DOI:10.13203/j.whugis20170358 文章编号:1671-8860(2019)09-1416-07

基于地名树的最佳空间尺度新闻事件地点提取方法

舒时立[1] 李 锐[1,2] 吴华意[1,2]

1 武汉大学测绘遥感信息工程国家重点实验室,湖北 武汉,430079
2 地球空间信息技术协同创新中心,湖北 武汉,430079

摘 要:从新闻纯文本数据中识别地名信息并确定对应的最佳空间尺度与事件所属地点,是准确抽取新闻事件发生地点的关键。针对上述目标,提出了基于隶属关系地名树的最佳空间尺度新闻事件地点提取方法。在完成地名实体识别和歧义消除的文本数据预处理工作的基础上,提出了一种顾及新闻结构的方法消除语义干扰等噪声的影响;通过引入虚父节点构建合理准确的隶属关系地名树,结合最小包围盒的概念实现了最佳空间尺度的选取,使用地名实体权重和实体相关性完成了事件地候选集推荐排序,从而合理定位事件发生地。实验证明,所提出的新闻文本地理信息抽取方法可以较高的准确率获取新闻所对应的最佳空间尺度和相应的事件地点。讨论和解决了新闻文本地理信息抽取涉及的空间尺度问题,使得新闻文本中抽取的地理信息具有更好的可用性和可解释性,在丰富地理信息数据来源的同时,可实现数量呈几何级增长的网络新闻自动地域划分,有助于人们对各类事件空间态势的关注与认知。

关键词:网络新闻;地名实体识别;隶属关系;空间尺度;地理信息

中图分类号:P208 **文献标志码**:A

网络新闻是一种重要的信息载体,新闻事件的地点作为新闻的五要素之一[1],是多源地理信息的重要来源[2]。抽取新闻文本中蕴含的地理信息,在地理信息变化检测[3]、舆情监测分析[4]等众多方面都有着重要的应用价值。面对每日产生海量新闻数据的现状,由人工进行判读标注的方法显然不合时宜。因此,设计和实现对新闻中所包含的地理信息进行自动提取的算法,完成海量信息的快速处理,可以为人们分析理解海量的新闻信息提供支持,并丰富地理信息的来源和内容。

新闻一般以纯文本的形式存在,属于典型的非结构化数据,所包含的地理信息一般难以直接提取或挖掘,所以进行地名实体识别是需要开展的第一步工作。地名实体识别是命名体识别这一自然语言处理基础工作的重要组成部分,特指识别出文本中所包含地名的过程[5]。国内外在命名实体识别领域开展了大量研究,所采用的方法可以分为规则方法[6]、统计模型方法[7]和混合方法[8-10]3类。若要将地名实体转化为切实可用的地理信息,还需要考虑歧义消除问题。在地名歧义消除方面,学者们提出的基于概念密度[11]、空间密度[12]、上下文与空间关系[13-14]等方法已经取得了较好效果。

新闻文本地理信息的抽取必须考虑提取结果的可用性,而可用性体现在所抽取的地理信息是否准确以及所属的空间尺度是否符合实际情况和人的认知两个方面[15]。现有研究一方面多利用地名相关关系和文本的语义结构来过滤干扰信息并综合有效信息,从而提高抽取结果的可用性。如钟翔等[16]通过引入文献计量的有关理论,提出了基于链接分析的核心地名提取方法;於建峰等[17]将文本语义结构等规则引入到地名识别与规范化过程中。另一方面,也有学者利用地名在文档、篇章中的关联关系,使用信息熵最大理论处理尺度问题。如Rafiei等[18]认为,使地名在文档中具有最大非均质性的尺度为最佳尺度;张毅等[19]基于地理参考树和信息熵进行地理

收稿日期:2018-07-20

项目资助:国家重点研发计划(2016YFB0502301);国家自然科学基金(41771426);中央高校基本科研业务费专项资金。

第一作者:舒时立,硕士生,主要研究方向为空间数据挖掘。shilishu@whu.edu.cn

通讯作者:李锐,博士,教授。ruili@whu.edu.cn

求焦以解决尺度问题,但该研究在地名重要性程度衡量上未考虑文本结构,在空间尺度方面未对文本涉及的地理关联区域问题作进一步探讨。

在新闻事件发生地的抽取方法上,现有研究对新闻文本中存在较多语义干扰这一现象缺少关注。另外,为满足新闻事件发生地抽取的实际应用需求,除解决新闻事件最佳空间尺度问题外,还应提供地理关联区域与有关事件地点的可靠排序推荐,以供有关人员参考使用。

本文考虑了新闻事件发生地的空间尺度问题,在完成地名实体识别的基础上,引入虚节点和最小包围盒等理论使空间尺度具有自适应性,基于地名隶属关系树完成了对新闻文本的最佳空间尺度选取与事件发生地提取工作,并结合地名实体权重和实体相关性提供了事件地点候选集排序推荐,使抽取得到的事件地点更加合理且具有较高的可用性。

1　地名识别与歧义、噪声消除

新闻是一种文本数据,为利用这种非结构化数据,必须对原始数据进行合理转化。因此,对新闻文本中含有的地名实体信息进行识别与标注,是开展地理信息抽取的基础工作。考虑到新闻的特点与目前网络新闻按地点划归的现状,本文选取的最小空间尺度到县区级为止,而在该级别下仍然存在着较多重名现象,为不影响结果的准确性,需要开展地名歧义消除工作。另外,针对文本数据特有的语义干扰问题,对噪声信息进行去除也是必要的。本节阐述了文本地名实体识别、歧义消除和噪声剔除方法,准确可靠地从新闻文本数据中抽取地理信息。

1.1　文本地名实体识别

要在文本数据中识别抽取地理信息,首先要开展地名实体识别工作,本文基于角色标注的层叠隐马尔可夫模型(hidden Markov model, HMM)[9]来实现这一目标。角色标注层叠HMM的基本思路是:首先使用HMM完成词汇粗切分,在粗切分的基础上再建立一层HMM,同时依据地名特征建立角色表,并将该角色表作为状态变量取值空间,将粗切分词汇作为观测序列,再次应用HMM,从而完成对地名命名实体的识别。

HMM是一种著名的有向图模型,主要用于顺序数据的建模。在词汇粗切分任务中,新闻文本中各单字构词情况的求解问题可抽象为已知观测变量序列(分句),需要求取一个与该观测序列最佳匹配的状态变量序列(单字状态)的概率计算问题,即:

$$\underset{C}{\operatorname{argmax}}\ P(C|B) \tag{1}$$

式中,$C=\{y_1,y_2\cdots y_n\}$为各个单字状态变量序列;$B=\{x_1,x_2\cdots x_n\}$为新闻文本字串;argmax表示求使得$P(C|B)$取到最大值的序列C。由贝叶斯公式可得:

$$\underset{C}{\operatorname{argmax}}\ P(C|B)=\underset{C}{\operatorname{argmax}}\ \frac{P(B|C)P(C)}{P(B)} \tag{2}$$

依据标点符号可将新闻全文切分为多个文本序列,针对每一个文本序列有$P(B)$一致,故可得:

$$\underset{C}{\operatorname{argmax}}\ \frac{P(B|C)P(C)}{P(B)}\Leftrightarrow\underset{C}{\operatorname{argmax}}\ P(B|C)P(C) \tag{3}$$

又由马尔科夫链的定义可得:

$$P(B|C)P(C)=P(y_1)\prod_{i=1}^{n}P(x_i|y_i)\prod_{i=2}^{n}P(y_i|y_{i-1}) \tag{4}$$

式中,$P(x_i|y_i)$为状态观测概率,即单字x_i为状态y_i的概率;$P(y_i|y_{i-1})$为状态转移概率,即由上一单字状态得到当前单字状态的概率。这些参数均可由语料库统计得到。

在文本粗切分中,x_i代表观测变量,y_i代表状态变量,则有观测序列$\{x_1,x_2\cdots x_n\}$(字符串)和状态序列$\{y_1,y_2,y_3\}$($y_i\in${词首,词间,词尾}),故式(4)依据各状态链的联合概率可表示为:

$$P(y_1\cdots y_n|x_1\cdots x_n)=P(y_1)P(x_1|y_1)\prod_{i=2}^{n}P(y_i|y_{i-1})P(x_i|y_i) \tag{5}$$

联立式(1)和式(5)可求解粗切分的结果。将HMM进行第2次应用完成地名角色标注,即将观测变量变更为粗切分后的字词,状态变量变更为地名角色集合(如地名的首间尾以及后缀等),其余原理和方法与粗切分时相同。

完成地名实体识别后,保存各个地名信息与它们在新闻文本中出现的相对位置,用于后续计算各个地名在新闻文本中的权重。

1.2　地名歧义消除与标准化

容易知道,由地名同名造成的地名歧义会使地名树产生错误树结构,而在中国地名命名现状下,重名现象之多使得地名歧义消除工作显得尤为重要。

由于新闻全文是一个不能割裂的整体,涉及的地名之间一般存在着相关关系,故歧义地名的

正确指示项会得到上下文的更好支撑。基于上下文信息的支持,通过比较文本中不存在歧义的地名实体为歧义地名各个指示项提供的支持程度的大小,可以进行地名歧义消除工作。在实现上,可将文本中包含的其他不存在歧义的地名实体根据隶属关系与存在歧义地名的每一个指示项构成地名组,比较各组包含的地名数量,以获得上下文支持数量较多的指示项为正确指示项。

为识别歧义地名并准确构建地名组,本文引入地名数据库以判断某地名是否存在同名歧义问题,并确定与文本中其他地名实体之间的隶属关系。针对本文的研究对象,使用带有行政区划代码的省市区三级地名数据库作为查询依据。

遍历某篇新闻中提取到的所有地名实体,查询标准地名库,对存在重名现象的地名进行歧义消除,使用区划代码代替地名,从而实现标准化处理,保证所获取的地名指示具有准确性和唯一性。

1.3 基于新闻结构的语义噪声剔除方法

在描述新闻事件时,除了事件发生地,有时也会涉及到其他地名。如报道黄冈新开通的高铁车次的新闻中提到,"……经由武汉开往广州、北京……",此例中最佳事件地点应判定为黄冈,而若不去除干扰项,尺度会被判定为全国,这是不合理的。

本文采用给各个地名实体赋予相应权重并设定阈值的方式来剔除干扰项,而权重的赋予应顾及新闻结构。新闻在文本结构组织上具有特殊性,其普遍遵循的规则是标题-导语-正文3级展开[1],其包含的地名信息可靠程度依次减弱,但空间信息的精确度却可能依次递增。

一个地名通常会在整篇新闻报告中出现多次。对于某一地名,若认为它不在标题和导语中出现而在正文中出现的次数小于阈值k时为干扰项,则应依据新闻结构为出现在不同位置的同一地名赋予不同的权重大小来综合度量其是否为干扰项。根据上述定义与新闻结构的特点,将正文、导语和标题的权重分别设为W、kW和$kW+1$,则对于该地名,全文权重加和的计算公式为:

$$W_{\text{sum}}=n_0(kW+1)+n_1kW+n_2W \quad (6)$$

式中,n_0、n_1、n_2分别表示该地名在标题、导语、正文3个部分出现的次数。

针对在§1.2中完成地名标准化的集合,对具有相同行政区划代码的元素按照式(6)计算W_{sum}值,若满足$W_{\text{sum}}\geqslant kW$,则加入候选集,完成干扰项的剔除(本文实验中阈值$k$定为2,$W$定为单位值)。

2 地名树构建

地名树的构建过程为:首先依据标准地名数据库,基于地名之间的行政隶属关系生成基础地名树;其次通过为地名树添加虚拟父节点,使结构更新后的地名树的尺度信息发生变化,补充了隐含的认知信息,使地名树更具可解释性;最后结合新闻文本结构和权重传递方法,为隶属关系地名树的每一个节点赋予恰当的权值,完成各个地名节点的核心度评估。

2.1 包含虚节点的隶属关系地名树构建

在中国现行的行政区划下,每一级别下辖的地区数量是不固定的,由此得出,在数据结构的选择上,构建地名树所选用的数据结构以多叉树为好。

为更加清晰地叙述地名树的构建方法,首先介绍本文引入的虚拟父节点的概念。若存在大于等于两个节点且隶属于同一高级别的行政区划,而地名数据集中不包含该区划,则向该地名树添加对应的虚拟父节点。添加虚拟父节点的意义在于生成了高一级别的行政单位,补全了新闻中暗含的地理信息和尺度信息,同时具有尺度放大效应,而对多叉树进行递归遍历可以获取更精确的地名信息,又具有尺度缩小的功能,因此本文方法生成的地名树在空间尺度上就具有了可变性。由于本研究的目的在于划分各地新闻,故最大尺度至省级,不宜向全国级别扩展,必要时应生成多个树结构,并依据各树节点的空间关系构造事件的地理关联区域(见§3.1)。

举例说明虚拟节点的作用。消除歧义和去除噪声后,若某新闻中包含有武昌区、洪山区(均隶属于武汉市)和宜昌市(与武汉市同隶属于湖北省)3个地名实体,则由以上3个地名首先生成武汉市这个节点,然后自适应生成湖北省这个虚拟父节点。可以发现,这则新闻描述的是湖北省内两个地级市之间的互动,添加虚拟节点后,构建的地名树补充了隐含信息,更加完整合理,亦符合人们的认知,详细结构如图1所示。

由于存在生成虚拟父节点这个关键步骤,故地名树的生成应由子节点(较低行政级别)向上扩展。为减少查询次数,应将数据项按照行政级别高低提前进行升序排序。基于行政隶属关系的带有虚拟父节点的地名树生成算法描述如下:

将排序后的数据加入候选集,生成一个树的初始节点,并把该初始节点的值设置为候选集的第一个数据项,同时将对应数据项从候选集中去除,完成初始化,此时地名树仅含有一个节点;依次遍历候选集,同时遍历多叉树,通过将候选集元素与各节点的值进行判断比较,若为同级别同隶属数据项,则生成右兄弟节点,若隶属于数据项,则生成父节点;按照县区、地市顺序,依据区划代码查询是否存在两个及以上节点同隶属于高级别行政区划且没有对应的父节点存在,若有则加入虚拟父节点,更新树结构,并再次进行候选集元素的树节点添加操作;重复上一步骤,直至树结构稳定;查询候选集,若为空,则输出树结构,若不为空,则从第一步开始生成新树,直至候选集为空。

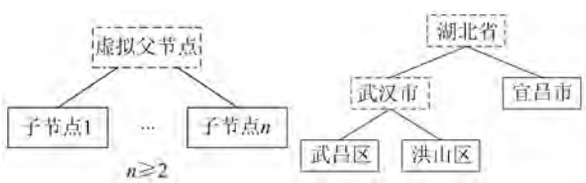

图 1　虚拟父节点生成示意图与实例
(虚线框为生成节点)

Fig.1　Sketch Map and Example of Virtual Parent Node Creation (Dashed Frames are New Nodes)

2.2　节点权重计算与虚拟节点权重获取

本文通过给节点赋予权重值来衡量地名节点的核心度。在权重赋予方法上,除考虑该节点地名在新闻文本中出现的次数之外,还需结合新闻的行文特点,使用其在新闻出现的位置加权求和计算。位置分为标题、导语和正文3种,它们的权重值依次降低。新闻标题在数据获取时通过HTML标签识别;导语的确定方法是以新闻正文的开头部分为搜索区域,以单个句子作为最小单元,选取同时具备时间、地点、主体对象要素(对应于中文命名实体对象)的单元作为目标结果。节点权重的计算按照式(6)给出。

本文引入的虚拟父节点实际并没有出现在新闻文本中,其权重由子节点传递而来,虚拟父节点的权重计算公式为:

$$P=\alpha\sum_{i=0}^{n}C_i \tag{7}$$

式中,C_i为某虚拟父节点P的孩子节点的权重;α为传递衰减参数。

父节点和子节点之间存在隶属关系,关系较为亲密,同时考虑子节点数量较多时(大于两个),父节点一般较子节点具有更高的核心度,故本文将α定为0.5,使得虚拟父节点的权重在子节点为2时取到子节点均值,且随子节点数增加而超过子节点权重。同理于地名树的构建过程,应按照行政级别由低到高的次序为虚拟父节点赋予权重。

3　尺度选取

本节探讨了新闻事件地理关联区域探测方法,并给出了最佳空间尺度选取算法和候选集排序推荐策略。

3.1　新闻事件地理关联区域探测

类比于空间自相关高值集聚的定义,若某新闻事件中存在两个或两个以上的地名满足以下条件:(1)地名权重值较大,(2)存在邻接或邻近关系,则认为这些地名在空间上构成一个地理关联区域,例如“武汉-长沙”构成了“中部城市群”地理关联。在新闻事件空间尺度的选择上,需要考虑地理区域关联现象。

地名权重值较大的衡量标准是,该地名的权重$W_i \geqslant kW+1$,即地名的核心度至少应达到标题地名的水平。结合实际数据,邻近和邻接关系针对于地市级以上计算,且邻近关系的判断方法为所属省级行政区划之间邻接。

由§2.1包含虚节点的隶属关系地名树的建立方法可得,应对生成了两棵及以上地名树的新闻报道进行地理关联区域探测。首先对各地名树筛选符合权重与级别标准的节点,然后将合格节点与其他地名树的合格节点进行空间关系判断,将符合关联条件的地名节点对联立,最后将包含同一节点的地名节点对进行组合,去除覆盖节点,形成最终的地理关联区域。区域的权重值由区域节点的权重加和得到,并参与到尺度选取和最终结果输出中。

3.2　尺度选择方法

在明确给出选取尺度的情况下,可以选取地名树的对应层级或由仅包含低行政级别的对应地名项行政代码生成对应级别的所属地名。仍以图1为例,例如某新闻网站按照所属城市显示本地新闻,则武汉地区可根据地市级尺度所包含的武汉和宜昌两个节点进行判断得到该则新闻与本地有关,继而将该则新闻划归到本地新闻栏目中。而本文所讨论的是未给出明确的尺度选取条件下,选取新闻文本对象所属的最佳尺度,这对于大量新闻文本空间信息的合理抽取和相关数据的可靠性分析无疑是有意义的。

新闻事件发生地的尺度选择策略遵循两条原则，一是信息完整性，二是空间精确性。对于获取到地理关联区域的新闻容易知道，该区域即是最佳空间尺度对应的信息抽取结果。针对更加普遍的情况，本文利用地名树在结构上产生的尺度可变性来选取合适的尺度。结合最小包围盒的概念，选取的地区级别应尽可能包含去除噪声信息后的所有涉及地名信息且不泛化，即尽可能精确地选取仅包含一个节点的行政层级作为最佳尺度，同时考察该层级节点的权重值，避免尺度过度放大的情况产生。

结合§2.2中对虚拟父节点的权重计算方式和§3.1中地名权重值的衡量标准得到，接受尺度放大节点的权重值 W_i 应满足 $W_i \geqslant 2\alpha(kW+1)$，容易证明，在该条件下，拥有两个以上重要子节点（$W_i \geqslant 2\alpha(kW+1)$）或3个及以上较重要子节点（$W_i \geqslant 3\alpha kW$）的虚拟父节点尺度会被选为最佳空间尺度。实现上，对地名树进行广度优先遍历，其停止的条件为找到接受尺度放大节点或该尺度节点权重大于所有子节点。例如，对应图1所示的情况，依据§2.2计算了各个节点的权重值，得到图2。

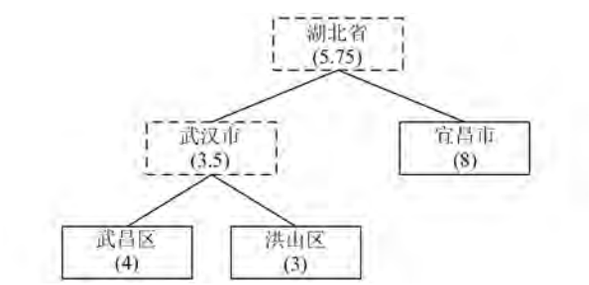

图2 赋权地名树

Fig.2 Weighted Toponymic Tree

初始化时最佳尺度为省级，以当前节点为基准进行条件判断，该节点的子节点数量为2，满足节点数判断条件，权重值为5.75，满足接受尺度放大条件并结束搜索。

3.3 候选集排序推荐

完成尺度选取后，最终结果以排序后的地名（地理关联区域）呈现，排序的多重条件优先次序为：地理关联区域＞尺度优＞权重值高＞行政级别低。各地名以区划代码为标准，标注所属行政级别信息，以满足查找和索引需求。

为满足地理舆情分析等新闻大数据的分析要求，在未给出明确的尺度选取标准和新闻数据应用目标时，以首项为事件发生地尺度选取和信息抽取结果参与数据分析。

仍以图2为例，依据尺度选取结果、节点权重值和地名树结构，得到排序结果为：湖北（省）-宜昌（市）-武昌（区）-武汉（市）-洪山（区）。参与新闻大数据分析的尺度和地理信息分别为省级和湖北省。

4 实验结果与分析

实验采用中国新闻网和搜狐网的社会与国内版块共1 000条新闻数据，包含字符数共370余万；含事件地理关联区域新闻100条，包含字符数40余万。该数据集是最佳尺度新闻事件地点提取方法的主要应用对象，具有较好的代表性。针对上述新闻数据集，统一使用本文提出的基于隶属关系的地名树结构的最佳尺度新闻事件地点抽取方法，对新闻事件发生地进行提取与验证。对于非地理关联区域的新闻数据集实验结果为：含地理信息数量872条，其中未提取27条，正确提取824条，错误提取21条；不含地理信息128条，未提取113条，提取15条。以查全率（R）、查准率（P）以及两者的调和平均值（F_1）作为评价指标来衡量本文方法的有效性，即：

$$R=\frac{\text{正确提取的数量}}{\text{含地理信息的总量}}=\frac{824}{872}\approx 94.5\%$$

$$P=\frac{\text{正确提取的数量}}{\text{提取总量}}=\frac{824}{824+21+15}\approx 95.8\%$$

$$F_1=\frac{2PR}{P+R}\approx 0.951$$

依据上述评价指标的计算结果可知，本文方法的查全率和查准率较高，且 F_1 值达到了0.951。对于包含地理关联区域提取的实验显示：正确提取数量94条，错误6条（区域错误5条），准确率为94%，能够准确合理地获取地理关联区域。

以上两项实验验证了本文提出的方法有较好的效果和较高的应用价值。针对实验中出现的错提和漏提两种情况，结合实际新闻数据进行分析，得出以下结论：

1）新闻事件发生地提取错误的情况主要由两种原因引起：（1）部分地名没有得到有效的识别或识别错误；（2）数据清洗工作仍待进一步提高。

2）对新闻事件发生地未提取情况进行分析可知，提取遗漏问题主要由3方面引起：（1）专用地名的识别；（2）著名机构名称的转化[20]；（3）细粒度地名的处理。

在新闻事件地理区域提取工作中，本文方法

仅给出了地名集合作为最终结果,没有进一步转化为便于人们认知的现行地理区块概念。

5 结　语

本文开展了地名实体识别、歧义消除和语义噪声剔除等工作,成功地将核心地名信息从新闻文本中剥离出来。提出了基于隶属关系的地名树结构的最佳空间尺度新闻事件地点抽取方法,通过引入虚拟父节点并基于最小包围盒概念,构建了新闻事件地理区域提取策略,使新闻事件空间尺度的选取具有自适应性,可以准确合理地从新闻文本中抽取新闻事件发生地。将该方法应用于实际新闻数据集,实验中得到的查全率与查准率较高,故具有较高的实际应用价值。另外可以推断,该方法在更加精细的空间尺度上依旧成立,具有较好的适用性。

目前,地名实体识别错误与遗漏、语义干扰等数据预处理阶段存在的问题是导致本文方法提取错误的主要原因。故在进一步研究中,可采取更加合理的地名实体识别算法与语义判断理解等措施,进一步提高准确率。另外,针对提取遗漏的情况,未来将考虑引入更加全面、精细的地名数据库与地名转换方法,从而更精确地完成新闻文本所包含的地理信息抽取工作。

参 考 文 献

[1] Li Liangrong. Introduction to Journalism [M]. Shanghai: Fudan University Press, 2013(李良荣. 新闻学概论[M]. 上海:复旦大学出版社, 2013)

[2] Sanderson M, Kohler J. Analyzing Geographic Queries [C]. International ACM SIGIR Conference, Sheffield, UK, 2004

[3] Ji Leijing. Semantic Change Detection of Geographic Information Based on Web Pages[D]. Nanjing: Nanjing Normal University, 2013(吉雷静. 面向网页文本的地理信息变化语义检测方法研究[D]. 南京:南京师范大学, 2013)

[4] Zhang Wei, Chen Xiaohui, Li Feng. Analysis of the Concept and Related Technologies of Geographical Public Opinion[J]. *Geospatial Information*, 2016, 14(3):5-6(张伟, 陈晓慧, 李锋. 浅析地理舆情的概念及相关技术[J]. 地理空间信息, 2016, 14(3):5-6)

[5] Tang Xuri, Chen Xiaohe, Zhang Xueying. Research on Toponym Resolution in Chinese Text[J]. *Geomatics and Information Science of Wuhan University*, 2010, 35(8):930-935(唐旭日, 陈小荷, 张雪英. 中文文本的地名解析方法研究[J]. 武汉大学学报·信息科学版, 2010, 35(8):930-935)

[6] Wu Y, Zhao J, Xu B. Chinese Named Entity Recognition Combining a Statistical Model with Human Knowledge[C]. ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition, Sapporo, Japan, 2003

[7] Sun J, Gao J, Zhang L, et al. Chinese Named Entity Identification Using Class-Based Language Model [C]. International Conference on Computational Linguistics, Taipei, China, 2002

[8] Zhang Xiaoyan, Wang Ting, Chen Huowang. A Mixed Statistical Model-Based Method for Chinese Named Entity Recognition[J]. *Computer Engineering & Science*, 2006, 28(6):135-139(张晓艳, 王挺, 陈火旺. 基于混合统计模型的汉语命名实体识别方法[J]. 计算机工程与科学, 2006, 28(6):135-139)

[9] Yu Hongkui, Zhang Huaping, Liu Qun, et al. Chinese Named Entity Identification Using Cascaded Hidden Markov Model [J]. *Journal on Communications*, 2006, 27(2):87-94(俞鸿魁, 张华平, 刘群, 等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006, 27(2):87-94)

[10] 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)

[11] Buscaldi D, Rosso P. A Conceptual Density-Based Approach for the Disambiguation of Toponyms[J]. *International Journal of Geographical Information Science*, 2008, 22(3):301-313

[12] Bensalem I, Kholladi M K. Toponym Disambiguation by Arborescent Relationships [J]. *Journal of Computer Science*, 2010, 6(6):653-659

[13] Wang Xingguang, Zhang Ruijie, Zhang Yi. Toponym Resolution Based on Geo-relevance and D-S Theory [J]. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 2017, 53(2):344-352(王星光, 张瑞洁, 张毅. 基于地理关联度和证据理论的地名消歧方法研究[J]. 北京大学学报(自然科学版), 2017, 53(2):344-352)

[14] Zhu Shaonan, Zhang Xueying, Li Ming, et al. Toponym Disambiguation Based on Administrative District Relation Tree[J]. *Geography and Geo-information Science*, 2013, 29(3):39-42(朱少楠, 张雪英, 李明, 等. 基于行政隶属关系树状图的地名消歧方法[J]. 地理与地理信息科学, 2013, 29(3):

39-42)

[15] Yu Li, Lu Feng, Zhang Hengcai. Extracting Geographic Information from Web Texts: Status and Development[J]. *Journal of Geo-information Science*, 2015, 17(2):127-134(余丽, 陆锋, 张恒才. 网络文本蕴涵地理信息抽取:研究进展与展望[J]. 地球信息科学学报, 2015, 17(2):127-134)

[16] Zhong Xiang, Gao Yong, Wu Lun. Extract Core Toponyms from Web Page Text Based on Link Analysis[J]. *Journal of Geo-information Science*, 2016, 18(4):435-442(钟翔, 高勇, 邬伦. 基于链接分析的网页文本核心地名提取方法[J]. 地球信息科学学报, 2016, 18(4):435-442)

[17] Yu Jianfeng, Wu Zhengsheng. Spatial Information Retrieval Based on Geographical Name Automatic Recognition in Text[J]. *Journal of Geomatics Science and Technology*, 2011, 28(3):227-230(於建峰, 吴正升. 文本地名自动识别的空间信息检索研究[J]. 测绘科学技术学报, 2011, 28(3):227-230)

[18] Rafiei J Y, Rafiei D. Geotagging Named Entities in News and Online Documents[C]. International Conference on Information and Knowledge Management, Indianapolis, USA, 2016

[19] Zhang Yi, Wang Xingguang, Chen Min, et al. A Semantics-Based Method for Extracting Geographic Scopes of Texts[J]. *Chinese High Technology Letters*, 2012, 22(2):165-170(张毅, 王星光, 陈敏, 等. 基于语义的文本地理范围提取方法[J]. 高技术通讯, 2012, 22(2):165-170)

[20] Zhang Chunju, Zhang Xueying, Ji Leijing, et al. Relation Mapping Between Generic Terms of Place Names and Geographical Feature Types[J]. *Geomatics and Information Science of Wuhan University*, 2011, 36(7):857-861(张春菊, 张雪英, 吉蕾静, 等. 地名通名与地理要素类型的关系映射[J]. 武汉大学学报·信息科学版, 2011, 36(7):857-861)

Extraction of News Location with Best Spatial Scale Based on Toponymic Tree

SHU Shili[1] *LI Rui*[1,2] *WU Huayi*[1,2]

1 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
2 Collaboration Innovation Center of Geospatial Technology, Wuhan 430079, China

Abstract: Online news provides users with access to current affairs in a timelier manner. As one of the key elements of news, news location also plays an important role in multi-source geographic information. The key technologies to extract news location from text data include the recognizing of toponymic information from text and the determination of the best spatial scale accordingly. To achieve these goals, we present an approach to extract news location with best spatial scale based on administrative district relation tree. A method is put forward to remove semantic interference on the basis of identifying place names to eliminate ambiguities. By introducing virtual parent node, the minimum bounding box, node weight and association relationship, it is possible to build toponymic trees accurately and select news locations with reasonable scale. The experiment shows that the method we proposed can get suitable scale and accurate location of news. In this paper, we discuss and solve the problem of spatial scale involved in the extraction of geographic information from news text, making geographical information more useful and interpretable. Apart from enriching geographical information, the method also makes sense in helping people form spatial awareness for all kinds of events, and helps tag exponentially-increasing online news by region.

Key words: online news; geographical names recognition; affiliation relationship; spatial scale; geographic information

First author: SHU Shili, postgraduate, specializes in geospatial data mining. E-mail:shilishu@whu.edu.cn

Corresponding author: LI Rui, PhD, professor. E-mail: ruili@whu.edu.cn

Foundation support: The National Key Research and Development Program of China, No.2016YFB0502301; the National Natural Science Foundation of China, No.41771426; the Fundamental Research Funds for the Central Universities.