基于数据挖掘的地理信息服务聚合研究

姜代炜

(广西基础地理信息中心,广西 南宁 530023)

摘　要: 互联网环境下的地理信息服务聚合是当前的研究热点之一。本文研究了网络爬虫、数据挖掘、行业信息空间定位以及服务聚合技术,智能解析了多个行业的数据资源,并将空间化后的行业地理信息数据进行了注册与发布,实现了行业数据与基础地理信息数据的服务聚合。最后,开发原型信息系统,验证了该方法的可行性与有效性。

关键词: 服务聚合; 网络爬虫; 数据挖掘

中图分类号: P208　　**文献标识码:** A　　**文章编号:** 1672-5867(2019)11-0078-04

Research on Geographic Information Services Aggregation Based on Data Mining

JIANG Daiwei

(Guangxi Geomatics Center, Nanning 530023, China)

Abstract: The aggregation of geographic information services is one of the current research hotspots under the Internet. this paper studies the technology of the web crawler, the data mining, the spatial location of industry information and the service aggregation, then analyzes the data resources of multiple industries intelligently, registers and releases the specialized industry geographic information data to realize the service aggregation of industry data and basic geographic information data. Finally, a prototype information system is developed for verifying the feasibility and effectiveness of the method.

Key words: service aggregation; web crawler; data mining

0　引　言

基础地理信息提供了电子地图、遥感影像、地名搜索等服务,满足了地图浏览、路线查找等基本需求,它是构建地理信息应用必不可少的基础服务资源。在我国,随着电子政务、数字城市、智慧城市建设的逐步推进,政府各职能部门对基础地理信息服务的需求越来越迫切[1]。然而,基础地理信息服务在面对不同类型用户的需求时却是单一、有限的,不能很好地满足实际应用的需求。一方面,公众服务、行业应用已普遍使用互联网,用户对于信息的感知度更加敏锐;另一方面,Web 2.0 时代的到来,使得网络信息资源急剧膨胀,它蕴含了大量、非空间化的地理信息,此类信息是一种巨大的信息战略资源,急需采集和利用[2]。因此,如何在海量的网页中快速、准确地抓取与地理信息相关的行业信息,如何使非空间化的行业信息空间可视化,并能够与已有的基础地理信息服务聚合,支持联合查询与协同分析,还有待研究[3]。

鉴于以上问题,本文设计了一种基于数据挖掘的地理信息服务聚合方法,实现了非空间化的、异构的行业信息网络化采集、净化与空间化,并与现有的基础地理信息服务进行了服务聚合,更好地挖掘了网络地理信息资源,以满足数字广西地理空间框架所倡导的更全面、更准确、更详细、更完整的地理信息服务目标。

1　总体思路

总体思路如图 1 所示。①借鉴搜索引擎的网络爬虫[4],在异构的网络环境中,对非空间化的行业地理信息进行自动采集;②使用数据挖掘的方法对行业数据进行清洗和整理;③使用地名地址匹配技术,将数据中包含的地名地址信息与现有的地名地址信息进行匹配,实现空

收稿日期: 2018-06-20

基金项目: 数字广西地理空间框架项目关键技术研究——行业数据与地理信息数据服务聚合技术研究(GXZC2014-G3-2233-GYZB)资助

作者简介: 姜代炜(1985-),男,广西全州人,注册测绘师,硕士,2012 年毕业于武汉大学摄影测量与遥感专业,主要从事 GIS 研发方面的工作。

间定位; ④根据行业应用需求,对业务流程稳固、数据和功能优化后的服务进行聚合,并将这类服务注册与发布,形成新的地理信息服务,提供唯一的统一资源定位符(Universal Resource Locator: URL) ,方便用户发现和使用。

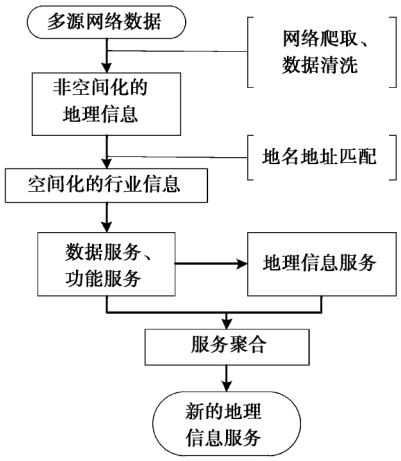

图 1　总体思路

Fig.1　General idea

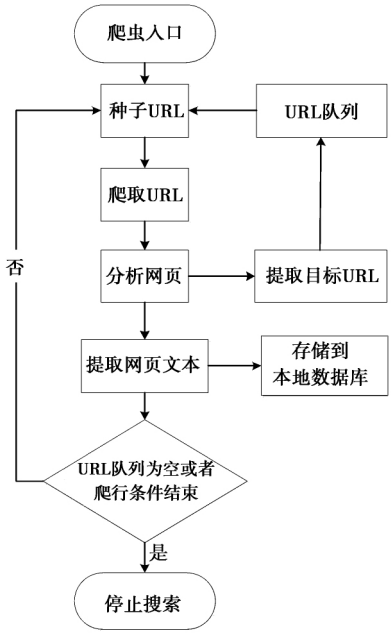

图 2　网络爬虫工作原理图

Fig.2　Web crawler working principle diagram

2　关键技术

2.1　数据爬取

网络爬虫(Web Crawler) 在互联网上漫游,可以对网络上几亿甚至于上百亿的网页进行爬取,是当今主流搜索引擎的信息采集利器。其工作过程是通过 URL 不断地从一个网页调到另一个网页爬取信息: ①人工选定一些 URL 作为起始点,以这些 URL 作为种子,根据 HTTP 协议向 Web 服务器进行网页的请求,开始对相关页面进行访问; ②对每一个被访问的页面,根据一定的爬行策略进行过滤,待页面正确获取后保存到本地; ③对被保存到本地的网页,进行网页文本分析处理,并提取出目标 URL 加入 URL 种子队列; ④继续从 URL 种子队列中选出一个 URL,重复①—③过程,直到满足网络爬虫停止的条件时结束。其工作原理如图 2 所示。

在搜索应用中,爬行策略至关重要。考虑到地名地址匹配,本文使用了最佳优先的、与主题相关的爬行策略[4],目标网页中包含的地名地址与基础地理信息的 POI 点具有一定的相似度,可以分别计算基础地理信息中的地名地址与 URL 网页文本解析得到的地名地址之间的相似度,取二者之间的均值作为当前文本的主题相关度,以 URL 队列中相关度最大者作为最优匹配结果,最优匹配的网页文本保存于本地数据库中。

2.2　数据挖掘

“当多元空间数据汇集一处时,就有可能存在高维数据、不同数据结构、不同的投影体系或度量单位等,以及噪声、误差的影像,势必需对空间数据清理。[5] ”已抓取的网页文本,其数据内容是繁杂的,部分数据是冗余的,甚至是完全无关的,它们的存在影响到有价值的信息发现。例如,有些文本中包含的数据信息是非线性的、粗粒度的,并不能直接使用,需要进行数据清洗,降低数据维度。而不同的行业数据,有着不同的特性,应通过数据挖掘归纳出行业数据的特征信息,发现行业数据的信息规律。

在数据挖掘方法的选择上,本文使用了支持向量机的监督分类方法。该方法建立在统计学习和经验风险最小的理论基础上,能够很好地解决小样本、非线性及高维模式识别方面的问题。利用支持向量机中核函数构建的超平面以区分不同行业的数据信息。在具体操作过程中,首先,从每个行业的数据中随机抽取 10 000 条以上的数据记录进行样本训练,统计各行业关键词出现的频率; 然后,构建训练样本的特征空间分词库; 最后,使用支持向量机的分类方法对未知文本进行筛选和分类。值得注意的是,对于中文网页文本的处理,使用中文分词方法[6]将文本区分为名词、动词、介词、连词、数词、标点符号、地名地址等(见表 1) 。

表 1　中文分词部分词库对照表

Tab.1　Chinese participle part vocabulary comparison table

类型	举例
名词	工业、游客、景区、增长率、生产总值……
动词	增加、增长、下降、核算、实现、预计……
连词	和、以及、与、因为、所以、由于……
标点符号	“、” “。” “;” “:” “!”……
介词	随着、通过、根据、从、由、和、把、将……
地名地址	广西、南宁、旅发委、广西日报社、广西融合集团……
方位词	东、南、西、北、前、后、左、右……
数词	462.82 亿元、8.3%、-29.3%、1.1 个百分点、500 人……

其中，介词、连词、标点符号等词类是与数据信息无关的噪声词类，予以去除；地名地址是空间定位的基础，存储于数据库中；数词是各行业数据的重要属性信息，与相应的名词、动词建立关键字联系。按照朴素贝叶斯统计方法以所有名词、动词为基底，统计各样本空间的动词、名词出现频率，以频率较高者作为该样本空间的特征向量，例如：广西发改委网页文本的特征词频率统计情况，如图 3 所示。

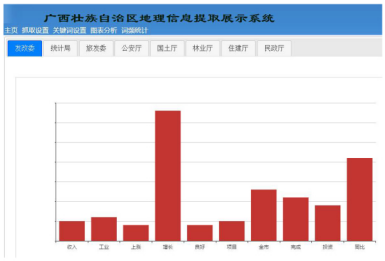

图 3 特征词频率统计图

Fig.3 Characteristic word frequency chart

2.3 地名地址匹配

无论从主题中还是从文本中抽取的地名地址都是非结构化的地理信息，均未含有空间地理坐标，需要进行空间定位。由于数字广西地理空间框架的基础地理信息数据已含有 600 000 条以上的地名地址数据，它存储了地理实体名称、地名地址名称及相应的空间坐标信息，可以使用该数据作为空间参考库，与网页文本的地理信息进行地名地址匹配，挂接各行业数据的属性信息，实现空间定位。

地名地址匹配包括精确匹配和模糊匹配。精确匹配用以对网页文本中具有详尽描述的地理信息进行空间定位，模糊匹配用以对网页文本中描述粗略或者不全的地理信息进行空间定位。在地名地址匹配过程中，网页文本中的地名地址描述与标准化的地名地址描述常常不一致（如：在网页文本中描述为“鹏程驾校”，而在标准化的地名地址描述为“广西壮族自治区南宁市江南区那洪街道金凯路鹏程驾校”），给地名地址匹配带来了一定的困难，需要将网页文本中的地名地址进行标准化处理。参考 l CH/Z 9002—2007 数字城市地理空间信息公共平台地名/地址分类、描述及编码规则，标准化的地名地址描述表现为一种树状的层次结构模型（如图 4 所示）。因此，在程序中将地名地址描述设计成一种可扩展的树状模型，对网络文本中的地名地址进行切分，对照树状模型由上而下依次匹配，当上级节点匹配成功时，搜索下级节点，再进行匹配，直到在地址参考库中找不到匹配的地名地址描述，最后根据权重情况确定该地名地址描述，将此时地名地址参考库中的坐标信息和行业数据的属性信息进行挂接，实现空间坐标定位。

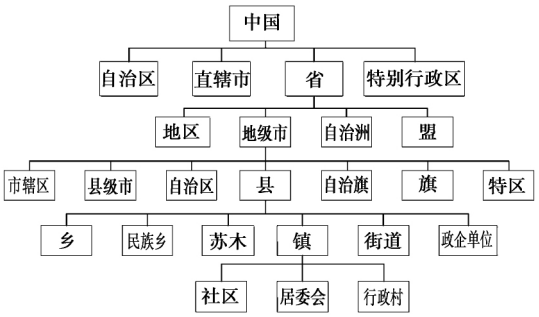

图 4 地名地址树状层次结构模型

Fig.4 Place name address tree hierarchy model

2.4 服务聚合

地理信息服务聚合通过地理信息服务之间的通信与协作，将分散、相对简单的细粒度服务组合成复杂的具有新功能的粗粒度服务，提高服务的利用率和可重用性，构建全新的应用，实现信息服务的增值[7]。经过网页文本中挖掘的行业数据，具有较高的时效性，它不仅是行业部门高度关注的事件，而且也是一种低成本、高效率获取的地理信息，将此类地理信息与基础地理信息服务聚合，可进一步地丰富地理信息的内容，体现行业数据的价值。

地理信息服务聚合需经过单一的地理信息服务到多种服务聚合的过程。遵循 OGC/ISO 的地理信息规范，使用数字广西地理空间框架服务引擎注册、发布空间化后的行业地理信息，并对服务描述的内容、功能、接口和访问方式等进行阐释，提供唯一的 URL 地址，方便用户搜索、发现和使用。

地理信息服务聚合包含服务端聚合与客户端聚合。服务端聚合在服务端完成，旨在叠加多源、异构的地理信息服务，作为一个整体返回给用户。目前，数字广西地理空间框架的“天地图·广西”已在服务端纵向实现了国家、自治区、市、县四级节点的信息服务聚合，并结合高分辨率对地观测系统广西数据与应用中心的需求，聚合了高分系列、资源系列和北京二号影像服务，提供影像查询、检索等功能聚合服务。客户端聚合在客户端完成，旨在聚合用户本身的业务服务和第三方地理信息服务，属于一种轻量级的聚合服务应用。为满足用户多样化的业务需求，将原先组件式的处理方法细化为原子级的处理方法，提供细粒度的服务调用、在线工具和开发工具等方法，按需组装业务功能。同时，在行业部门数据基础上，将非空间化的业务数据空间化，关联相应的地理对象，提供空间信息、图文信息及关联信息的查询，满足行业部门专业信息融合、业务功能定制与基础地理信息服务的集成。如图 5 所示的河池市精准脱贫攻坚指挥系统，其地理信息服务调用了数字广西地理空间框架与数字河池地理空间框架的第三方服务，专题信息通过采集和空间化处理，在客户端发布并调用，功能服务则使用原子级的处理方法进行多样化的组装，完成了战区分布、战场研判、战果监督、战果展示等功能的集成。

图 5　河池市精准脱贫攻坚指挥系统服务聚合图

Fig.5　Hechi City accurate poverty alleviation command system service aggregation map

3　实验结果

在 Java Script 语言环境下,本文使用以上方法开发了一套原型信息系统。该系统将各行业门户网站爬取的数据(广西发改委、统计局、旅发委、公安厅、林业厅、住建厅等)以 REST 服务方式进行了注册、发布,以富客户端的方式实现了与数字广西地理空间框架基础地理信息服务的服务聚合。系统调用了天地图·广西的矢量地图服务,将各行业门户网站获取的数据在前端进行直观展示,并提供空间查询、统计等功能服务。实验结果表明,本文方法是可行的。

4　结束语

在互联网环境下,本文利用数据爬取、数据挖掘、行业信息空间定位以及服务聚合技术,通过挖掘与行业息息相关的地理信息,将行业数据进行空间可视化处理,并进行了注册与发布,完成行业数据与基础地理信息数据的服务聚合,以满足各种行业对地理信息个性化的需求。该种方法采集的行业地理信息数据,具有时效性、准确性和空间分布特征,同时也能够将行业数据和基础地理信息数据有效融合,可进一步丰富数字广西地理空间框架的数据资源,为用户提供更为翔实、便捷、有价值的信息服务。

参考文献:

[1] 龚健雅.地理信息系统基础[M].北京: 科学出版社,2003.

[2] 王克永.面向网页文本的地理信息要素提取与空间定位方法研究[D].泰安: 山东农业大学,2015.

[3] 段兵营.搜索引擎中网络爬虫的研究与实现[D].西安: 西安电子科技大学,2014.

[4] 陈睿嘉,康志忠,张卫涛,等.基于网络爬虫的导航深度服务信息自动采集[J].测绘工程,2015,24(1):17-24.

[5] 李德仁,王树良,李德毅,等.空间数据挖掘理论与应用[M].北京: 北京科学出版社,2006.

[6] 潘正高.基于规则和统计相结合的中文命名实体识别研究[J].情报科学,2012,30(5):708-712.

[7] 张珊.REST 式 GIS 服务聚合研究及软件开发[D].上海: 华东师范大学,2011.

[编辑: 张　曦]

(上接第 77 页)

由于目前从理论上无法严谨地解释不整平时竖盘读数观测值误差方程可以采用整平时竖盘读数误差方程的原因,所以仍有待后续的研究。

参考文献:

[1] 徐宜敏.全站仪免置平测量技术及其算法模型研究[D].南昌: 南昌大学,2013.

[2] 王鹏,刘成龙,杨希.无碴轨道 CPⅢ自由设站边角交会网平差概略坐标计算方法研究[J].铁道勘察,2008(3):26-29.

[3] 王兆祥.铁道工程测量[M].北京: 中国铁道出版社,2010.

[4] 张忠良,杨友涛,刘成龙.轨道精调中后方交会点三维严密平差方法研究[J].铁道工程学报,2008(5):33-36,71.

[5] 罗远刚.三维平差技术在高铁轨道控制网测量中的应用研究[D].成都: 西南交通大学,2014.

[6] 郭剑琴,宣伟,余锐,等.全站仪不对中不整平条件下的测量[J].地理空间信息,2012,10(5):114-116.

[7] 吴迪军,何广源,熊伟.对现行规范中光电测距倾斜改正公式的探讨[J].测绘通报,2013(10):73-75.

[8] 施一民.三角高程测量的公式论证及应用[J].测绘通报,2003(1):1-3.

[9] 王穗辉.误差理论与测量平差[M].上海: 同济大学出版社,2010.

[10] 王磊,刘成龙,杨雪峰,等.高速铁路自由设站 3 维整体平差计算及精度评定[J].测绘科学技术学报,2011,28(4):258-261.

[11] 中华人民共和国铁道部.TB10601—2009 高速铁路工程测量规范[S].北京: 中国铁道出版社,2009.

[12] 李毛毛.无砟轨道 CPⅢ控制网数据处理方法研究及其软件的集成[D].成都: 西南交通大学,2008.

[13] 崔希璋,於总倓,陶本藻,等.广义测量平差[M].第 2 版.武汉: 武汉大学出版社,2012.

[14] 朱洪涛,徐宜敏,吴维军.全站仪免置平自由设站及其测量方法[J].铁道标准设计,2013(6):25-28.

[编辑: 张　曦]