

基于网络爬虫的地名数据库维护方法

张春菊, 张雪英*, 朱少楠, 徐希涛

(南京师范大学虚拟地理环境教育部重点实验室, 南京 210046)

摘要: 目前,我国地名数据库建设存在大、中颗粒度地名集中,小颗粒度地名较为缺乏,地名资料陈旧、时效性较低,简称、别名等非标准地名信息和地名的相对位置信息缺失等问题。而地名数据库的更新维护工作主要通过人工测绘手段完成,存在周期长、成本高、效率低等缺点。针对这一问题,本文以现有地名数据库和空间关系词汇为基础,基于 Google 搜索引擎服务,提出一种以网页资源为数据源,利用网络爬虫技术和地名识别技术,进行地名数据库更新维护的方法。首先,设计以地名为主题的网络爬虫,实现非结构化的网页数据中海量空间敏感网页文本的主动获取;然后,采用 HTML DOM 技术解析空间敏感网页并应用 CRF 地名识别模型自动识别网页文本中地名;最后,设计相关算法进行网页文本中地名信息的自动解析,实现新地名和地名空间位置信息的获取,进行地名数据库的更新维护。以“南京师范大学仙林宾馆+西北”为空间检索实例,验证了此方法的可行性。

关键词: 地名数据库;网络爬虫;地名识别;主题相关性

DOI: 10.3724/SP.J.1047.2011.00492

1 研究背景

地名是人们赋予宇宙中特定地理实体的代号,是区别某一特定地理实体与其他地理实体的一种标志。地名数据库利用现代数据库技术,采用数字、文字、图像、声音等多媒体形式,对地名相关信息进行存贮、组织和管理^[1-2]。地名数据库是地名公共服务的基础,为国家行政管理、经济建设、国内外交往等提供不可或缺的基础信息资源。特别是随着社会信息化的发展,以地名信息为基础的基于位置服务(Location-Based-Service, LBS)的需求日益增长,并在日常生活中潜移默化地改变着人们的生活。例如,寻找餐馆、旅店、娱乐中心、购物中心等常规的寻址问路,以及弘扬地名文化、旅游文化、畅享虚拟城市游戏、共享网络社区交流等多元化的空间位置服务。而建立信息完备、时效性强的地名数据库是实现 LBS 高效服务的前提和保障^[3]。

长期以来,欧美等国家地名命名比较规则,地名数据库内容较为规范,地名数据库的构建、更新维护较为容易。比较典型案例有亚历山大数字图书馆地名数据库(ADL)^[4]、美国地名信息系统

(GNIS)^[5]、澳大利亚地名数据库(GOA)^[6]等。这些地名数据库具备比较完善、实时的地名描述信息,提供免费共享服务,已经成功应用于国家的政治、外交、军事、经济和公众服务等各个领域。我国地名数据库建设起步较晚,主要由民政部门和测绘单位承担。1979年至1986年期间,民政部门开展了第一次全国地名普查工作,2009年至2012年间的第二次全国地名普查试点工作正在启动。本次普查内容侧重于现代地名信息数据库的建设,信息采集内容主要包括试点区的地名及相关属性信息的清查、不规范地名的标准化、重要地理实体的地名标志设置等。2003年民政部颁发“关于建立地名数据库有关问题的通知”,全国各省市都加快了当地地名数据库建设的步伐,县级以上行政单位基本建立了本地地名数据库。自1994年以来,国家测绘局相继建成了全国1:100万、1:25万和1:5万地名数据库^[7-8]。该数据库将国家地形图上各类地名注记及其汉语拼音、属性要素等录入计算机,与地形数据库通过技术结构连接实现相互访问,或作为独立的关系型数据库运行。目前,各省、自治区、直辖市正在开展省级1:1万地名数据库的建

收稿日期: 2011-04-07; 修回日期: 2011-06-22.

基金项目: 国家自然科学基金项目(40971231)。

作者简介: 张春菊(1984-),安徽人,博士研究生,主要从事地理信息智能处理的理论与方法研究。E-mail: zcjtzw@sina.com

* 通讯作者: 张雪英(1970-),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的研究。

E-mail: zhangsnowy@163.com.

设(部分已完成)。

通常情况下,人们对位置信息描述时地名颗粒度较小,而且习惯使用地名的别称、简称、地名属性、相对位置关系等相关信息进行描述。尽管民政部门 and 测绘单位采用现代测绘技术手段,建立了各级别的国家和地方地名数据库,并进行了地名数据库维护更新的相关工作^[9]。但是地名数据库建设不能够满足社会需求,存在较多亟待解决的难题。主要包括以下几个方面:(1)地名信息采集缺乏统一规范,信息描述非标准化;(2)大、中颗粒度地名集中,小颗粒度地名和非标准地名信息较为缺乏;(3)地名资料陈旧、时效性较低;(4)地名的属性信息描述不完善;(5)地名的相对位置描述信息缺失;(6)地名数据库更新维护主要采用人工测绘手段,周期长、成本高、效率低。因此,对地名数据库进行高效持续的更新维护具有十分迫切的需求。

近年来,随着网络资源的日益丰富,以及网页资源的更新速度和参与者的日益增多,大量的中文信息以电子文档的形式出现在人们面前。据调查显示,人类社会 80% 以上的信息资源与地理空间相关。作为人类信息资源表达的载体,网页文档中蕴含着丰富的地理空间信息^[10]。本文基于 Google 搜索引擎服务,以网页资源为数据源,利用网络爬虫技术从非结构化的网页数据中主动获取海量空间敏感的网页文本,应用地名识别和空间位置解析技术实现地名数据库的快速更新维护,可以有效解决当前地名数据库建设与社会需求之间的矛盾。

2 地名数据库模型

地名数据库是一定区域内的地理要素及其相互联系和各种地名特征的数据集合,包含地名、类型和位置 3 个基本征及其他信息的列表^[11]。地名信息随着时间而演变,蕴含着丰富的地名时空信息。然而,目前已建立的大部分地名数据库,关注了历史地名信息的描述,但忽略了时空信息,而且描述的范围、完备性和丰富性等方面存在差异,缺乏统一的数据结构。本文参照文献^[12]中地名时空演变数据模型的框架,构建地名数据库模型。该模型能够较为完善地表达地名实体的基本信息、要素分类体系、空间位置信息、时间信息、空间关系等时空一体化属性信息,具有完整性、合理性和丰富性。从网页文本中解析地名的时空属性及非时空

属性信息,特别是地名的时空演变过程,是一个相当复杂的过程。因此,本文侧重于新地名和地名空间位置信息(经纬度坐标和相对位置)的获取。其中,新地名是指地名数据库未收录的地名。

3 基于网络爬虫的地名数据库维护

基于网络爬虫的地名数据库维护主要包括网络爬虫的空间敏感网页获取和网页文本中地名信息的自动解析两部分内容(参见图 1)。首先,以现有地名数据库和空间关系词汇为基础,基于 Google 搜索引擎服务,采用可定制网络爬虫技术获取空间敏感网页;然后,采用 DOM 方法解析网页,并使用 CRF 地名识别模型对网页文本中的地名进行识别;最后,应用地名数据库匹配和空间位置信息获取的相关算法,进行网页文本中新地名的获取和地名空间位置信息的解析,更新维护现有地名数据库。基于网络爬虫的地名数据库维护和空间敏感网络爬虫子模块都是动态循环的过程。

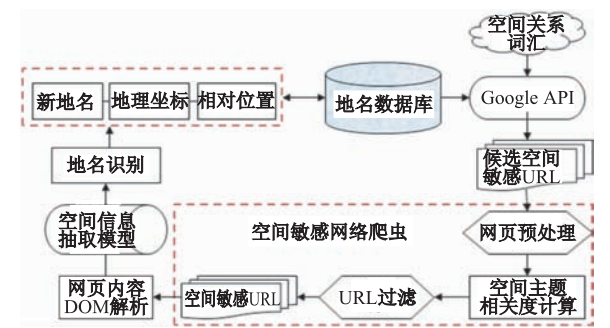


图 1 基于网络爬虫的地名数据库维护流程
Fig. 1 Updating flow of toponym database
based on web crawler

3.1 空间敏感网络爬虫的搜索

网络爬虫是搜索引擎中用来探索与下载网页资源的程序,搜索过程中程序自己判断下一步抓取的对象,有一定的智能性,各个网页之间的链接将互联网组成了一个网状结构^[13]。网络爬虫的抓取策略有 IP 地址搜索策略、广度优先、深度优先和最佳优先等几种搜索策略^[14]。随着互联网的不断发展和日益普及,互联网上的信息量呈现爆炸式的增长。传统网络爬虫获取的网页涉及领域范围过于广泛,而在特定领域的查询上则不够深入和专业化,整个采集过程中主题性不够突出。本文中的空间敏感网络爬虫是指以构筑空间信息领域的信息资源库为目标,遍历现有地名数据库,基于 Google

搜索引擎服务,按照“地名”或者“地名+空间关系词汇”形式的空间检索主题,在互联网上智能地搜集与该空间主题相关的信息资源,进行候选空间敏感网页和空间检索主题相关性的预测、判断和筛选。空间敏感网络爬虫在搜索的过程中,只需选择与空间检索主题相关的页面,无须对整个 Web 进行遍历,具有准确、针对、深入、高效的特点。

3.2 基于网络爬虫的空间敏感网页获取与解析

3.2.1 空间敏感网页获取

本文选用目前比较成熟的 Google 搜索引擎服务,基于空间查询关键字获取候选空间敏感网页^[15]。候选空间敏感网页是空间敏感网络爬虫进行爬行的起始页面,其质量直接决定了爬行主题的内容相关度。为了聚焦地名相关信息,采用 Web 服务的形式调用 Google Search API 实现。首先,注册 Google 账户。然后,遍历现有地名数据库,取出地名并将其作为种子地名,利用 Google 搜索引擎的主题搜索功能,以“地名”(如“北京”)或者“地名+空间关系词汇”(如“南京+东南”)的检索词形式获取相关网页,即候选空间敏感 URL(统一资源定位符)。用户可根据 Google 搜索结果定制候选页面的数量。

上文获取的候选空间敏感网页,系采用空间敏感网络爬虫技术进行网页过滤。首先,分析候选 URL 页面,采用正则表达式剔除语法标记、纠正不合格语法及去掉重复的网页地址,采用中科院研究所研制的 ICTCLASS 软件对网页页面进行分词预处理。然后,网络爬虫将候选 URL 页面置为着陆页面,遍历候选 URL 页面并分析页面上所有的链接、网页标题、网页正文,计算待选 URL 页面与空间检索词的空间主题相关度。Native Bays、神经网络、实例映射模型、向量空间模型等都是常用的网页文本主题相关度计算方法^[16],本文选用计算简洁、效率较高的向量空间模型算法。其将空间检索词的个数作为向量空间的维数,每个空间检索词的权值作为每一维分量的大小,通过挖掘候选 URL 页面中空间主题关键字及其出现频率,计算文本空间主题向量,进而计算候选 URL 页面与空间检索词的主题相关度(参见公式 1)。

$$\text{Sim}(D_1, D) = \text{Sim}(D_2, D) * \alpha + \text{Sim}(D_3, D) * \beta$$

(1)

式中, D_1 为空间检索主题, D 为待选 URL 页

面, D_2 和 D_3 分别为待选 URL 页面的正文和标题, $\text{Sim}(D_1, D)$ 为待选 URL 页面与空间检索词的空间主题相关度, $\text{Sim}(D_2, D)$ 和 $\text{Sim}(D_3, D)$ (参见公式 2) 分别为待选 URL 页面的正文和标题与空间检索词的空间主题相关度, α, β 分别为 $\text{Sim}(D_2, D)$ 和 $\text{Sim}(D_3, D)$ 的权值 ($\alpha < \beta$)。

$$\text{Sim}(D_2/D_3, D) = \frac{x_1w_1^2 + x_2w_2^2 + \cdots + x_nw_n^2}{\sqrt{w_1^2 + w_2^2 + \cdots + w_n^2} \sqrt{x_1^2w_1^2 + x_2^2w_2^2 + \cdots + x_n^2w_n^2}}$$

(2)

式中, $w_1, w_2, \cdots, w_n (i=1, 2, \cdots, n)$ 为空间检索词的主题向量, n 表示空间检索词的个数, w_i 为每个检索词的权重; $x_1w_1, x_2w_2, \cdots, x_nw_n (i=1, 2, \cdots, n)$ 为待选 URL 页面空间主题向量, x_i 为待选 URL 页面中各空间检索词出现的频率, x_iw_i 表示该页面对应向量的每一维分量。最后,设置空间主题相关度阈值,根据待选 URL 页面与空间检索词的空间主题相关度过滤待选空间敏感 URL 网页。

3.2.2 空间敏感网页解析

HTML 页面是由标题、标签、链接、图片和样式等组合而成,经过浏览器解析以网页的形式呈现给用户。这种图文并茂的网页适合浏览,但是如果需要进一步对网页结构和内容进行分析和挖掘,则必须将原始网页转换成结构清晰的格式。HTML Document Object Model(文档对象模型,HTML DOM)则是专门适用于 HTML/XHTML 的文档对象模型^[17]。HTML DOM 把文档表示为节点(node)对象树,网页中的各个元素都看作一个对象,从而使网页中的元素可以被计算机语言获取或者编辑。这种 DOM 树状结构被定义为父节点、子节点和兄弟节点互相联系的对象集合,节点对象不仅代表文档中的 HTML 元素,而且包含文档内的所有内容,如属性、注释和数据等。

本文采用 HTML DOM 技术进行空间敏感网页解析。首先,对 HTML 文档进行规范化表达,包括标记间的匹配、标记的正确嵌套、标记的属性值是否在引号中等。然后,利用 DOM 方法解析网页,形成一棵以 HTML 为根节点的结构明晰、层次好的 DOM 标记树,树中的每个节点由网页中的所有标记属性对构成(见图 2)。最后,去除 HTML 标签、样式标签、网页脚本等信息,生成网页文本。Web 文档具有多样性,存在部分未被识别的不规范标签和影响地名识别的文本,此类情况可通过人工辅助完成。

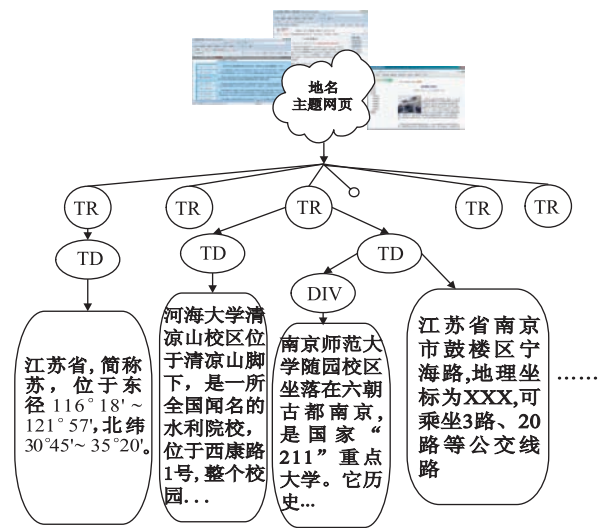


图 2 空间敏感网页 DOM 解析结果示意图

Fig. 2 Analysis result of web pages with a DOM tree method

3.3 网页文本中地名信息解析

3.3.1 地名识别和新地名获取

地名识别是信息抽取领域的重要研究内容, 目前, 主要有规则方法、统计方法和机器学习方法^[18]。近年来, 各种统计模型的机器学习方法在地名识别领域受到了广泛的关注, 如条件随机场模型(CRF)、最大熵模型、隐马尔科夫模型、最大熵隐马尔科夫模型等^[19-20]。本文选用条件随机场模型识别网页文本中的地名, 并检验地名识别结果的有效性。

条件随机场是一种以给定的输入节点值为条件来预测输出节点值概率的无向图模型, 在观测序列的基础上对目标序列进行建模, 重点解决序列化标注的问题, 已经被成功应用于生物医学、计算机语言学 and 语音识别等领域。假设 O 是一个“输入”随机变量的集合, 其值可以被观察, S 是一个“输出”随机变量的集合, 其值要求模型能够预测。这些随机变量之间通过指示依赖关系的无向边连接, 让 $C(O, S)$ 表示这个图中团的集合, CRF 模型在给定 O 的情况下, 将一系列输出随机变量值 S 的条件概率定义为与无向图中各个团的势函数的乘积成正比:

$$P_A(s | o) = \frac{1}{Z_o} \prod_{c \in C(s, o)} \Phi(s_c, o_c) \tag{3}$$

$$\Phi(S_c, O_c) = \exp(\sum_{k=1}^K \lambda_k f_k(s_c, o_c)) \tag{4}$$

式中, $\Phi(S_c, O_c)$ 表示团 C 的势函数, 一般定义为团的所有特征的带权指数形式, Z_o 是一个归一化因子, 每一个 f_k 是一个布尔特征值, 它依赖于状态

S 输入观察序列 O 的任何特征, Φ_k 是特征参数。

例如, 网页文本中的描述“坐落在上海市浦东新区世博园内的中国馆主体造型雄浑有力, 宛如花冠高耸, 屋顶酷似九宫格。”, 其基于 CRF 模型的地名识别结果为: “坐落在上海市/LOCATION 浦东新区/ LOCATION 世博园/ LOCATION 内的中国馆/LOCATION 主体造型雄浑有力, 宛如花冠高耸, 屋顶酷似九宫格。”(标签“/ LOCATION”表示识别后的地名)。

CRF 模型的地名识别结果, 采用地名数据库匹配的方式, 从网页文本中获取新地名, 即现有地名数据库中未收录的地名。

3.3.2 地名空间位置信息获取

本文中地名空间位置信息获取侧重于地名的空间坐标信息和相对位置信息, 包括空间位置信息的获取及其与地名的关联。地名空间位置信息的获取以上文中地名识别结果为前提。网页文本中蕴含的地名空间位置信息, 一般包括“地名+地理坐标”和“地名+相对位置”两种形式。

(1) “地名+地理坐标”形式

分析候选空间敏感网页发现, 省、市、县的黄页和专业性地名网站中的地名大都是以“地名+地理坐标”形式成对出现。在将 HTML 网页解析成 DOM 树时, 如果当前网页中只出现一个地名和地理坐标对, 则直接将标签内的地名和地理坐标关联起来。如果当前网页中出现多个地名和地理坐标, 在抽取出地名和地理坐标的同时, 必须实现地名和地理坐标之间的正确关联。针对此类现象, 本文基于两个假定条件, 即“DOM 树中两个实体离得越近, 相关性越高”和“两个关联实体在地理语义层面没有冲突”, 设计了地名和地理坐标关联的具体算法如下:

Step1: 对于每一个叶子节点上的地理坐标, 后序遍历直至找到第一个满足上述假定条件的地名。如果地名和地理坐标相关联, 将这个“地名-空间位置”对存储在新建树的节点中, 同时将所有未关联的节点传递至父节点, 转到 Step2;

Step2: 内部节点中, 从左向右接受子节点传递过来的所有未关联的节点, 将这些节点有序插入到文本字符串中。同样, 按照步骤 Step1 的方法将这些节点关联起来, 将所有未关联的节点传递至父节点, 转到 Step3;

Step3: 如果不是根节点, 继续步骤 Step2, 否则

停止遍历。

(2)“地名+相对位置”形式

一般情况下,网页文本中颗粒度较小、显著度较低的地名,其空间坐标往往处于缺失状态。此类地名的空间位置信息描述主要通过其与其他地名间的相对位置关系(即空间关系)实现。特别是,方向关系是人们日常生活中使用频率较高的空间关系类型。自然语言文本中,空间关系词汇对空间关系表达具有强烈的指示作用^[21],而且地名和地名的相对位置信息经常成对出现在网页文本中。例如,“五大淡水湖中的巢湖素为长江下游、淮河两岸的鱼米之乡。”描述中,“巢湖”、“淮河”和“长江”是地名,“下游”和“两岸”为描述相对位置信息的空间关系词汇。因此,本文在候选空间敏感网页文本中,以句子为单位,采用地名和空间关系词汇就近匹配的方式获取参照物地名和目标物地名之间的相对位置信息。其中,空间关系词汇的获取通过空间关系词汇词典匹配实现(见表 1)。

表 1 空间关系词汇词典示例
Tab. 1 Examples of spatial relation terms

空间关系类型	空间关系类型	词汇样例
拓扑关系	包含/包含于	分支,支脉,并入,合并,隶属,直辖,划归
	相接	起点,发源地,濒临,相邻,紧依
	相离	相离,相隔,相距,相间,隔,相望
	交叠	横贯,横渡,流贯,纵贯,贯穿
	相等	相等,语意为,改为,改置,别名
	围绕	环绕,围绕,界山,边境线,界线
方向关系	前	前端,前头,面前,顶上,前,前边
	后	后端,后头,背后,底下,后,后边
	左	左端,左头,左岸,左下头,左前端
	右	右端,右头,右岸,右下头,右前端
	上	上游,上源,上端,上头,上,上边
	下	下游,下游,下端,下头,下,下边
	内	内,内部,内侧,以内,里,里边
	外	外,外部,外面,外边,外头,外界
	东	东面,东方,以东,之东,东端,东
	西	西端,西头,西侧,西源,西边,西
	南	南部,南支,南段,南麓,南隅,南
	北	北面,北方,北东东,北西西,以北
	中	中部,中心,中游,中上游,中下游
	东北	东北面,东北方,东北,东北偏东
	东南	东南面,东南方,东南,东南偏东
	西北	西北部,西北境,西北角,西北头
	西南	西南部,西南境,西南麓,西南端
距离关系		距离,离,相隔,相离,相距

4 应用实例

本文选用“南京师范大学仙林宾馆+西北”为空间检索词实例,利用上文所述的空间敏感网页获取和解析方法及网页文本中地名信息解析方法,进行地名数据库的更新维护。

首先,基于 Google 搜索引擎服务,以“南京师范大学仙林宾馆+西北”为空间检索词,实现空间主题网页聚焦,将获取的网页作为候选空间敏感网页,定制候选页面的数量为 100,见图 3。

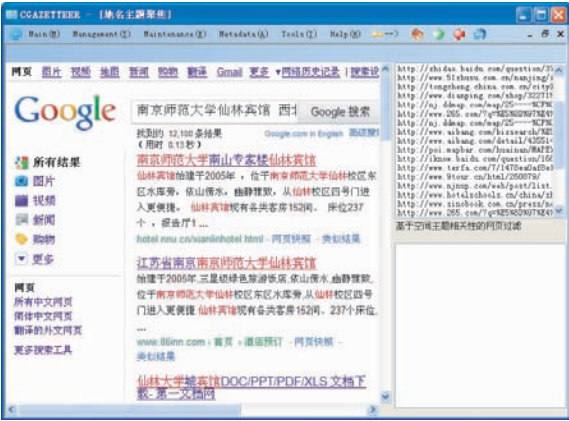


图 3 空间敏感 URL 获取

Fig. 3 Obtaining of space-sensitive URLs

然后,在候选 URL 页面纠错、标准化、分词等预处理的基础上,计算待选 URL 页面与空间检索词“南京师范大学仙林宾馆+西北”的空间主题相关度,并根据相关度过滤待选 URL 页面。各参数取值情况为:空间检索词的个数 $n=3$,地名检索词的权重 $w_1=0.3$ 、 $w_2=0.3$,空间关系词汇的权重 $w_3=0.4$,待选 URL 页面正文和标题的空间主题相关度权重 $\alpha=0.4$ 、 $\beta=0.6$, α 、 β 可以进行适当调整。候选 URL 网页过滤后,其前 50 位 URL 参见图 4。

在采用 DOM 技术实现空间敏感网页树形结构解析的基础上,利用 CRF 地名识别模型识别网页文本中地名,并进行地名有效性检验,图 5 中的“/LOC”标签表示识别后的地名。地名识别结果,采用地名数据库匹配的方式,从网页文本中获取新地名(见图 5);采用网页文本中地名信息解析的方法,获取地名空间位置信息,即地名的地理坐标和相对位置信息(参见图 6)。最后,将获取的新地名和地名的相对位置信息存入地名数据库。

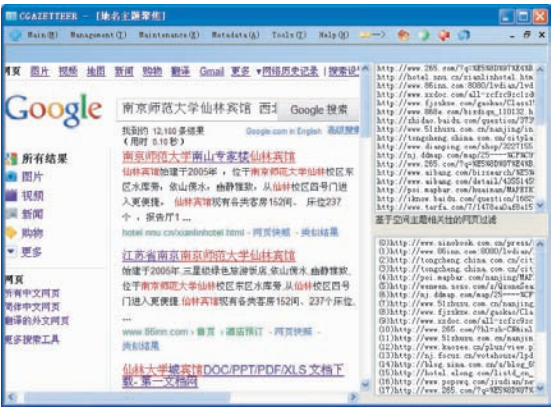


图 4 空间敏感网页过滤

Fig. 4 Filtration of space-sensitive web pages



图 5 基于 CRF 的地名识别

Fig. 5 Place name recognition based on CRF

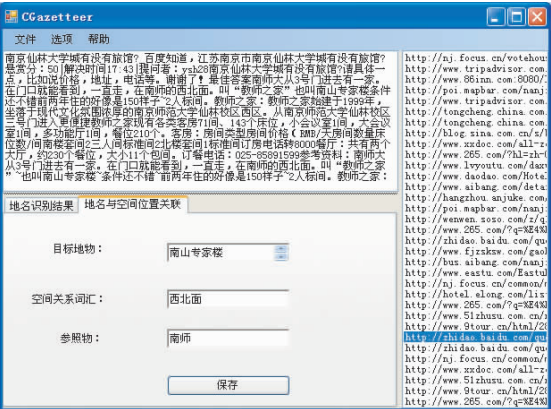


图 6 地名空间位置信息获取

Fig. 6 Obtaining of place name location information

从上述实例可以看出,以“南京师范大学仙林宾馆+西北”为空间检索主题,基于 Google 搜索引擎获取的相关网页涉及领域范围较为广泛。经过空间敏感网络爬虫的主题相关性预测、判断和筛选

后,实现了网页的空间主题聚焦,从而为网页文本中地名信息的解析提供了准确且针对性数据源。应用地名识别技术和空间位置信息解析方法,可以有效获取网页文本中新地名(特别是小颗粒度地名)和地名空间位置信息,实现地名数据库的更新维护。例如,通过“南京师范大学仙林宾馆+西北”的空间检索主题,可以捕获“南山专家楼”和“南师大”的相对位置关系。同时,如果直接以小颗粒度地名(如“南山专家楼”)为空间检索主题,则可以更加针对性地获取该地名的空间位置信息。

5 结语

针对目前地名数据库建设存在更新维护周期长、成本高、效率低等问题,探讨了一种网络爬虫的地名数据库维护方法。该方法基于 Google 搜索引擎服务,以网页资源为数据源,利用网络爬虫技术获取海量空间敏感的网页文本,应用地名识别技术进行网页文本中地名识别和地名空间信息解析,从而更新维护地名数据库。实例分析表明,该方法可以有效、低耗地获取网页中新地名及其空间位置信息,为地名数据库更新维护提供了一种新方法。

对此,今后研究工作将集中在如下两个方面:(1)挖掘网页文本中地名的时间属性及非时空属性,如网页的发布时间和网页中因事件驱动而致使地名发生变化的时间,可能带来地名信息的滞后问题,难以保证地名数据库的完备性与一致性。(2)基于互联网地图的地名及其空间位置信息获取。近年来,GoogleMap、GoogleEarth、OpenStreetMap 等公众参与性互联网地图可以提供实时、准确的地名数据来源,特别是地名的空间坐标信息,同时,都提供开放的 API 接口,可以互联网地图为数据源进行地名及其空间位置信息获取,并与网页文本中地名信息融合,实现地名数据库快速、实时的更新维护。

参考文献:

[1] Goodchild M F, Hill L L. Introduction to Digital Gazetteer Research[J]. Geographical Information Science, 2008, 22(10):1039 - 1044.

[2] 张雪英,张春菊,闫国年. 地理命名实体分类体系的设计与应用分析[J]. 地球信息科学学报,2010,12(2):220 - 227.

- [3] 陈钻, 万庆, 吴杰. 基于 XML 的无线位置服务地理信息服务器的实现[J]. 地球信息科学, 2004, 6(4): 100 - 104.
- [4] <http://www.alexandria.ucsb.edu/adl/>.
- [5] <http://nhd.usgs.gov/gnis.html>.
- [6] <http://www.ga.gov.au/place-name/>.
- [7] 狄琳, 欧阳宏斌. 全国 1:25 万地名数据库的设计与建立[J]. 测绘通报, 2010, 10: 32 - 33.
- [8] 陈春华. 1:5 万地名数据库到 1:1 万地名数据库转换的研究与开发[J]. 测绘通报, 2006, 5: 71 - 72.
- [9] 张保钢, 杨伯钢, 孔俊元. 北京市地名数据库的维护更新[J]. 北京测绘, 2010, 3: 28 - 30.
- [10] Palkowsky B and MetaCarta I. A New Approach to Information Discovery-Geography Really Does Matter [C]. In Proceedings of the SPE Annual Technical Conference and Exhibition, 2005.
- [11] Hill L L. Core Elements of Digital Gazetteers: Place Names, Categories, and Footprints[C]. Research and Advanced Technology for Digital Libraries. Berlin, Germany: Springer, 2000, 280 - 290.
- [12] 李金良, 张雪英, 樊晓春. 汉语地名时空信息一体化表达[J]. 地理与地理信息科学, 2010, 26(6): 6 - 10.
- [13] 陈丛丛. 主题爬虫搜索策略研究[D]. 山东大学, 2009.
- [14] 李勇, 韩亮. 主题搜索引擎中网络爬虫的搜索策略研究[J]. 计算机工程与科学, 2008, 30(3): 4 - 6.
- [15] 陈财森, 王韬, 郑伟. 基于搜索引擎调用的主题搜索设计与实现[J]. 计算机工程与设计, 2008, 29(21): 5627 - 5629.
- [16] Diligenti M, Coetze M, Lawrence S, *et al.* Focused Crawling Using Context Graphs[C]. In Proceedings of the 26th International Conference on Very Large Data-Bases, Cairo, 2000, 527 - 534.
- [17] 刘秉权, 王喻红, 葛冬梅, 等. 基于结构树解析的网页正文抽取方法[C]. 黑龙江省计算机学会 2007 年学术交流年会, 2007, 14 - 17.
- [18] 周俊生, 戴新宇. 自然语言信息抽取中的机器学习方法研究[J]. 计算机科学, 2005, 32(3): 186 - 199.
- [19] 张小衡, 王玲玲. 中文机构名称的识别与分析[J]. 中文信息学报, 1997, 11(4): 21 - 32.
- [20] 王志强. 基于条件随机域的中文命名实体识别研究[D]. 南京理工大学, 2006.
- [21] Bowerman M. Learning How to Structure Space for Language: A Cross linguistic Perspective [M]. // Bloom P, *et al.* (Eds.). Language and Space, Cambridge, MA, USA: MIT Press, 1996, 385 - 436.

Method of Toponym Database Updating Based on Web Crawler

ZHANG Chunju, ZHANG Xueying, ZHU Shaonan, XU Xitao

(Key Laboratory of Virtual Geographical Environment, Ministry of Education,
Nanjing Normal University, Nanjing 210046, China)

Abstract: Generally, toponym database provides description information on place names and its spatial location and feature type. It provides basic information for national administration, economic development, domestic and foreign exchanges, etc. It is a basis for public place name services, particularly for Location-Based-Service (LBS) with a growing demand. Therefore, a toponym database with complete and timely place name information is a premise and guarantee for efficient LBS services. However, currently, there are some problems about place names in our national toponym database. Most of the place names are with a big particle size, and small particle sized and non-standard place names are in shortage, and there are no relative position descriptions of place names in toponym database. Moreover, toponym database updating is based on manual surveying with disadvantages of long cycle, high cost, low efficiency and time consuming. In this paper, a new method for toponym database updating is explored on the technology combination of search engine, web crawler and place name recognition. Firstly, a mass of space-sensitive web pages are obtained by a web crawler which is based on Google search engine and a spatial search subject of "place name" or "place name + spatial relation terms". Secondly, after analysis of web pages with a DOM tree method, place name recognition is completed based on Conditional Random Fields (CRF) recognition

model. Finally, automatic spatial location interpretation of place names is completed from candidate web texts which include new place names and spatial location information of place names. This paper also presents a case study with a spatial search subject of “Nanjing Normal University, Xianlin hotel + north-west”. The experiment result shows that this method is feasible and effective. However, timely and accurately locating of place names in web pages are in challenge, because publishing time of web pages and change time of place names driven by events in web pages are not considered in this paper. This may result in potential lag of place name information and can’t ensure the completeness and consistency of toponym database. In recent years, public participation internet maps can provide accurate and real-time place name source, especially coordinate information, such as GoogleMap, GoogleEarth, OpenStreetMap, etc. Our future work will focus on time attribute interpretation of place names from web pages and obtaining of place names as well as their coordinates from internet maps. Moreover, an integration of place names from different data sources will provide a more effective toponym database updating.

Key words: toponym database; web crawler; place name recognition; theme correlation