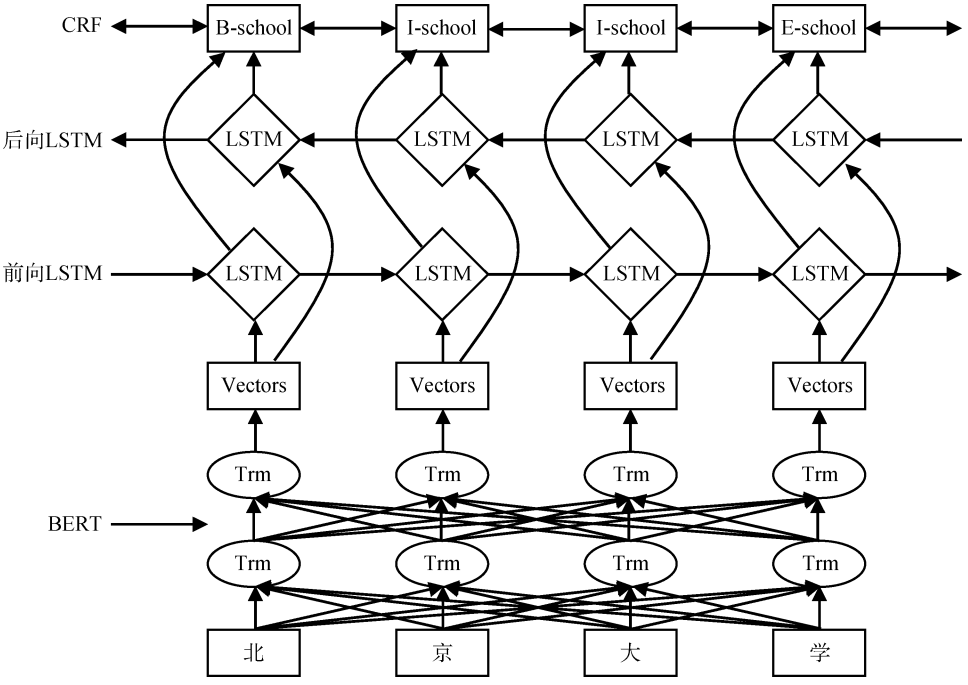

图2　BERT-BiLSTM-CRF模型示意图

Fig.2　The Architecture of BERT-BiLSTM-CRF Model

Network，RNN)模型在处理长序列数据时可能会发生梯度消失或爆炸的现象，由一个记忆单元和更新门、输出门、遗忘门三个门限机制控制记忆单元中信息的更新、传递和遗忘，从而使有用的信息经过较长的序列也能保存在记忆单元中。t时刻细胞的状态更新h_t可由公式(2)–公式(7)得到。

$$i_t = \sigma(W_i[h_{t-1};\ x_t] + b_i) \tag{2}$$

$$f_t = \sigma(W_f[h_{t-1};\ x_t] + b_f) \tag{3}$$

$$o_t = \sigma(W_o[h_{t-1};\ x_t] + b_o) \tag{4}$$

$$g_t = \tanh(W_c[h_{t-1};\ x_t] + b_c) \tag{5}$$

$$c_t = f_t c_{t-1} + i_t g_t \tag{6}$$

$$h_t = o_t \tanh(c_t) \tag{7}$$

其中，σ表示sigmoid激活函数；tanh表示双曲正切激活函数；W_i表示输入门权重；W_f表示遗忘门权重；W_o表示输出门权重；b_i、b_f、b_o分别为偏置项。

在应用中，LSTM只能获得前向的信息而无法有效利用后文信息，BiLSTM由正、反序两个LSTM构成，分别存储前文和后文的隐藏信息，然后计算两个隐藏单元的结果作为最终输出。

CRF是模型的最后一部分，即在给定一组输入随机变量条件下给出另一组输出随机变量的条件概率分布模型，考虑了标签之间的相互依赖性获得全局最优的标签序列。本文使用的是线性链条件随机场，可以用条件概率$P(Y|X)$表示。其中，X为输入变量，Y为预测序列，则有公式(8)。

$$P(Y|X) \propto \exp\left(\sum_{k=1}^{K} w_k f_k(y,x)\right) \tag{8}$$

其中，f表示特征函数；w表示特征函数对应的权重。在训练时，采用极大似然估计得到条件概率模型$P(Y|X)$。

4　系统设计与实现

4.1　履历数据获取与简单抽取

基于数据准确性与权威性的考虑，本文选用中国政要资料库[17]和地方党政领导人物库[18]中收录的干部履历作为研究示例数据。上述数据库分别由中国共产党网、中国经济网建立，收录了中央机构、国务院机关、各省市区重要机构现任领导干部的履历信息。

在数据获取层，首先使用Python中的requests、beautifulsoup包爬取上述资料库中的履历文本数据，

并使用URL与人名进行去重。共爬取到中央及地方干部的561条人物履历数据，保存为JSON格式文件。

资料库中的履历数据具有一定的结构化特征：首段一般遵从形如“XXX，性别，X族，XXXX年XX月生，AABB人，……，XX大学（学院、学校）XX专业毕业（学习、学历）……”的写作格式，故采用正则表达式匹配的方式即可快速抽取匹配人物姓名、籍贯及毕业院校；次段起为任职信息，一般遵从形如“XXXX年（XX月）-XXXX年（XX月）XX省委书记、代省长”的写作格式，同理可简单地使用空格及“年/月”字符切分抽取出任职的起始、结束时间。部分干部的履历信息没有提供任职时间段，则将该段任职期记为Unknown。

4.2　训练数据和模型参数设置

为获得质量较高的训练语料，本文爬取中文维基百科中58位中国干部的生平和履历文本数据，人工标注了1 571个机构、职位实体，后使用百度LAC（Lexical Analysis of Chinese）工具[19]、BIOES的标注模型进行命名实体识别标注（B代表一个实体的起始；I代表一个实体的内部；E代表一个实体的结束；S代表单字构成的实体；O代表文本中非所关注的实体部分）。标签类型包括姓名（Name）、地点（Address）、机构（Organization）、职位（Position）等4类。标注完毕后，共产生5 810条以句为单位的数据，分为4 841条训练数据、969条测试数据。

本文模型具体使用参数方面，BERT中文预训练向量采用谷歌开源数据，参数如下：BERT-Chinese一共12层，隐层为768维，采用12头模式，共110 M个参数。训练时，最大序列长度为60，rain_batch_size为64，其他参数均为默认值。BiLSTM层设置隐藏层维数为64。

4.3　基于规则的关系构建

在信息抽取模块中，从履历获取的人物、籍贯、毕业院校是一一对应的，可以直接构建“人-出生于-地点”“人-毕业于-机构”的语义关系。在实体识别任务中，同一个句子中会识别出多个实体，图谱构建需要生成“人-（以某职位）任职于-机构”三元组关系，但由于有些履历表述不够规范、机构合并、党政任职存在时间差等原因，句中机构、职位实体间并非一一对应，可能存在机构表述简略、一人在一机构任多职等情况。为了解决上述关系构建难题，本文定义了6条关系构建规则，应用到所有任职经历语句中，以期正确构建任职关系并修正模型的错误标注，如图3和表1所示。

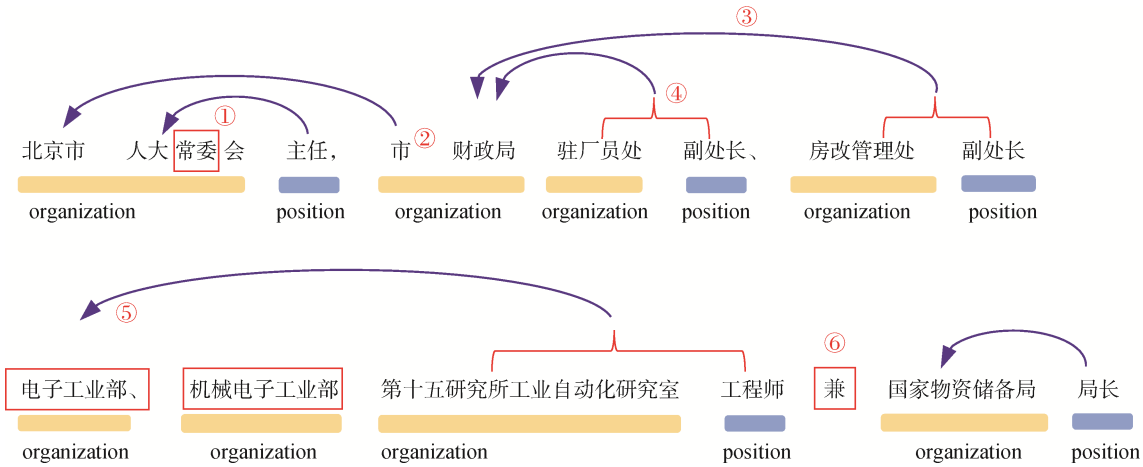

图3　关系构建规则示意图

Fig.3　Schematic Diagram of Relation Extraction Rules

除了机构任职语义外，曾于何地任职也是揭示人物职业轨迹的重要语义信息，如从任职于“北京市人大常委会”可知曾工作于北京。为了抽取机构属地的语义信息，参考中国行政区划，预先构建了包含省级、地级、县级三级行政单位的地名列表，利用该列表，由县级、地级再到省级的行政单位去匹配模型

表1　关系构建的6条规则

Table 1　Six Rules for Relationships Extraction

序号	规则	规则阐释
①	嵌入在机构标签内部的职位标签更改为机构标签	如“北京市人大常委会”中的“常委”会识别为职位，“常委”两字的标签修正为机构
②	机构地点指代消解	如“市财政局”中的市指代前半句“北京市人大常委会”中的“北京市”
③	机构职位一对多关系	机构数少于职位数的情形下采用职位向左最近匹配的方法构建职位与机构的关系
④	机构粒度处理	(1)在任职信息中，“、”后一般为职位信息，该职位与前句的机构有关； (2)“，”后一般为区别于前句的新机构任职信息； (3)抽取出以“，”分割分句下的第一个机构，该分句内其他的的机构定义为子机构，子机构与职位进行合并； (4)若识别出子机构且子机构内出现地点信息，如“中国银行|辽宁省分行”，则不对识别出的两机构切分处理
⑤	机构名变迁	如“电子工业部、机械电子工业部”存在机构名包含(变迁)情况，职位与位置靠前的机构进行配对
⑥	兼任职务处理	若“兼”字后面仍识别出机构，则需要在“兼”字处对句子进行切分，抽取出兼职所属机构的信息

所识别出的机构地理位置信息。对于未能成功匹配到地点的机构，首先尝试匹配百度百科中该机构词条中的地点。若机构尚无词条，则使用百度地图应用程序接口(API)查询获取该机构所属城市信息。若仍无法获取机构的属地信息，则用NaN进行标记。基于上述方法，可以构建出“某机构-位于-某地”的属地语义关系。

4.4　实体共指消解

现实履历文本中因抽取误差、不规范表述、机构变迁等主客观因素影响，同一实体或存在多个名称，需要进行共指消解规范实体表述。表2总结了履历实体抽取中涉及的多表述对应同一实体的情况及相应的共指消解操作。

4.5　存储与可视化

基于实体识别与关系构建的结果，以“(实体，关系，实体)”“(实体，属性，属性值)”两类语义三元组的方式组织所获取的履历信息。获取人、地点、机构

表2　多表述类型及解决方案

Table 2　Types of Coreference Resolution

类型	示例	操作
表述省略	部分机构实体有简称与全称多种表述，如“中国石油化工集团公司”在某履历中简写为“中国石化总公司”	若一机构名为另一机构名子字符串，剔除“集团”等停用词后且文本编辑距离在2以内，标记为同一实体并统一为全称
名称变更	在不同历史时期，同一机构实体使用不同名称，如“中国长江三峡集团公司”与“中国长江三峡工程开发总公司”为同一公司在不同时期的名称	若一机构名各字符顺序存在于另一机构名中，剔除“集团”等停用词后且文本编辑距离在5以内，化作字向量[20]并计算余弦相似度，大于0.9阈值则标记为同一实体并统一为时代靠后的名称表述
机构变迁	时代发展所导致的组织机构撤销、重组及调整现象，如“国土资源部“等部门重组为“自然资源部”	该情况实例数量较少，但难以自动化辨识。需人工借助外部知识更正名称表述

三类共8 174个实体，如表3所示；建立“出生于”“毕业于”“任职于”“位于”“属于”这5类共20 162条关系，如表4所示。将如上所述三元组以CSV格式存储，并批量导入Neo4j图数据库，最终生成可视化的知识图谱。

表3　实体及其属性描述

Table 3　Entities and Their Attribute Descriptions

实体名	属性	属性取值	数量
人	姓名	干部姓名	561
地点	地名	地点名，如“河北”	3 317
	等级	地点行政区域等级，如“省级”	
机构	机构名	机构称谓，如“河北省委”	4 296

5　结果与应用

5.1　模型性能及识别结果

为检验BERT-BiLSTM-CRF模型的有效性，采用准确率P、召回率R和F_1值三个指标对该模型针对不同类型实体识别的实验结果进行评价，评价指标的计算如公式(9)-公式(11)所示。

$$P=\frac{\text{正确识别的实体数目}}{\text{所有识别出的实体数目}}\times 100\% \tag{9}$$

$$R=\frac{\text{正确识别的实体数目}}{\text{所有标记的实体数目}}\times 100\% \tag{10}$$

表4　关系及其属性描述

Table 4　Description of Relationships and Their Attributes

关系名	关系语义	头实体	尾实体	属性	属性取值	数量
出生于	某人出生于某地，如某某出生于五峰县	人	地点	–	–	548
毕业于	某人毕业于某校，如某某毕业于北京大学	人	机构	–	–	515
任职于	某人任职于某机构，如某某任职于河北省委	人	机构	开始时间 结束时间	任期开始时间 任期结束时间	12 241
位于	某机构位于某地，如北京大学位于北京	机构	地点	–	–	3 544
属于	某地属于某地，如石家庄属于河北	地点	地点	–	–	3 314

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{11}$$

识别结果如表5所示。并将所有类型实体识别下的评价指标与BiLSTM-CRF模型IDCNN-CRF（Iterated Dilated Convolutional Neural Network-Conditional Random Field）两个模型进行对比，结果如表6所示。可以看到，对于各类型实体识别，该模型都具有较好的效果；且在三个模型中，BERT-BiLSTM-CRF模型的准确率P、召回率R和F_1值表现最佳且显著高于其他两个模型。

表5　BERT-BiLSTM-CRF模型各类实体识别结果评价

Table 5　Evaluation of Various Entity Recognition Results of BERT-BilSTM-CRF Model

实体类型	准确率/%	召回率/%	F_1值/%
地点	81.93	78.29	80.07
机构	78.84	81.8	80.29
职位	90.74	87.53	89.11
姓名	90.55	94.24	92.36

表6　各模型效果比较

Table 6　Model Performance

模型	准确率/%	召回率/%	F_1值/%
IDCNN-CRF	77.29	76.76	77.02
BiLSTM-CRF	78.86	76.91	77.87
BERT-BiLSTM-CRF	84.85	84.51	84.68

将训练好的模型应用到实际的履历文本数据中，共抽取出4 296个不同的机构实体和12 241个职位信息。实体识别总体效果较好，基本能够准确地分出机构和职位。相比于另外两个模型，BERT-BiLSTM-CRF模型在实体的起止位置判断上较准确。同时，观察抽取结果发现尚存在少量未被准确抽取的条目，总结如下。

（1）部分词语容易识别为单独机构。例如，“中共”“空军”等词会被标注为一个机构，忽视了后续文字，如“中共北京市委”“空军第一航空预备学校”的后半部分未被识别为机构，但同时“中共”“空军”两个机构也实际存在，且统辖前述两个机构。即此类错误情况的出现表明，由于上位机构的存在，会对下属机构、子机构的识别造成一定影响。

（2）机构具体的地理位置信息是部分机构识别的重要依据。例如，任职信息“省委610办公室综合处负责人”未说明具体省份，导致模型无法识别该机构。

（3）部分在“，”“、”等分句后较简略的机构名较可能也无法很好地被模型识别，如任职信息“市人大”，无法识别的原因可能与训练数据较少存在一定关系。

5.2　图谱应用

本文使用Neo4j图数据库作为履历知识图谱存储与可视化展示的工具，通过Neo4j交互式界面可以供用户轻松地对图谱进行探索，也可以通过Cypher查询语言对数据库进行检索。

以下示例中使用字母代号对涉及的干部姓名进行脱敏处理。如图4所示，使用Cypher语言查询名为Hu的干部的履历信息，并展开与其相连的节点所呈现的局部图谱。查询所用的Cypher语句为“match（n）--（m：人）where m.personID='Hu' return n，m；”图4中橙色节点代表人，蓝色节点代表机构，红色节点代表地点。由图谱可以清晰获知，Hu为五峰县人，与Liu1、Li1为宜昌同乡；毕业于北京大学，与Hou等为校友关系；在工作上，Hu先后在西藏、广东、内蒙古、北京等地任职，期间与Yang等存在同机构共事关系，其中与Lin有两地共事经历。基于对图谱的解读，可以知道Hu有着丰富的地方、中央任职经验，

且与其他干部存在合作经历，这可以作为后续人事调动的参考依据。

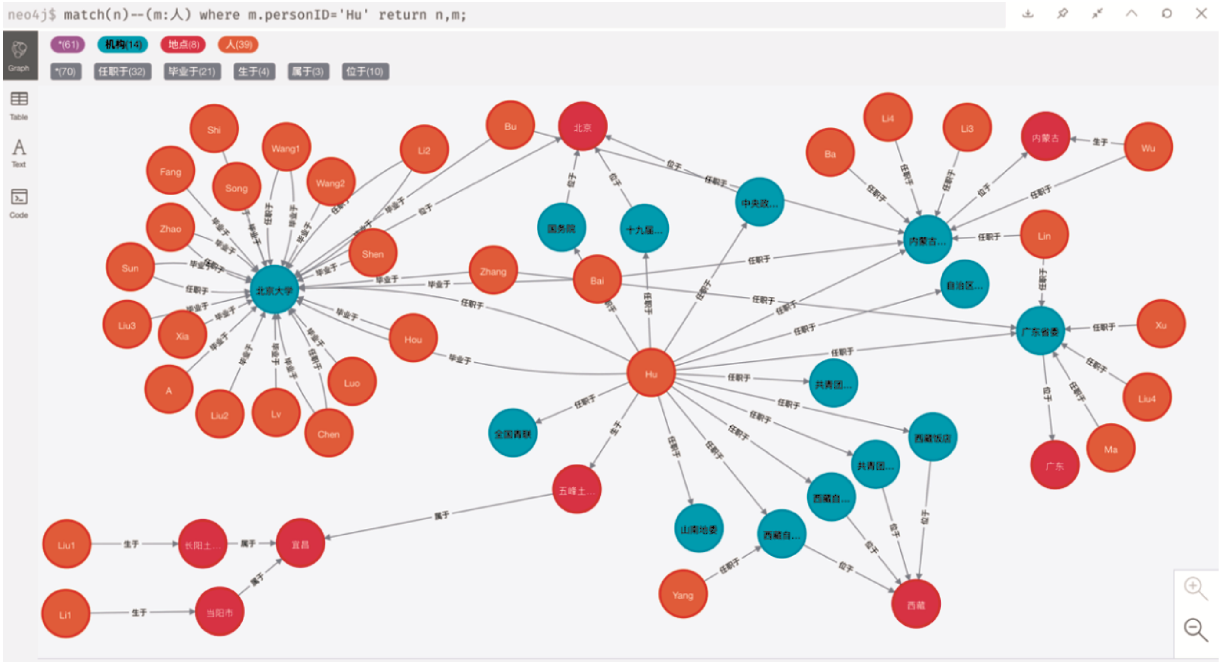

图4　知识图谱应用示例

Fig.4　Application Examples of Knowledge Graph

不仅如此，此履历知识图谱还支持其他角度的查询与探索，包括：

（1）以机构为核心的查询，如查询所有曾在广东省委这一机构有任职经历的干部，可探究共事关系、上下级关系、继任关系等；

（2）以地点为核心的查询，如查询同一时段内在广东省下辖各机构任职的所有干部，可探究同地任职的共事关系；

（3）以关系为核心的查询，如查询所有的“毕业于”关系，可探究各高校人才的培养情况等。

总体而言，本文设计的知识图谱具有开放灵活的特点，能够支持多角度的查询、分析与洞察。

6　结　语

本文基于公开的履历文本数据，构建了一个履历知识图谱，实现了履历分析技术与知识图谱技术的结合。同时，图谱以可视、可交互的方式呈现人物履历信息，弥补了现有履历资料库信息分散、无法查询、无法揭示深层关联的不足，也为其他基于履历数据的研究与决策提供了一个直观而有效的方案和工具。

同时，也需要看到，本文在以下方面尚有优化的空间：其一，现有机构实体之间相对独立，缺乏机构层级间关系描述，进而一定程度上制约了人物之间上下级等关系的揭示；其二，由于机构名称变更、履历写作不规范等因素，部分同一机构实体在不同履历中写作不同的机构名称，现有图谱仍无法完全精确识别其中关联，导致将其视为不同的机构实体。针对上述问题，未来仍可聚焦领域知识引入关系补全、实体对齐等任务，进一步完善基于履历数据的知识图谱构建工作。

参考文献：

[1]　田瑞强, 姚长青, 潘云涛, 等. 基于履历数据的海外华人高层次科技人才流动研究: 社会网络分析视角[J]. 图书情报工作, 2014, 58(19): 92-99.(Tian Ruiqiang, Yao Changqing, Pan Yuntao, et al. Using the Curriculum Vitae for Career Mobility Research of Chinese Overseas Highly-Talent: From the Perspective of Social Network Analysis[J]. Library and Information Service, 2014, 58

(19): 92-99.)

[2] 马秀玲, 饶帅. 少数民族地区基层公务员晋升的影响因素研究——基于县处级正职领导干部的履历分析[J]. 西北民族大学学报(哲学社会科学版), 2016(4): 53-63.(Ma Xiuling, Rao Shuai. On Influence Factor of Promotion of Basic Unit Public Servants in Ethnic Area——Case Study of CVs of County-level Principals [J]. Journal of Northwest Minzu University (Philosophy and Social Sciences), 2016(4): 53-63.)

[3] Hamman J A. Career Experience and Performing Effectively as Governor[J]. American Review of Public Administration, 2004, 34(2): 151-163.

[4] Sun J J, Cole M, Huang Z Y, et al. Chinese Leadership: Provincial Perspectives on Promotion and Performance[J]. Environment and Planning C: Politics and Space, 2018, 37(4): 750-772.

[5] 任宁. 大规模真实文本中的人物职衔信息提取研究[D]. 北京: 北京语言大学, 2008.(Ren Ning. Personal Position and Title Information Extraction in Large-Scale Real Texts[D]. Beijing: Beijing Language and Culture University, 2008.)

[6] 谷楠楠, 冯筠, 孙霞, 等. 中文简历自动解析及推荐算法[J]. 计算机工程与应用, 2017, 53(18): 141-148, 270.(Gu Nannan, Feng Yun, Sun Xia, et al. Chinese Resume Information Automatic Extraction and Recommendation Algorithm[J]. Computer Engineering and Applications, 2017, 53(18): 141-148, 270.)

[7] Dong F, Wang J N. Personal Information Extraction of the Teaching Staff Based on CRFs[C]//Proceedings of 2015 International Conference on Network & Information Systems for Computers. 2015: 615-617.

[8] 祖石诚, 王修来, 曹阳, 等. 基于新型文本块分割法的简历解析[J]. 计算机科学, 2020, 47(S1): 95-101.(Zu Shicheng, Wang Xiulai, Cao Yang, et al. Resume Parsing Based on Novel Text Block Segmentation Methodology[J]. Computer Science, 2020, 47(S1): 95-101.)

[9] Gaur B, Saluja G S, Sivakumar H B, et al. Semi-supervised Deep Learning Based Named Entity Recognition Model to Parse Education Section of Resumes[J]. Neural Computing and Applications, 2021, 33: 5705-5718.

[10] 曹烃. 体育科研论文合著状况分析——基于知识图谱的CSSCI文献计量分析[J]. 北京体育大学学报, 2012, 35(9): 49-54.(Cao Ting. Analysis on the Co-author Status of the Sports Scientific Research Thesis——A Study Based on the Knowledge Map of CSSCI Literature Metrological Analysis[J]. Journal of Beijing Sport University, 2012, 35(9): 49-54.)

[11] 杨海慈, 王军. 宋代学术师承知识图谱的构建与可视化[J]. 数据分析与知识发现, 2019, 3(6): 109-116.(Yang Haici, Wang Jun. Visualizing Knowledge Graph of Academic Inheritance in Song Dynasty[J]. Data Analysis and Knowledge Discovery, 2019, 3(6): 109-116.)

[12] 王晓萍, 郭梦洁, 岳婧雯. 基于关系图谱的人岗关系研究[J]. 大数据, 2020, 6(6): 129-139.(Wang Xiaoping, Guo Mengjie, Yue Jingwen. Research on Person-Position Relationship Based on Relation Graph[J]. Big Data Research, 2020, 6(6): 129-139.)

[13] He Y, Yun H Y, Lin L. The Character Relationship Mining Based on Knowledge Graph and Deep Learning[C]//Proceedings of the 5th International Conference on Big Data Computing and Communications (BIGCOM). 2019: 22-27.

[14] Huang Z H, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[OL]. arXiv Preprint, arXiv: 1508.01991.

[15] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [OL]. arXiv Preprint, arXiv: 1810.04805.

[16] 王子牛, 姜猛, 高建瓴, 等. 基于BERT的中文命名实体识别方法[J]. 计算机科学, 2019, 46(S2): 138-142.(Wang Ziniu, Jiang Meng, Gao Jianling, et al. Chinese Named Entity Recognition Method Based on BERT[J]. Computer Science, 2019, 46(S2): 138-142.)

[17] 中国政要资料库[EB/OL]. [2021-01-30]. http://cpc.people.com.cn/GB/64162/394696/index.html.(Database of Chinese Politicians [EB/OL]. [2021-01-30]. http://cpc. people. com. cn/GB/64162/394696/index.html.)

[18] 地方党政领导人物库[EB/OL]. [2021-01-30]. http://district.ce.cn/zt/rwk/index.shtml.(Database of Local Party and Government Leaders[EB/OL]. [2021-01-30]. http://district.ce.cn/zt/rwk/index.shtml.)

[19] Jiao Z Y, Sun S Q, Ke S. Chinese Lexical Analysis with Deep Bi-GRU-CRF Network[OL]. arXiv Preprint, arXiv: 1807.01882.

[20] Li S, Zhao Z, Hu R F, et al. Analogical Reasoning on Chinese Morphological and Semantic Relations[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 138-143.

作者贡献声明：

沈科杰：数据采集与清洗，实体识别及关系抽取模块研究，论文起草；
黄焕婷：设计构思，存储与可视化部分研究，论文起草；
化柏林：研究方案设计，论文修改和修订。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

支撑数据由作者自存储，https://github. com/ShenKJ/Using-open-resource-data-to-construct-a-resumes-knowledge-graph。
[1] 沈科杰, 黄焕婷. model-train.json. BERT-BiLSTM_CRF模型训练数据.

[2] 沈科杰, 黄焕婷. model-test.json. BERT-BiLSTM-CRF模型测试数据.

[3] 沈科杰, 黄焕婷. resume_data.json. 由中国政要资料库、地方党政领导人物库爬取的履历数据.

收稿日期:2021-02-11
收修改稿日期:2021-04-21

Constructing Knowledge Graph with Public Resumes

Shen Kejie Huang Huanting Hua Bolin
(Department of Information Management, Peking University, Beijing 100871, China)

Abstract: [**Objective**] This paper constructs knowledge graph based on the public resume data with natural language processing technology, which provides new tool for traditional data analysis. **[Context]** The proposed method could automatically extract profesional backgrounds and job information from resumes, and then obtain the relationship of working experience and colleagues in the organizations. The visualized knowledge graph could provide decision support for talent selection, personnel appointment and removal tasks of enterprises and institutions. [**Methods**] First, we used crawler to obtain the resume data and used the BERT-BiLSTM-CRF model to recognize entities. Then, we established the relationship between entities by defining rules and integrating the external domain knowledge. Finally, we used neo4j graph database to store and visualize data. [**Results**] The accuracy of the BERT-BiLSTM-CRF model with the entity recognition task was 84.85%. The constructed knowledge graph, which included resumes of 561 people, 8,174 entities in 3 categories, and 20,162 relationships in 5 categories, could support multi-angle queries and data mining. [**Conclusions**] This proposed model explores the internal relationships among resumes and provides a novel way to analyze resumes. However, there are few precise entity alignment processing and the establishment of relationships among institution entities.

Keywords: Rusume Analyse Knowledge Graph NER Characters Knowledge Graph