面向网页文本的地理要素变化检测

王 曙[1],吉雷静[2],张雪英[2*],赵仁亮[3],陈晓丹[2],余 浩[4]

(1. 英国利兹大学地理学院,利兹 LS2 9JT; 2. 南京师范大学虚拟地理环境教育部重点实验室,南京 210046;
3. 中国国家基础地理信息中心,北京 100830; 4. 南京邮电大学计算机学院,南京 210003)

摘要:地理要素变化检测已成为国家地理信息"十二五"规划和全国地理国情普查的重要组成部分。网页文本中蕴含海量的地理要素信息,尤其是新闻、政府、社交平台等网站的网页文本更新频繁,可为地理要素变化检测提供现势性的数据源。本文针对网页文本中地理要素变化的语言描述特点,构建了表达地理要素变化的语义知识库,设计了搜索引擎和通用主题相结合的网页爬虫,实现了相关网页文本的高效获取;采用规则模型和条件随机场模型,分别进行网页文本中地理要素变化信息抽取,包括地理要素名称、位置(地名)、时间和属性等。实验结果显示,本文设计的网页爬虫具有较高的相关网页文本获取能力,地理要素变化信息抽取的准确率能够达到70%以上,但是,语义知识库的完备程度对于信息抽取性能具有较大影响。研究成果表明,以网页文本为数据源的地理要素变化信息获取方法,能提供一种快速检测地理要素变化的新途径,与实地调绘和遥感影像检测等方法结合应用具有较好的优势互补性,可作为有力的辅助手段解决地理要素的持续更新和实时更新问题。

关键词:网页文本;地理要素变化;信息抽取;网页爬虫;文本解析

DOI:10.3724/SP.J.1047.2013.00625

1 引言

地理要素对地物现状描述的准确性和时效性直接影响地理信息服务质量。地理信息的核心就是数据,而数据的生命力在于数据更新保障,具体体现在数据的现势性、准确性和共享性等方面[1]。目前,地理要素变化检测主要采用遥感影像检测和实地调绘两种手段[2-3],但是,通过遥感影像不能进行定量的属性信息检测[4-5],而实地调绘则存在工作量大、周期长、费用高等问题[6]。近年来,地理要素更新检测成为制约地理信息共享、服务和应用的重要瓶颈[7]。国家测绘地理信息科技发展"十二五"规划和首次全国地理国情普查纲要,明确将地理要素变化检测作为国家地理信息建设的核心组成部分[8]。

互联网在信息时效性方面具有显著优势,在信息传播方面扮演着越来越重要的角色,各行各业都设法将网络变成其信息的发布平台和汇聚地。调查研究表明,未来的人机交互95%将采用文本语言,且80%以上的文本中涉及地理要素描述[9]。网络文本信息具有时效性好、地理信息丰富和更新不受时间地点限制等优点[10],逐步成为新型的、重要的地理信息数据源。然而,网络文本信息存在语言描述结构复杂,质量水平参次不齐等问题,使得网页文本中地理要素信息获取和利用,具有相当高的难度和挑战性[11-12]。

2 网页文本地理信息检测应用分析

2.1 自然语言中地理信息解析

近年来,语言学、人工智能和地理信息系统等领域对自然语言中地理信息解析方法进行了较为深入研究,主要包括时间、空间位置(地名、空间关系)和地理属性等。其中,时间和地名属于实体识

收稿日期:2013-06-07; **修回日期**:2013-07-05.

基金项目:国家测绘科技项目"网络地理信息变化检测技术研究";国家自然科学基金项目(40971231);"863"计划项目(2007AA12Z221)。

作者简介:王 曙(1989-),男,山西人,硕士生,主要从事地理信息智能数据处理、地理信息系统等方面研究。
E-mail:shuwang8951@hotmail.com

***通讯作者**:张雪英(1970-),女,四川人,博士,教授,主要从事地理信息系统、地理信息智能处理和服务等方面研究。
E-mail:zhangsnowy@163.com

别，主要采用规则（基于词典和句法模式库）和机器学习（条件随机场和隐马尔可夫等模型）两种方法；空间关系和地理属性属于关系信息抽取，主要采用规则方法。

（1）时间识别：文献[13]通过制定时间表达式规则集，研发了CTEMP时间分析器，实现了基于规则的中文时间短语抽取和归一化；文献[14]借助文本标注手段，将时间短语分为一般时间短语和事件时间短语，针对不同时间短语类型采用不同识别方法，时间信息抽取的准确率和召回率达到了90%。近年来，条件随机场的时间识别方法得到了众多学者们的青睐，有效减轻了繁重的规则库构建工作，而且可移植性较好，但是，其准确率和召回率低于规则方法[15-16]。

（2）地名识别：地名词典匹配是最早使用的、比较简单的地名识别方法，但是，在较大程度依赖地名词典的数据质量，而且不能识别未登录地名，实际应用中匹配性能较低。为此，隐马尔可夫[17]、最大熵[18]、条件随机场[19]、支持向量机等机器学习模型逐步引入地名识别研究。虽然在一定程度上可以提高准确率，然而往往需要大规模的标注语料库。研究表明，语言知识库和混合机器学习模型，都是提高地名识别性能的有效手段[20-21]。

（3）空间关系抽取：早期美国国家地理信息与分析中心（NCGIA）就对自然语言中的空间关系语言进行了系统研究[22-23]，主要应用词性标注和句法分析工具实现文本中空间关系抽取[24]。近年来，国内针对中文文本中空间关系信息抽取方法进行了较为系统的研究，特别是空间关系描述词汇和句法模式，以及规则的空间关系抽取方法[25-27]。在此基础上，文献[28]进一步探讨了支持向量机的空间关系抽取模型，并与规则方法进行了比较。

（4）地理属性抽取：由于文本中属性信息描述具有较强的规律性，包括属性名称和属性值，因此，属性抽取主要采用关键词库和规则库相结合的方法[29-30]。为了克服人工建立规则库的技术瓶颈，也有学者探讨了条件随机场模型在属性抽取中的应用方法[31]。一般来讲，属性信息抽取主要针对“属性名称-属性值”对。地理信息描述更加复杂，可以概括为“地理实体-属性名称-属性值”的三元组形式。但是，研究方法仍然比较类似，包括规则方法[32]和机器学习方法，或者两者相结合的方法[33]。

2.2 面向文本的地理要素变化检测

21世纪初期，人们已经开始利用专业单位、社会力量和新闻等途径收集文献资料，通过人工阅读方式获取地物变化信息[34]。随着各种媒体的发展，报纸、网络等逐步成为地物变化信息发现的一种新手段。特别是近年来，随着互联网的日益普及和智能信息检索技术的发展，人们开始关注从互联网中获取地物变化信息。文献[35]提出利用搜索引擎关键词定制模式（地理要素名+表示变化的动词），实现与地物变化信息描述相关的新闻网页获取和可信度排序，并以杭州地区为例，开发了Web的地物变化检测系统，为区域的地物变化检测提供了新方法。文献[36]认为，网页文本中地理要素变化包括3个层次：地理位置（具有变化发生的地点）、变化主题（发生变化的地物）和变化情况（具体发生变化的内容）。通过在.Net平台上开发的原型系统，初步实现了省市县级政府网站内的地理要素变化相关网页获取和HTML解析。上述研究表明，从网页文本中获取地理要素变化信息具有一定的可行性，而且具有较强的实际应用需求。然而，相关研究尚处于初步探索阶段，主要针对特定政府网站和新闻网页的关键词搜索和可信度计算，缺乏对网页文本中相关语义信息的深层次挖掘。目前，文本中地理信息解析方法已经取得了较为先进的研究成果，为面向网页文本的地理要素变化信息检测提供了有力的理论和技术支撑。

3 网页文本中地理要素变化检测

3.1 地理信息检测方法与技术流程

在借鉴自然语言处理和地理信息解析技术的基础上，本文提出了面向网页文本的地理要素变化信息检测方法。具体技术流程如图1所示，主要包括以下几个步骤：

（1）语义知识库构建：地理要素的语义表达内容繁多、结构多样并且关联复杂，为了在网页文本获取和地理要素变化信息抽取时，提供主题语义约束和关联依据，针对网页文本中描述地理要素变化的语言特点，归纳总结相关信息表达的词汇和语义关系，并利用本体工具对知识库进行构建、管理、操作和维护。

（2）网页文本获取与解析：网页文本作为地理要素变化检测的重要数据源，其数据量巨大，语义

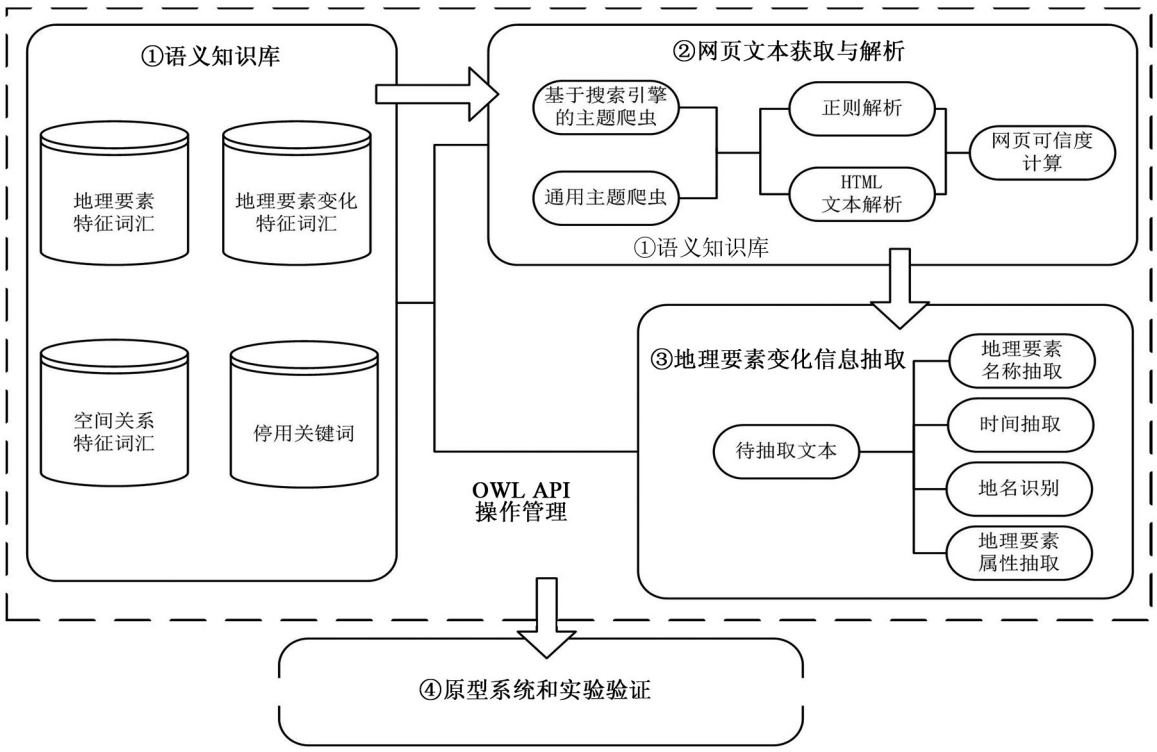

图1　技术流程图

Fig.1　The technique flowchart

表述复杂,为地理要素变化检测提出了巨大挑战。只有将网页文本进行规律性获取并解译,才能从HTML中得到所需的文本信息。因此,需要利用网络爬虫、HTML解析和可信度评价模型,有效解决相关网页获取问题。

(3)变化信息抽取:由于中文语义表述灵活,从文本中抽取地理要素变化信息相当困难。因此,需要在分别实现变化地理要素名称、时间、位置和属性等信息抽取的基础上,进一步实现地理要素名称与其他信息之间的关联关系。信息抽取将根据不同的信息类型分别采用规则模型和机器学习模型。

(4)实验验证:在上述研究基础上,设计相应的原型系统,并对相关模型和算法进行实验验证。

3.2　语义知识库构建

语义知识库构建包括2个步骤:(1)对知识库中的词汇进行分类总结,构建词汇库内部和词汇库之间的知识关系;(2)基于本体的思想创建各类词汇库,利用Protégé 4.2平台上实现知识库的构建、操作、使用、管理和维护。

3.2.1　领域特征词汇库及其语义关系

网页文本中地理要素变化描述需要回答4个核心问题:“什么地理要素发生变化?”;“地理要素发生了什么类型的变化?”;“地理要素变化何时发生?”;“地理要素变化在什么位置”。相应地,这些描述可以归纳为4类信息单元:发生变化的地理要素(通过名称识别)、变化类型(描述变化的动词)、时间、空间位置(通过地名和空间关系表达)。因此,语义知识库需要对表达它们的语言特征进行概括和总结,具体包括地理要素特征词汇、地理要素变化词汇、空间关系词汇和停用关键词。由于搜索结果中往往存在大量不相关网页,对这些网页的剔除亦可以从反面识别出地理要素的变化信息,这些不相关的网页信息可概括为停用关键词。

(1)地理要素特征词汇:根据“GBT 13923-2006地理信息要素分类与代码”的分类标准,将地理要素划分为4个大类,其中,特征词汇为一级词汇,共125个;典型要素为二级词汇,共552例。表1列举了部分特征词汇和其对应的典型要素。

(2)地理要素变化词汇:一般以中文中的动词形式表示,通过网页获取、统计和人工归纳方法,文中共总结要素变化动词91个。从事物发展角度,将划分为新生、变化和消亡3种类型,并将要素变化特

表1 地理要素特征词汇示例

Tab.1 The sample vocabulary of geographical information features

要素类型	特征词汇	典型要素
水系	河	河段、时令河、干涸河、河床、干河、河道、河流
	湖	湖泊、常年湖、时令湖、干涸湖
居民地及设施	站	发电站、水站、污水处理站、处理站、抽水站、观测站、气象站、水文站、监测站、地面站、试验站、实验站
	场	挖掘场、盐场、饲养场、养殖场、打谷场、贮草场、游乐场、剧场、体育场、球场、游泳场、垃圾场、殡葬场
交通	桥	车行桥、单层桥、双层桥、并行桥、引桥、桥墩、人行桥、过街天桥、缆索桥、级面桥、拱桥、亭桥、廊桥、溜索桥、栈桥
管线	线	输电线、架空线、配电线、通信线、管线、电力线、供电线、照明线、电车线、电信线、给水管线、地上管线、地下管线、架空管线、排水管线
	管道	油管道、地上管道、地下管道、架空管道、水管道

征词汇和要素特征词汇之间的对应关系进行了映射与关联，形成了每一个要素变化动词都有地理要素特征词汇与之对应。例如，地理要素特征词汇“河”，对应的变化特征词汇为“通航”和“开挖”等。

（3）空间关系词汇：包括拓扑关系词汇、距离关系词汇和方向关系词汇。本文针对文献[28]中空间关系词汇进行了筛选和过滤，同时将空间关系词汇和地理要素特征词汇进行了映射，形成了每个地理要素特征词汇都有相对应的空间关系词汇，如“河”对应着“汇入”“贯通”等词汇。

（4）停用关键词：为了剔除关键字相关但语义不相关的信息，比如规定、通告、通知、通报、公告、告示和技术规程。

知识关系可以描述和组织知识库所包含的各类事物，可分为词汇库内部的知识关系和词汇库之间的知识关系。词汇库内部的知识关系分为层次关系和继承关系，并将词汇库中的词汇进行分类，使得词汇之间具有从属和属性继承的关系，如地理要素特征词汇分为4大类，各大类包含若干一级词汇，一级词汇又包含若干二级词汇；词汇库之间的知识关系主要表达为它们之间的映射关系。例如，地理要素特征词汇“河”与地理要素变化词汇“通航”以及空间关系词汇“汇入”的语义关联。

3.2.2 知识库管理

本体模型提供了丰富清晰的知识结构表达能力，其结构简化为三元组，$O = <Class, Property, Individual>$，每个个体分别具有三元组中的Class、Property和Individual属性，并通过子类subClassOf建立等级层次关系，通过子属性Object Property和Data Property建立继承关系和属性关联关系。

本文的语义知识库利用本体结构软件Protégé 4.2来构建：首先，将各个词汇库分别作为父类，每个父类中利用subClassOf建立子类，在每个子类也是若干个以本体概念结构为基础的个体；然后，利用Protégé提供的属性关联机制（ObjectPropertyOf），使两个类之间建立映射关系，如每个地理要素特征词汇“路”都会有对应的地理要素变化词汇“竣工”等词进行描述，也就是说地理要素特征词汇具有属性并且这种属性是由地理要素变化词汇提供的。Protégé利用本体模型构建语义知识库中复杂的层次关系、继承关系和属性关联。知识库存放在Protégé生成的OWL2文件当中，可采用Protégé OWL API对知识库进行操作、使用、管理和维护。

3.3 网页文本获取与可信度的评价

3.3.1 网页文本获取与解析

网络爬虫是一种按照一定规则自动抓取互联网信息的程序或脚本，是目前网页获取最常用的工具。本文针对地理要素变化网页在网络中分布范围广的现象，构建基于搜索引擎的网络爬虫；针对地理要素变化信息在部分主题网页中集中分布，而搜索引擎对主题网站检索效果差的问题，构建通用主题爬虫。两种爬虫相结合，具有较好的优势互补性能。

（1）搜索引擎的网页获取：①通过搜索引擎机制，应用检索模板和关键词获取高相关度的种子URL；②采用多线程的方式对种子网页内的URL进行遍历爬取；③进行网页类型判断，区分主题型网页和一般型网页，主题型网页具有较多资源链接存入待爬序列，一般型网页进行内容爬取；④进行网页去重处理，去除重复的URL地址；⑤进行主题过滤，通过去除停用词汇网页的方式，得到筛选后相

关网页。

(2)通用主题爬虫的网页获取:通用主题爬虫与搜索引擎爬虫实现的功能一致,只是在种子URL获取方式和主题过滤方法上稍有差别。其种子URL可以由用户自行添加或删除,进行网站内或多网站的爬取。根据所得URL再进行网页采集,网页类型判断,网页去重,主题过滤等步骤获取网页文本。主题过滤是通过地理要素词汇进行判断,例如,网站名称中含有"水利"等词汇,说明网站对应的地理要素类型为"水系"。因此,只需遍历"水系"类别中的所有特征词汇,就可对其主题进行较为准确的判断。

为了避免获取一些明显不相关的网页,可以定制停用网站,两种爬虫可以自动略去这些网站的爬取,比如百度百科、百度文库、维基百科、互动百科和淘房网等网站。

通过网络爬虫获取的网页文本通常以HTML用来结构化网页中的标题、段落、列表等,适合以网页形式进行内容的浏览[35]。为了进一步挖掘文本中的内容信息,需将其解析为纯文本信息。本文利用正则解析和HTML parser相结合的方法,对获取的网页进行解析。正则解析的原理是进行标签匹配或字符匹配,可解析HTML中无固定位置出现的一些结构化信息,例如,应用正则表达式解析网页文本中的元数据。HTML parser功能强大,通过管理和操作树状的HTML结构,实现对HTML内容的解析,如段落文本的获取。

3.3.2　网页文本可信度的评价

在网页文本获取之后,可通过网页可信度分析,进一步去除相关度很低的无用网站。与PageRank的评价体系相比[35],文献[37]提出的可信度计算模型更加全面客观。本文结合地理要素变化检测需求,对该模型进行了修正。具体定义如下:

$$C(i)=\lambda_1 f_{tw}(i)+\lambda_2 f_{pg}(i)+\lambda_3 f_k(i)+\lambda_4 f_d(i)+\lambda_5 f_m(i) \quad (1)$$

其中,$C(i)$为可信度;$f_{tw}(i)$为地理要素特征词汇的类别等级;$f_{pg}(i)$为该信息的PageRank值;$f_k(i)$为网页类型;$f_d(i)$为网页发布时间;$f_m(i)$为网页元数据中是否包含主题词的权值。λ_1、λ_2、λ_3、λ_4和λ_5分别为其对应系数:λ_1指地理要素特征词汇等级,等级较低要素特征词汇的权重小于具有较高等级典型要素的权重,分别为1、2、3;λ_2指PageRank值等级,Google提供了查询网站PageRank值的API,网站PageRank值分为0–10共11个等级,数值越大其对应的网站重要性越高;λ_3指网页类型等级,政府网站(.gov)、商业网站(.com和.net)、非营利性网站(.org)和其他类型的网站(.edu等)的权值从高到低依次为5、3、1、1;λ_4为网页发布日期的等级,其权值按照距今10天、距今30天和大于30天由高到低依次为5、3、1;λ_5为网页元数据权重等级。根据不包含要素特征词汇、仅包含要素特征词汇、仅含要素变化词汇及包含要素特征词汇和要素变化词汇的不同权重,分别赋值为1、2、3、5。

3.4　地理要素变化信息的抽取

通常情况下,地理要素变化信息集中在一个句子之内。因此,本文以句子为单位,进行地理要素变化信息的关联范围,确定句内是否含有地理要素及其对应变化类型的词汇,若没有则分析下一句,若有则搜索上下文对相关信息分别进行抽取。

3.4.1　地理要素的抽取

本文以句子为单位,依据变化要素与变化类型之间的关联关系对要素进行抽取。具体包步骤为:(1)将待抽取文本进行分句;(2)利用ICTCLAS分词软件将待抽取文本进行分词;(3)定位到句子的动词,并与知识库中的变化要素词汇进行对比;(4)按照距离该动词距离由近到远的顺序,找出变化名词,并于知识库中的名词进行匹配;(5)遍历整个文本,抽取所有满足条件的组合;(6)按照"GBT 13923–2006地理信息要素分类与代码"中要素大类分类标准,将变化事件进行归类。以图2为例,文本根据扫描到的动词"迁移"为依据,扫描到句内待匹配的名词"面积"和"管线",再利用要素词汇和要素变化词汇之间的关系进行匹配,最终扫描到地理要素"管线"。

3.4.2　时间信息的抽取

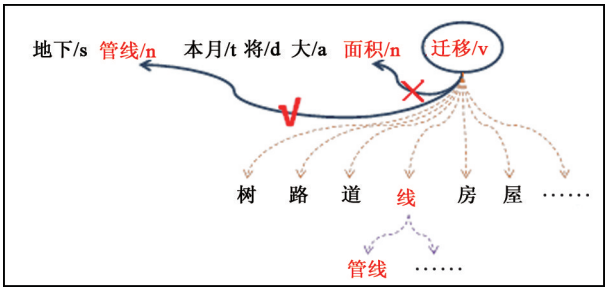

图2　地理要素变化信息抽取示意图

Fig.2　The schematic processes of change information extraction of geographic features

地理要素时间变化的语义表达相对复杂，主要存在两种描述方式：一是确切的时间表达，例如“中国首条穿越长江的地铁28日将通车”；二是预估的时间表达，例如，“地铁4号线一期年底前全线动工，预计2015年底前全线通车”。此外，中文文本表达方式的不完整性和多样性使时间要素的直接抽取相对困难。时间抽取通过基础触发词和规则模版相结合的方法，建立时间表达式并对检测到的时间要素进行抽取。时间推理解决句内时间不明确或缺失的问题，例如，“年内”无法直接作为时间要素进行抽取。通过大量分析句内和段内的时间语义表述，将时间信息分等级：句内出现的时间信息等级设置为“高”；前一个句内出现的时间信息等级设置为“低”。在判断过程中，以等级较高的时间要素为准，抽取对应的时间要素。例1：“扬子晚报讯秦淮区第17届人民代表大会第2次会议昨天召开。扬子晚报记者获悉，4片区、20个小区地下管网改造将启动，下半年完成。夫子庙景区的贡院街沿街立面店出新招，明年6月完工……”。文本中抽取出的时间有“昨天”、“下半年”和“明年6月”，以检测到的地理要素和地理要素变化类型为基础，根据句内的时间推理，得到变化时间应为“下半年”。

3.4.3 地名抽取

本文利用文献[33]和[38]中提出的条件随机场模型进行地名识别，信息抽取准确率和召回率达到70%以上。算法流程主要包括以下步骤：将待抽取文本以句子为单位，进行句子划分；利用分词软件将待抽取文本进行分词和词性标注；运用条件随机场模型算法对拆分文本进行地名识别；最后，将识别的地名进行输出和入库。

3.4.4 地理要素变化属性抽取

属性抽取是地理要素抽取的一个重要环节，涉及属性类型多且复杂。本文采用触发词和规则相结合的方法进行属性信息抽取。触发词利用知识库中的地理要素变化特征词汇，并可针对应用领域不同进行完善。属性要素的表达方式相对较固定，若采用机器学习方法需要大规模的语料库进行训练和测试。因此，本文采用人工总结的方式进行属性要素抽取规则总结。部分属性抽取规则如表2所示。

表2 地理要素变化属性抽取规则示例

Tab.2 Samples of syntactic rules for description of change attributes of geographical features

规则	实例
word1+*+word2+*+{(m+q)||(m||(m+q)+”至”||”到”||”~”||”-“||”—“+m+q)}	拓宽/word1 后将/* 达到/word2 500/m m^2/q
word1+*+word3+*+word2+*+{(m+q)||(m||(m+q)+”至”||”到”||”~”||”-“||”—“+m+q)}	改造/word1 后的望江路中段/* 宽度/word3 应该/* 是/word2 双向/* 6/m ~ 8/m 车道/q
v+*+{(m+q)||(m||(m+q)+”至”||”到”||”~”||”-“||”—“+m+q)}+*+“的”+word1	完成/v 了/* 6000/m m^2/q 小区/* 的改造/word1
word1+*+{(m+q)||(m||(m+q)+”至”||”到”||”~”||”-“||”—“+m+q)}	增宽/word1 了/* 2/m 车道/q

注：word1、word2和word3分别表示等级为1、2、3的触发词，“*”表示任意字符串，m表示数词，a表示量词，v表示动词

4 实验评估分析

本文以Microsoft Visual Studio 2010和Eclipse为平台，应用C#和Java开发了面向网页文本的地理要素变化检测系统（如图3），包括系统登录、电子地图、知识库管理、网页检索和空间分析等功能，并以南京地区为例进行实验评估分析。

实验选取http://www.njdaily.cn/（南京日报）和http://www.njjt.gov.cn/（南京交通运输局）两个种子URL，抓取层数设置为3层，进行主题爬虫的定向检索。其中，南京日报网站中处理的总链接记录为2295条，南京交通局网站中处理网页总链接数为336条。在进行搜索引擎的混合检索时，只对检索关键字和Google返回结果数量进行限制，检索关键字涵盖了4大地理要素特征词汇类型和7种检索组合关系（如表3），每个检索词返回记录条数为50。

（1）网页获取性能分析

实验中定向检索和混合检索的HTML解析率、查全率和查准率如表4所示。HTML解析率达95%以上；两种检索方式查全率和查准率都在70%以上，且定向检索的准确性优于混合检索。其中，HTML解析率未能达到100%的原因主要是个别网页

图3 面向网页文本的地理要素变化信息检测系统操作界面

Fig.3 The interface of geographic information change detection system based on web pages

表3　网页检索关键词组合关系

Tab.3　The retrieval expressions for related web pages

序号	检索关键词组合关系
1	地名+要素特征词汇+变化特征词汇
2	要素特征词汇+变化特征词汇
3	地名+要素特征词汇
4	地名+变化特征词汇
5	要素特征词汇+变化特征词汇+空间关系词汇+地名
6	要素特征词汇+空间关系词汇+地名
7	变化特征词汇+空间关系词汇+地名

为动态生成,或未注释其编码方式,因此,不能对其进行解析。网页的漏查和漏滤是影响查全率和查准率的2大主要因素。例如,“4月16日扬州市政府对文昌路LED路灯改造……”,其中,标题正文都涉及“路”、“改造”等相关词汇,但其主题与研究内容无直接相关,却被爬虫所获取。其原因主要在于中文文本语言表述的模糊性,且知识库不能对检索领域进行全面覆盖。

表4　网页获取和文本解析性能

Tab.4　The performance of web obtaining and text parsing

检索类型	HTML解析率(%)	查全率(%)	查准率(%)
定向检索	100	75	83
混合检索	95	70	74

注:查全率=处理后相关信息数目/人工判断相关信息数目;查准率=处理后相关信息数目/处理后信息总数目

(2)地理要素变化信息抽取性能分析

地理要素抽取实验准确率如表5所示。地理要素变化信息抽取分为对地理要素、变化类型、时间、地点和变化属性5个部分(见表6)。

表5　地理要素变化信息抽取的准确性

Tab.5　The accuracy of change information extraction of geographic features

检索类型	要素抽取	变化类型	变化时间	变化地点	变化属性
定向检索	83%	100%	83%	100%	73%
混合检索	78%	90%	71%	68%	70%

实验发现,地理要素抽取和要素变化类型抽取结果较为满意,准确性基本保持在80%以上,其中抽取误差主要集中于对关键词的误判断,比如人名、地名当中出现“路”、“河”等要素词汇,且后续文字出现相应的变化类型与之匹配;变化时间和变化地名的抽取可达到一定的正确率,但诸如“今后”、“近日”、“南京”等词汇涉及范围广泛,易引起语义不确定性;另外,属性抽取和关联由于上下文语义的模糊性,导致识别出错误信息,如样例1当中的变化属性“30m”,实际上是其工程的改造宽度,并非扩宽之后的宽度(52~60m)。综上所述,系统中的误差主要来源于网络文本中中文语义描述存在较强的模糊性和不确定性,并且知识库在全面性方面存在缺陷,需要进一步完善。另外,网页文本中识别出的地名,在与变化地理要素关联之后,还需要与地图进行空间匹配。本文实验选用南京市1:100万地名数据,每一条记录包括ID、地名、拓扑描述(点线面)、定位信息和其他信息。在实际应用中,地名词典可根据具体涉及的区域和尺度进行选取。

表6 地理要素变化信息抽取示例

Tab.6 Samples of change information extraction of geographic features

样例1	网页文本	近日，位于南京市主城东南部杨庄地区的石杨路(友谊河路—科技园二号路)拓宽改造工程全线贯通……负责此项道路拓宽工程的白下住建局局长徐仁彪介绍：本次拓宽改造主要实施南半幅约30m宽路段，道路全长2548m。建成后的道路红线宽52~60m，道路组成也更加安全规范……				
	抽取信息	地理要素	变化类型	时间	地名	属性
		路	拓宽	近日、11月	南京、杨庄、石杨路、友谊河路、科技园二号路	30m
	变化信息	变化要素	变化类型	规范化时间	变化地名	变化属性
		路	拓宽	2012年11月	科技园二号路	30m
样例2	网页文本	农历蛇年来临之际，南京工程局管线工程处根据吹填施工的现场需求，深入调研，大胆创新，周密部署，在青岛董家口工地尝试将700m沉管沉入18m水深的航道，并一举成功……铺设水域的水深达-18m以上，加上潮汐影响……				
	抽取信息	地理要素	变化类型	时间	地名	属性
		管线	铺设	今后、2月8日	南京、青岛	-18m
	变化信息	变化要素	变化类型	规范化时间	匹配地名	变化属性
		管线	铺设	2013年2月8日	南京	-18m

5 结论

本文探讨了一种面向网页文本的地理要素变化信息检测方法，研发了相应的原型系统，进行了实验验证分析。后续的研究工作将集中在3个方面：(1)完善网络爬虫和HTML解析技术，使得从网页上获取文本信息成为一个通用的平台接口；(2)完善知识库以提高地理要素变化检测的准确性，并论证变化信息的属实性；(3)添加与其他数据源的比对接口，如矢量地图和遥感影像，形成一套较为完善的地理要素变化检测工具，满足地理要素的持续更新和按时更新需求。

参考文献：

[1] 钱育华.数字城镇的数据更新[J].地球信息科学,2002,4(3):64-67.

[2] Chen J, Zhao R L ,Wang D H. Dynatmic updating system for national fundamental GIS: Concepts and research agenda[J]. Geomatics World, 2007,5(5):4-9.

[3] Heipke C. Updating geospatial databases from images[C]. // Li Z L, Chen J, Baltsavias E (Eds.). Advances in Photogrammetry, Remote Sensing and Spatial Information Sciences:2008 ISPRS Congress Book, Boca Raton: CRC Press, 2008,355-362.

[4] Badard T. Towards a generic updating tool for geographic databases[C]. Proceedings of GIS/LIS'98 Annual Exposition and Conference. USA, 1998，352-363.

[5] 陈军,王东华,商瑶玲.国家1:50000数据库更新工程总体设计研究与技术创新[J].测绘学报,2010,39(1):7-10.

[6] 王迪伟.基于PDA的1∶10000比例尺地形图野外调绘[J].测绘通报,2010(7):59-61.

[7] 李冰,曹宏文,曹歆宏.大比例尺地理信息数据库建设刍论[J].科技创新与生产力,2010,195(7):83-85.

[8] 王帅.初探首次全国地理国情普查[J].3S News Weekly, 2013(5):30-33.

[9] Palkowsky B, MetaCarta I. A new approach to information discovery——Geography really does matter[C]. Proceedings of the SPE Annual Technical Conference and Exhibition, 2005.

[10] Ai T. Constraints of progressive transmission of spatial data on the web[J]. Geo-spatial Information Science, 2010, 13(2):85-92.

[11] 容伟杰.网络信息存在的几大问题[J].图书馆学研究,2003(2):48-49.

[12] 孙瑞英.网络数据内容分析研究[J].图书馆学研究,2005(5):35-39.

[13] Wu M L, Li W J L, Lu Q, *et al.* CTEMP: A Chinese temporal parser for extracting and normalizing temporal information[C]. Natural Language Processing—IJCNLP 2005, Second International Joint Conference, Korea, 2005:694-706.

[14] 赵国荣.中文新闻语料中的时间短语识别方法研究[D].太原:山西大学,2006.

[15] 逯万辉,马建霞.基于条件随机场模型的复杂时间信息抽取研究[J].现代图书情报技术,2011(10):29-33.

[16] 宋洋,徐蔚然.基于条件随机场的事件起止时间表达式的识别[J].中国科技论文在线,2012(1):1-8.

[17] 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.
[18] 钱晶,张杰,张涛.基于最大熵的汉语人名地名识别方法研究[J].小型微型计算机系统,2006,27(9):1761-1764.
[19] 张雪英,闾国年,李伯秋,等.基于规则的中文地址要素解析方法[J].地球信息科学学报,2010,12(2):9-16.
[20] 李丽双,党延忠,廖文平,等.CRF与规则相结合的中文地名识别[J].大连理工大学学报,2012,52(2):285-289.
[21] 李丽双,黄德根,陈春荣.SVM与规则相结合的中文地名自动识别[J].中文信息学报,2006,20(5):27-51.
[22] Mark M D, Comas D, Egenhofer M J, *et al.* Evaluating and refining computational models of spatial relations through cross-linguistic human-subjects testing[C]. // Frank A and Kuhn W (Eds.). Spatial Information Theory – A Theoretical Basis for GIS, International Conference COSIT, Semmering, Austria, Lecture Notesin Computer Science. Berlin: Springer-Verlag,1995,553-568.
[23] Egenhofer M J. Locational SQL: Syntax extensions, surveying engineering program[D]. Orono: University of Maine,1987.
[24] Coyne B, Sproat R. WordsEye: An automatic text-to-scene conversion system[C]. Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, New York, 2001:487-496.
[25] Le X Q, Yang C J, Yu W Y. Spatial concept extraction based on spatial semantic role in natural language[J]. Geomatics and Information Science of Wuhan University, 2005,30(12):1100-1103.
[26] Zhang X Y, Lv G N. Natural-language spatial relations and their applications in GIS[J]. Geo-information Science, 2007,9(6):77-81.
[27] 朱少楠,张雪英,张春菊.地理空间关系描述的句法模式识别[C]. Proceedings of 2010 International Conference on Broadcast Technology and Multimedia Communication, 2010(4):355-357.
[28] 蒋文明.面向中文文本的空间方位关系抽取方法研究[D].南京:南京师范大学,2010.
[29] 高文利. IERDL——基于关键词驱动的信息抽取系统的规则描述语言[J].软件导刊,2009,10(8):67-69.
[30] Ravi S, Paşca M. Using structured text for large-scale attribute extraction[C]. Proceedings of the 17th ACM Conference on Information and Knowledge Management, ACM, 2008:1183-1192.
[31] Soderland S. Learning information extraction rules for semi-structured and free text[J]. Machine Learning, 1999, 34(1-3):233-272.
[32] 刘臻熙.中文文本中地理实体属性信息抽取方法研究[D].南京:南京师范大学,2010.
[33] 张春菊.面向特定时间的Web文本时刻属性信息挖掘方法[D].南京:南京师范大学,2013.
[34] 周立,邓云青.城市地理信息系统数据更新方式研究[J].地理空间信息,2008,6(5):45-47.
[35] 曾文华,黄桦.基于网页信息检索的地理信息变化检测方法[J].计算机应用,2010(4):1132-1134.
[36] 闫会杰,赵巍.服务于地理信息数据动态更新的网络蜘蛛[J].测绘技术装备,2012(2):21-22.
[37] Metzger M J. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research[J]. Journal of the American Society for Information Science and Technology, 2007,58 (13):2078-2091.
[38] 李丽双,党延忠,廖文平,等.CRF与规则相结合的中文地名识别[J].大连理工大学学报,2012,52(2):285-289.

Change Detection of Geographic Features Based on Web Pages

WANG Shu[1], JI Leijing[2], ZHANG Xueying[2*], ZHAO Renliang[3], CHEN Xiaodan[2] and YU Hao[4]
(1. *School of Geography, The University of Leeds, LS2 9JT, United Kingdom;* 2. *Key Laboratory of Virtual Geographical Environment, Ministry of Education, Nanjing Normal University, Nanjing* 210046, *China;* 3. *National Geomatics Center of China, Beijing* 100830, *China;* 4. *School of Computer Science, Nanjing University of Posts and Telecommunication, Nanjing* 210003, *China*)

Abstract: Geographic features change detection has became a vital component of the national geographical information 12th Five-Year-Plan and the national geographic general survey. In web pages, billions of geographic feature changes were contained, especially in government official websites, news homepages, social portals and etc. The web pages of these websites update frequently, which could provide the latest data for geographic infor-

mation change detection. Considering the complex characteristics of the web geographic information description, this paper did some valuable achievements. First of all, the geographic information knowledge base was established by summarizing the geographic information words and phrases, which could give the great supports to geographic information semantics change detection. Then, the web geographic information was obtained using two kinds of web crawler technologies. Combining the Google Custom Search crawler and general topic crawler, the web geographic information obtainment could be more complete in both scope and depth. Thirdly, the geographic information was parsed and extracted from the web text, which showed users the related features, place names, times and attributes. Last but not least, the prototype system was finally developed and the results were analyzed. The experiments indicated that the accuracy of related web pages obtainment and features change detection were over 74% and 70% respectively. In addition, the results of geographic information change detection highly relied on the integrity of knowledge base, which need to be completed further. Moreover, the uncertainty and fuzziness of web geographic information also limited the change detection results. Therefore, the web page based geographic information change detection could be a supplementary method of geographic information change detection. Combining the traditional surveying detection and remote-sensing imagery detection methods, it could solve the problems of continuous updating and timely updating of geographic information efficiently.

Key words: web text; geographic feature changes; information extraction; web crawler; text parsing

***Corresponding author:** ZHANG Xueying, E-mail:zhangsnowy@163.com