面向网络的空间信息提取系统研究

毛 曦 李 琦 刘 帅 朱亚杰

(北京大学地球与空间科学学院 北京 100871)

摘 要 随着网络技术的不断发展,互联网已经成为一个海量、复杂多样的数据源,特别是随着Web2.0与社交网络的兴起,每个网民都可视为一个空间传感器,其源源不断地将周围的空间信息发布在网上,互联网中的空间信息日益丰富。提出了面向网络的空间信息提取系统,在从Web页面中所包含的半结构文本或自由文本中识别出完整位置的基础上,提取出与该位置相关的专题属性信息,并将其结构化和空间化。通过系统实例的研究,验证了本系统的可行性。

关键词 空间信息提取,信息提取,地址地理编码

中图法分类号 TP391 **文献标识码** A

Web-oriented Spatial Information Extraction System

MAO Xi LI Qi LIU Shuai ZHU Ya-jie

(School of Earth and Space Science, Peking University, Beijing 100871, China)

Abstract With the development of network technology, the Internet has become a complex and diverse data sources, especially the rise of Web2.0 and social networking, each Internet users can be seen as a spatial sensor from spatial information angle, massive spatial information surrounding them has been published on the Internet. Spatial information on the Internet becomes increasingly rich. This paper presented the Web oriented spatial information extraction systems. In this system, we recognize geographic sense address from non-structured data in the Web pages. Then we extracted attribute information related to that address, and map this attribute information to spatial entity by address geocoding. Finally we proposed a prototype to prove the feasibility of the system.

Keywords Spatial information extraction, Information extraction, Address geocoding

随着网络技术的不断发展,互联网已经成为一个海量的、复杂多样的数据源[1]。特别是随着Web2.0与社交网络的兴起,人们已经不再仅仅是网络信息的消费者,更是网络信息的提供者,通过微博或社交网站等工具,每个网民可以很方便地将自己在某地所看或所拍的事或物发布到网上。从空间信息学的角度来看,每个网民都可视为一个空间传感器,其源源不断地将各自周围的空间信息发布在网上。因此,这部分以非结构化网页形式存在的空间信息也已逐渐成为空间信息的一个重要来源。著名地理信息学家Michael Goodchild教授指出,通过社交网络与众包(crowdsourcing)活动所产生的地理空间数据将足以和专业数据相媲美,它将解决许多传统空间信息获取技术无法解决的问题[2]。Flickr网站便是一个很典型的例子,用户可以看到很多关于某一个地点的有价值的图文描述,而这些信息都是全球网民们通过网站的地理标注功能(Geotagging)自愿上传的。因此,面对网络中所存在的如此丰富的空间数据,如果我们能够利用信息提取技术从中抽取出所需的空间信息,将极大地提高空间信息获取的效率,大大降低空间专题属性信息获取的成本。

在众多的空间信息定义中,公认的观念是:空间信息通常包含3个方面的内容:空间位置信息、专题属性信息、时间信息。简单地说,空间位置信息描述的是在哪里,专题属性信息描述的是该位置有什么地物或发生什么事情,时间信息则表示在什么时间。因此,相对于其他的信息,空间信息最大的特点是与实际的地理位置密切相关。现有的信息提取算法虽然在命名实体识别、句法分析、篇章分析与推理等方面有了长足的发展,为从非结构化的网页数据中提取空间信息提供了强有力的技术支撑,但是还不能完全满足于空间信息的提取,具体表现在:

1. 现有的信息提取中命名实体识别技术的作用对象是词语和短语,而空间信息提取的对象则是具有完整地理意义的文本串。比如,对于"北京市海淀区颐和园路5号"这个短语,现有的命名实体识别可能只将"北京"、"海淀"、"颐和园"这几个词作为地名识别出来,而不能将整个短语作为一个详细地址进行识别。

2. 现有的信息提取中所抽取的信息还是以文本的形式表示,而空间信息提取要求所抽取的信息是和具体的地理坐标绑定的,而且在不同空间尺度下具有不同的空间形态。比如在小尺度的地图中"学校"显示的是一个点,而在大尺度的地

本文受国家863项目(2009AA122101)资助。

毛 曦(1983-),男,博士后,主要研究方向为空间信息搜索引擎等,E-mail:maoxi.1122@gmail.com;李 琦(1955-),女,教授,主要研究方向为数字城市等;刘 帅(1982-),男,博士生,主要研究方向为数字地球与数字城市等。

图中则显示的是个面。

因此，本文提出了面向网络的空间信息提取系统，在从Web页面中所包含的半结构文本或自由文本中识别出完整位置的基础上，提取出与该位置相关的专题属性信息，并将其结构化和空间化。

1 系统流程

针对空间信息在非结构化数据中的分布特点以及空间信息的内容，将整个空间信息提取的过程分为网页数据库构建、空间位置信息识别、专题属性信息提取以及地址地理编码这4个步骤来进行，如图1所示。

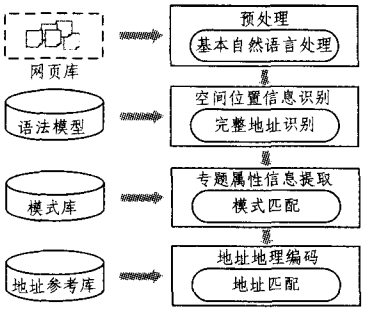

图1 系统流程

1)网页数据库构建：建立网页数据库，对这些网页文档进行分词处理，统一段落格式，并提取出文档基本信息；

2)空间位置信息识别：对网页文档进行空间命名实体的识别，从文档中提取出描述专题属性信息的完整的地址和时间；

3)专题属性信息提取：根据专题属性信息模型的要求，从网页文本中抽取出专题属性信息的每个侧面，如餐厅人均消费的具体数值；

4)地址地理编码：将以文本形式存在的空间位置映射成实际的地理坐标，实现专题属性信息的空间化。

1.1 网页数据库的构建

空间信息提取的数据源通常需要包含空间信息且内容真实、权威，这就要求网页文档搜集算法有较高的准确率且能分析网站数据的真实性。搜索的准确性可以采用垂直搜索保证，但是，垂直搜索需要一些种子链接作为搜索的起点，这些种子链接是搜索准确性的保证；至于网页数据的真实性和权威性，一般采用PageRank算法来衡量，即该网页被引用的次数越多那该网页的数据就越具有真实性。类似，论文被引用的次数越多就越权威，但是，计算PageRank需要一个庞大的网页数据库和超级计算资源。

因此，为了解决上述的问题，我们采用垂直搜索和元搜索相结合的方式进行网络文档的搜集。基本流程为：通过元搜索获取空间信息主题关键词在Google等搜索引擎中返回的结果中排名靠前的网站，因为Google等搜索引擎返回的结果是按照PageRank算法来排序的，它返回的前几个结果具有较高的可信度；然后以这些网站的链接作为种子链接通过垂直搜索进行搜索。

1.2 空间位置信息识别

空间位置信息识别的主要难点在于如何能够识别一段完整的地址，而不是仅仅判断一个词是否是地址或地名。针对现有命名实体在空间位置信息识别上的不足，实现准确的空间位置信息识别需要研究新的识别方法。本系统在现有命名实体识别的基础上，结合完整中文地址语法规则，提出了完整地址地名识别算法，具体的实现方法在2.2节中阐述。

1.3 专题属性信息提取

专题属性信息在网页数据中的分布存在分散、噪声大、规律不显著等特征，给属性信息的提取带来了不小的难度。按照内容来说，专题属性信息一般可分为实体属性与事件属性两大类。为了能够尽可能充分、正确地从网页数据中提取专题属性信息，我们借鉴了一般信息提取的技术。针对这两类属性信息，设计了实体属性信息提取模型与事件属性信息提取模型，分别用来描述对这两类信息提取所关心的侧面。实体属性信息提取模型主要描述常用兴趣点的基本构成要素，比如，针对“餐厅”这个兴趣点我们一般需要提取“人均消费”、“菜系”等内容；事件属性信息提取模型主要描述所发生事件的属性信息，比如，某事件的时间、人物等。专题属性信息提取模型中所包含的每个要素都是专题属性信息提取所要抽取的信息。

针对提取模型中的每个要素我们总结出一些常见的提取模式。这些模式用于匹配句子中的语言结构的一组语法规则，由固定词汇和提取变量所组成，其中的提取变量就是我们所关心的信息。表1所列的是关于突发事件伤亡人数的部分提取模式实例。系统根据这些提取模式完成对于专题属性信息模型中各个要素的提取。因此，在本系统中，专题属性信息提取的基本流程为：

1)对于信息提取模型中每个要素的所有提取模式，在文本中定位模式中的固定词汇，如果匹配成功，转到步骤2)；否则，转到下一个提取模式。

2)根据该提取模式，从文本中提取相应的值赋给变量，完成一次提取，并回到步骤1)。一般来说，我们主要关心的是名词性短语，因为大多数我们感兴趣的专题属性信息多是用名词短语或数值来表示的。

表1 专题属性信息提取模式

类别	模式	示例
突发事件人数	[数字] 人 死亡	1000 人死亡
	死亡 人数 为 [数字]	死亡人数为 500
	造成 [数字] 人 死亡	造成 100 人死亡
	死亡 人数 上升 至 [数字]	死亡人数上升至 100

1.4 地址地理编码

通过空间位置信息识别与专题属性识别所获取的空间信息都以文本的形式存在，还没有空间化。需要利用地址地理编码技术将这些所识别的中文地址映射成实际的空间实体(点、线、面)，从而实现专题属性信息与实际地理坐标的绑定。我国历史悠久、公众使用习惯和人文文化变化等导致中文地址表述比较复杂。在人们的日常地址表述中除了使用“行政区划+街道+门牌号”的标准地址外，还存在大量相对表述方式和使用具有醒目标牌的标志性地物的模糊地址表述方式，例如，“中关村附近”、“海龙大厦向东100米”等等。这些复杂的表述方式给中文地址地理编码带来了很大的难度。因此，针对这两类地址我们分别设计了相应的地址地理编码方式。

对于标准的地址进行地址地理编码的过程实质上是地名库检索的过程，其基本流程为：首先在标准地址地名库中查找待编码地址，如果有正确匹配的则返回该地址空间范围，如果

没有正确匹配的则进行地址插值。在插值方面，主要使用基于地址门牌和位置相关性的插值方法来进行[3]，首先判断待插值门牌的奇偶(判别在街道哪侧)，然后找到与它临近的已知位置的门牌号码进行内插定位。

至于模糊描述的地址，可以看到这些地址一般包括空间定位词和空间方位词。模糊描述地址的地理编码算法的基本过程为：根据地址地名库在地址中识别出空间定位词，然后根据空间方位词确定该地址和空间定位词之间的空间关系；最后，计算出该地址的空间范围。

2 关键技术实现

2.1 中文标准地址地名库的构建

中文标准地址地名库是本系统的基础，空间位置信息识别与地址地理编码都依赖于地址地名库。中文标准地址地名库包含中文地址与空间数据两部分内容，输入地址关键词返回最匹配的地址及其空间范围。因此，在中文标准地址地名库的建立过程中，我们充分借鉴文本数据库和空间数据库的技术来进行构建。

1)中文标准地址地名库概念模型

在国外，地址一般都可以用“号码＋道路＋城市”的线点结构表示，然而，我国城市地址的各个组成部分分别由不同政府部门管理，规划之间没有整体的协调性。目前我国没有任何一个政府部门有地址命名的完整权限，也没有负责地址标准化和规范化的政府部门[4]。这些都给标准地址地名库的构建造成了不少的困难。在借鉴国外点-线模式地址的基础上，结合中文地址的特点，在本地址库中，采用空间层次模型对中文地址进行描述(见图 2)。在该模型中，根据空间上的层次关系，将标准地址共分成 6 个级别，即国家、省、市、区县(乡镇)、街道、门牌(楼)号，上一个层级地址的空间范围包含下一个层级地址，每个标准的地址至少包含 1 个级别的地址，并且都对应着相应的空间范围。此外，中文地址中还存在一些专属地名，比如，“泰山”等。由于这类地址通常直接与特定空间实体相对应，因此我们将其放在最后一层。

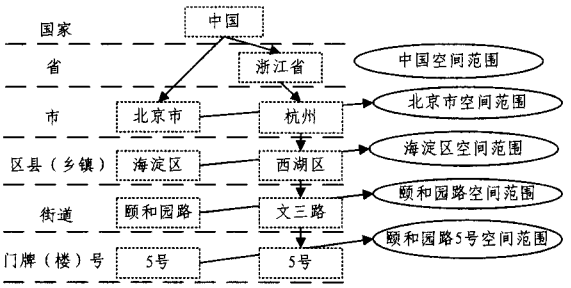

图 2 中文地址空间层次模型

2)地址地名索引

中文地址都以文本的形式存在，文本倒排索引被广泛地应用于文本检索。因此，中文地址库索引的基本思路是在一般文本倒排索引的基础上再增加所对应的空间范围，具体结构由 3 个部分所组成，分别是地址关键词列表、空间范围列表和地址地名列表(见图 3)，每一个地址关键词对应该关键词的地址或地名，而每个地址或地名对应着该地址地名所表示的空间范围。

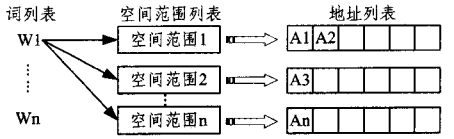

图 3 地址地名索引

3)地址匹配算法

在地址匹配算法中，主要借鉴空间向量模型(VSM)中文档相似度作为地址匹配的度量。对于所输入地址，首先计算输入地址与地址库中地址地名的相似度；然后，返回前 N 个最近似的地址地名；最后，从地址地名索引中返回这些地址地名所对应的空间范围。

2.2 完整地址识别技术

所谓完整的地址，是指一段能够完整描述空间位置或方位的文本串，能够被映射成具体的空间实体或范围，可以是词也可以是短语。现有的地名识别技术主要是判断某个词是否是地名，因此，其所识别出来的地名不能完整地描述文本中所要表达的真实空间实体。例如，“北京大学附近”这个短语，现有的地名识别方法一般将北京大学作为地名或机构名识别出来，而忽略了“附近”，然而，“北京大学”和“北京大学附近”在空间范围上具有不小的差别。

针对上述问题我们看到，无论是标准的中文地址或模糊描述的地址都具有常用词汇，将这些词称之为地址特征词汇，如“市”、“街”等；而且，特征词汇相互之间具有很强的相关性。因此，在本系统中，在统计出中文地址特征词汇的基础上，建立语法模型，然后，找出地址特征词汇间最优的组合作为完整地址。

1)地址统计语法模型的构建

由于地址特征词汇的相关性，本系统采用 n-gram 模型来描述地址语法规律。首先，根据语料库中的地址统计出中文地址的常用特征词汇，根据地址特征词汇所对应的空间范围，将其分为点域、线域、面域、空间关系这 4 种类型，表 2 所列的是部分地址特征词汇；然后，统计出各个类别中地址特征词汇两两之间的共现概率。

表 2 中文地址常用特征词汇

类别	词汇
面域	省、自治区、市、区、乡、镇、村、庄、里、小区
线域	街、路、道、巷、弄
点域	幢、栋、楼、号
空间关系	附近、以(东、南、西、北)、对面、交界处

2) 基于统计语法模型的地址识别流程

根据以上建立的中文地址统计模型，空间位置识别的过程是一个计算文本中所出现的地址特征词间最优组合的过程。具体为：

(1) 在文本中查找到地址特征词，如果找到，则将该词设为当前词汇；如果没有，则停止；

(2)查看当前词汇后面的词是否是地址特征词，如果不是，则执行步骤(3)；如果是，地址特征词则计算两个词的共现概率，并将该词组作为当前词汇继续执行步骤(2)；

(3)比较所计算的所有概率，返回最大的组合作为所识别的中文地址，执行步骤(1)。

(下转第 264 页)

IEEE Transactions on Computational Intelligence and AI in Games, 2010, 2(1): 17-26

[3] Littman M L. Friend-or-foe q-learning in general-sum games[C]// Proceedings of the Eighteenth International Conference on Machine Learning. Williams College: Morgan Kaufman, 2001: 322-328

[4] Greenwald A, Hall K, Serrano R. Correlated-q learning[C]// Proceedings of the Twentieth International Conference on. Washington DC, 2003: 242-249

[5] 赵凤强，徐毅，李广强. 基于岛屿群体模型的多目标演化算法研究[J]. 计算机科学，2010，37(12)：190-192

[6] 宋梅萍，顾国昌，张国印，等. 一般和博弈中的合作多 agent 学习[J]. 控制理论与应用，2007，24(2)：317-321

[7] HuJun-ling, Wellman M P. Nash Q-learning for general-sum stochastic games[J]. Journal of Machine Learning Research, 2003, 4(11): 1039-1069

[8] Vassiliades V, Cleanthous A, Christodoulou C. Mutltiagent Reinforcement Learning: Spiking and Nonspiking Agents in the iterated Prisoner's Dilemma[J]. IEEE Transactions on Neural Networks, 2011, 22(4): 639-653

[9] Aumann R J, Hart S. Computing equilibria for two-person games. Handbook of Game Theory with Economic Applications [R]. Amsterdam: Elsevier, 2002

[10] Murty K G. Computational complexity of complementary pivot methods[C]//Mathematical Programming Study 7 Complementarily and Fixed Point Problems. Amsterdam: North-Holland Publishing Co, 1978: 61-73

[11] Howard N. Paradoxes of Rationality[M]. Theory of Metagames and Political Behavior, MIT Press, Massachusetts: Cambridge, 1971

[12] Thomas L C. Games, Theory and Application [M]. Halsted Press, Chichester, 1984: 129-149

[13] Cao Yong, Li Ren-hui. Conflict Analysis between Tacit Knowledge Sharing and its Exclusivity Based on Meta-game Theory[C]// 2009 International Symposium on Information Engineering and Electronic Commerce. IEEC, 2009: 31-35

[14] Sharma R, Gopal M. Hybrid Game Strategy in Fuzzy Markov-Game-Based Control[J]. IEEE Transactions on Fuzzy Systems, 2008, 1(16): 1315-1327

[15] Milinski M, Sommerfeld R D, Krambeck H J, et al. The Collective-risk Social Dilemma and the Prevention of Simulated Dangerous Climate Change[J]. Proceedings of the National Academy of Sciences, 2008, 105(7): 2291-2294

(上接第 231 页)

3 应用实例

基于本文的思路和涉及到的关键技术，在一系列 CyberSIGStudio 服务器[6-8]的支撑下，设计并实现了面向网络的空间信息提取原型系统。下面以“地震事件”为例来阐述本系统的功能。

用户对话框中输入“地震事件”，系统将所提取的地震实例在页面的左侧显示，并在页面右侧的地图中标出事发的地点(见图 4)。在图中，左侧列表显示的结果是系统从网络中获取到的最近发生的 3 项地震实例。

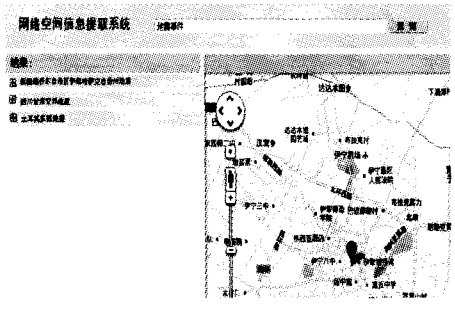

图 4 空间信息提取结果示意图

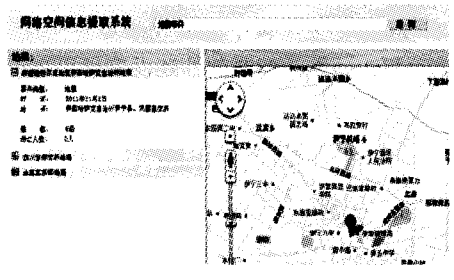

图 5 空间信息提取结果示意图

当用户点击页面左侧的具体事例时，系统将显示该事件的具体侧面，包括时间、伤亡人数、震级等信息。图 5 展示是当用户点击“新疆维吾尔族自治区伊犁哈萨克自治州地震”时，该事件的详细信息被展示出来。

结束语 本系统充分挖掘了网络中存在的大量空间信息，增加了空间数据获取的有效途径，能够有效地降低空间数据获取的成本，为空间数据服务系统提供技术与数据的支持。针对空间信息在网络中存在的特点，本系统分别提出了空间位置信息识别算法与属性信息提取算法，利用统计模型提高了地址识别的准确度，并进一步利用地址地理编码技术将其空间化。

在地址地名库的构建当中，如何有效地对复杂多样的地址地名进行建模与索引是需要进一步研究的课题。由于地址地名库中的标准地址都是短语文本，因此研究更加高效的匹配方法十分必要。此外，研究如何适应不同比例尺的地址地理编码技术也很有意义。

参 考 文 献

[1] McCurley S. Geographic Mapping and Navigation of the Web [C]// The 10th www Conference. Hong Kong, 2001

[2] http://p2pfoundation.net/Michael_Goodchild_on_Volunteer_Mapping's_Role_in_Geospatial_Science

[3] 朱建伟，王泽民. 地理编码原理及其本地化解决方案[J]. 北京测绘，2002，12(2)：24-27

[4] 马皓明. 中文地址地理编码研究与原型系统实现[D]. 北京：北京大学，2007

[5] Ricardo B. Modern Information Retrieval[M]. New York: Addison Wesley, 1999: 110-112

[6] 林绍福，李琦，董宝青. 数字城市应用服务平台体系结构研究[J]. 计算机科学，2002，29(12)：98-102

[7] 李琦，甘杰夫. 数字城市空间信息与服务集成交换平台系统分析与设计[J]. 计算机科学，2005，32(9)：123-126

[8] 史文勇，李琦，林宇. 数字城市核心系统平台的服务总线设计[J]. 计算机科学，2006，33(3)：279-282