高技术通讯 2012 年 第22 卷 第2 期:165 ~ 170

doi:10.3772/j.issn.1002-0470.2012.02.009

# 基于语义的文本地理范围提取方法①

张 毅② 王星光 陈 敏 刘 瑜③

(北京大学遥感与地理信息系统研究所 北京 100871)

**摘 要** 为了能够处理网页文档中的地理信息,提出了一个新颖的自动提取文本地理位置的方法。该方法通过一个三阶段的地理语义处理过程,实现了文本的多尺度地理标注。首先,在地理知识库的支持下,识别文本中的地名;其次,基于地理的和非地理的语义消除地名歧义并且应用证据理论合成排歧证据;最后,基于相关认知理论构建文本的地理参照树,再根据实体间的语义关系计算得到焦点地理实体,从而确定文本的地理位置。以上算法在地理信息检索原型系统 GeoSeracher 中得到实现,评估结果表明其具有较高的准确度。

**关键词** 地理信息检索(GIR),文本地理范围,证据理论

## 0 引 言

搜索引擎用户的查询需求常常与地理位置有关,约 14.8% 的查询使用地名对检索进行限制[1,2],这类查询称为地理相关查询。但目前这类查询的结果往往不理想,主要问题是:(1)地名歧义造成误检,例如,检索“伦敦”时,那些内容包含了“伦敦”(全世界有 30 多个)的网页都会被检索到;(2)语义关系的忽略导致漏检,例如,检索“上海”时,那些内容中只包括“沪”或者“申”而不包括“上海”的网页会被漏检;(3)无法理解用户查询,例如,搜索引擎不能理解“XX 以东”、“附近 XX 米以内”之类的查询,从而返回无意义的网页。造成以上问题的根本原因是目前的搜索引擎无法理解词汇所表达的语义,它既不懂用户的查询意图,也不理解网页的内容,只是根据关键词匹配进行查询处理。为了提高搜索质量,近十年研究者提出和发展了地理信息检索(geographic information retrieval, GIR)技术[3]。和全文检索不同,GIR 能够在语义层次上处理用户的地理相关查询。其基本思路是:首先识别和分析文本中的地理信息以确定用户查询和网页文档的地理范围,然后根据查询和网页在空间上的相关程度返回和排序检出结果。显而易见,提取文本的地理范围是 GIR 最重要的技术。

对文本进行地理标注必须消除地名歧义和能够从多个地名中确定文本的地理位置。关于地名解歧,目前的方法可分为两类:(1)基于地理知识库提供的相关知识实现解歧;(2)基于机器学习的方法消除歧义。由于缺乏足够的训练集,GIR 广泛使用启发式规则消除地名歧义。最常见的简单规则是缺省规则,即选择缺省地理实体作为歧义地名所指[4-12]。缺省实体通常由地理实体的重要程度决定,而影响重要性的因素有人口[6-9]、类别[9-12]等。常用的复杂规则有地理相邻规则和空间包含规则[4,6-11,13],它依据的假设是同一文本中出现的地理实体通常在空间上相近。关于文本的地理参照,Woodruff 和 Plaunt 在 GISPY 系统中实现了一个基于多边形叠加的文本地理范围提取方法[14],该方法首先识别出文本中的地名;其次得到每个地名所有可能所指实体的地理范围,表示为多边形,并且多边形的高度由指称它的地名在文本中出现的次数决定;然后将所有多边形叠加在一个空白栅格地图上;最后叠加结果的最高部分就代表了文本的地理范围。Silva 等人则认为,文本的地理范围就是文本中最重要地理实体的地理范围,地理实体的重要性是由地理实体之间的空间关系以及它们在文本中出现的次数决定的。并且他们发展了一个基于 PageRank 的提取算法,用于度量地理实体的重要性[9]。本文认为,提取文本的地理位置也就是从文本中提取地理语义的过程,类似于人的阅读活动。因此,如果在提

① 863 计划(2007AA120502)和国家自然科学基金(41171296)资助项目。

② 男,1971 年生,博士;研究方向:地理信息科学;E-mail: zy@pku.edu.cn

③ 通讯作者,E-mail: liuyu@urban.pku.edu.cn

(收稿日期:2011-06-20)

取算法中集成了相关认知理论,就能够提高文本地理参照的准确度。基于此观点,本文提出了一个基于语义的自动提取文本地理位置的方法,它是 GIR 原型系统 GeoSearcher 研究工作的一部分。

## 1 提取文本地理位置的基本流程

GeoSearcher 是我们自行开发的一个 GIR 原型系统,它能够在语义层次上处理用户的地理相关查询。GeoSearcher 有两个应用流程:检索流程和索引流程。其中,索引流程是指自动抓取 Web 中的网页,分析和提取网页中的地理语义并建立文本索引的过程;检索流程则包括从用户发送地理相关查询请求到查询结果返回给用户的全过程。

图 1 显示了 GeoSearcher 索引流程的软件架构。该架构首先通过网络爬虫从互联网中抓取网页并存储在数据库中,等待系统对其进行处理;然后语义分析器对预存的网页进行地理语义分析,从中提取主题特征和地理位置从而形成文本的文档视图;最后索引器基于文档视图分别建立空间和全文索引,以支持地理相关查询。

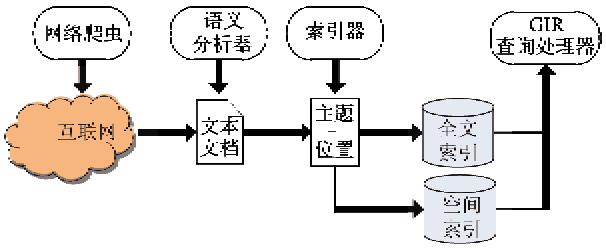


**图 1 对网页进行语义索引的软件架构**

GeoSearcher 语义分析器的主要功能就是提取文本的地理位置,图 2 显示了提取的基本流程。首先,对网页进行预处理,将网页文本的字符序列转化为词汇序列。预处理主要包括中文分词,排除停用词等。然后,识别和分析文本中的地理信息,以确定

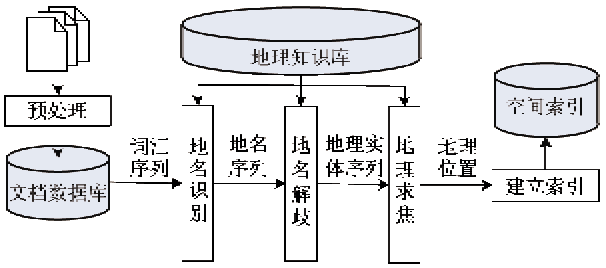


**图 2 提取文本地理范围的处理过程**

文本的地理位置。本文提出了一个三阶段的地理语义处理过程来实现文本的自动地理标注,即地名识别,地名解歧和地理求焦。

(1)地名识别

经过预处理后,文本表示为一个词汇序列。其中,词汇在序列中的顺序与它在文本中的次序相同。地名识别的任务是从词汇序列中识别和标注地名。目前存在两种实现方法:机器学习和基于地名库。GIR 多采用后一种方式。GeoSearcher 也是基于地理知识库确定文本中的地名。通过地名识别,文本的词汇序列表征转化为地名序列。其中,地名在序列中的顺序同它在文本中出现的次序一致。

(2)地名解歧

在文本的地名序列中,可能存在歧义地名。语义分析器是根据歧义地名上下文中的线索来消除歧义。所谓上下文线索是指地名序列中出现在歧义地名前后的其它地名所指的地理实体。GeoSearcher 通过分析歧义地名的所有可能所指和其上下文中其它地理实体之间的语义关系来确定歧义地名的实际所指。地理实体之间的语义关系是由外部地理知识库提供的。通过地名解歧,文本的地名序列表征转化为地名所指的地理实体序列。

(3)地理求焦

经过前两个处理过程,文本表示为一个地理实体序列。那么哪个地理实体才能代表整个文本的地理位置而用于空间索引呢? SPIRIT 采用的策略是对所有地理实体建立索引[5]。但是,显而易见不是每个地理实体都对文本起到地理参照作用。例如,“长江流经青海、四川、西藏、云南、重庆、湖北、湖南、江西、安徽、江苏、上海 11 个省市,是世界第三大河流,仅次于尼罗河和亚马逊河”。根据认知语言学中主体和背景的概念,在这段文本中,“长江”是主体词,而其它地理实体则是背景词,用于凸显主体词在某个方面的特征。其中,青海等 11 个省级行政区凸显了长江的地理位置,具有地理参照作用;而尼罗河和亚马逊河只是凸显了长江的长度。因此,本文基本采纳 Silva 等人的观点,即文本的地理范围是由文本中最重要的地理实体决定的,称其为文本的焦点地理实体,不同之处是本文认为只有起到地理参照作用的地理实体才能用于文本的地理参照。地理求焦处理就是通过分析文本的地理实体序列,根据它们之间的语义关系以及在文本中出现的次数,确定文本的焦点地理实体,从而获得文本的地理范围,用于建立文本的空间索引。

## 2　提取文本地理位置的算法实现

### 2.1　常识地理知识库和地名识别

地理信息系统以地图的观点对外部世界进行建模,是目前应用最广泛的空间数据处理工具。但是,这种基于定量地理坐标的软件系统难以处理定性的文本地理信息。人们能够依据人脑中的地理知识理解文本中的地理语义。根据地理认知研究,记忆中的地理知识是常识的、多态的、定性的、片段的、不完整的、不精确的,以及按照层次结构组织的。为了支持文本的地理语义处理,我们发展了一个常识地理知识库(common sense geographic knowledge base, CSGKB)。它以认知的视角表达外部世界,反映了人脑中的地理知识。首先,CSGKB 记录了组成地理空间的地理实体,表征为一组特征属性的集合,包括名称、类别、地理覆盖等。与传统地理信息系统不同,实体的地理覆盖是可选的并且可以是粗略的表示。其次,CSGKB 存储了地理实体之间的语义关系,包括上下位关系、相邻关系、穿越关系、重叠关系等。目前,CSGKB 大约记录了 100 万个地理实体以及 102 万个语义关系[15]。基于 CSGKB,GeoSearcher 通过简单查询实现了地名识别。

### 2.2　基于证据理论的地名解歧

#### 2.2.1　原理

不同地理实体出现在同一文本中的根本原因是它们之间存在着语义关联。GeoSearcher 地名解歧的基本原理是:如果歧义地名的某个所指和其上下文中其它地理实体的语义关联程度越高,则它是歧义地名实际所指的可能性就越大。

那么解歧需要考虑哪些语义关联呢? Leidner 总结了 15 条解歧规则[16],它们大致可以被分为 3 类:语法、语用和语义。语法是语言相关的,而本文主要考虑语义。常用的语义规则包括空间包含规则和地理相邻规则,它们所依据的理论是地理学第一定律[17],即"地表上的所有事物和现象在空间上都是关联的,距离越近关联程度越强,距离越远关联程度越弱"。地理关联可能是地名解歧中最重要的语义关联。但是,同一文本中地理实体之间的关联不仅仅只是空间上的联系。例如,前例中长江和尼罗河、亚马逊河同时出现不是因为它们在空间上相近,而是因为它们在某些方面相似,即都是世界级大河。因此除了地理关联,本文解歧方法还考虑了非地理语义关联。

#### 2.2.2　算法

经过地名识别处理,得到了文本的地名序列。它的内容包括地名、地名在文本中的位置以及地名的所有可能所指。如果一个地名的所指不唯一,那么它就是歧义地名。本文的排歧算法经过以下步骤:(1)确定歧义地名的上下文,它由歧义地名前后若干地理实体组成;(2)分别计算歧义地名的每个可能所指和其上下文中每个地理实体的语义关联度;(3)合成歧义地名的每个可能所指和其上下文中每个地理实体的语义关联度,得到每个可能所指和上下文的"总"的语义关联度(此步骤称为证据合成);(4)根据解歧原理,选择关联度最大的所指作为歧义地名的实际指称;(5)运用唯一意义规则(即同一个地名在同一论域中所指相同),将解歧结果扩散到地名序列中所有的相同地名。按照顺序依次对地名序列中的歧义地名进行解歧处理后,得到文本的地理实体序列。显而易见,解歧处理的两个关键问题是语义关联度的计算和证据的合成。

(1) 关联度计算

关联函数 $\mathrm{Rel}(g_1,g_2)$ 用于度量地理实体 $g_1$ 和 $g_2$ 之间的关联程度。它的值域在[0,1]之间,是由 $g_1$ 和 $g_2$ 之间的地理语义和非地理语义关联组合而成,公式为

$$\mathrm{Rel}(g_1,g_2) = a \cdot \mathrm{Rel}_g(g_1,g_2) + \sum b_i \cdot \mathrm{Rel}_{s_i}(g_1,g_2) \quad (1)$$

其中,$a$ 和 $b_i$ ($a>0, b_i>0, a+\sum b_i = 1$) 是权重系数; $\mathrm{Rel}_g$ 表示地理关联度,它由地理实体之间的空间关系决定; $\mathrm{Rel}_{s_i}$ 表示非地理语义关联度,例如类型关联、功能关联等。基于拓扑关系可以定量两个实体之间的地理关联程度[18]。相对而言,非地理语义关联度的计算困难。首先,选择哪些类型关联同文本的主题有关;其次,在缺乏相关知识库的情况下,关联程度难以定量。Hecht 和 Raubal 提出了一种基于维基百科的语义关联计算新方法,该方法可以用于度量公式(1)中的非地理语义关联度[19]。

(2) 证据合成

假设地理实体 $g$ 是歧义地名 $p$ 的一个可能所指,$\Omega$ 是 $p$ 上下文中所有地理实体的集合,$e$ 是 $\Omega$ 中任意一个地理实体,那么,地理实体 $e$ 和 $g$ 之间的关联度 $\mathrm{Rel}(g,e)$ 代表了 $e$ 提供的关于 $p$ 指称 $g$ 的证据,本文将其解释为可信度。上下文 $\Omega$ 提供的 $p$ 指称 $g$ 的证据则是融合了 $\Omega$ 中每个地理实体提供的证据。在歧义地名 $p$ 的所有可能指称中,同上下文

Ω 关联度最高的那个就是 $p$ 的实际指称。本文基于 D-S 证据理论[20,21]实现了解歧证据的合成，具体算法在文献[18]中进行了详细介绍。

### 2.3 基于地理参照树的地理求焦

#### 2.3.1 原理

如前例所示，有些地理实体是由于地理参照作用而出现在文本中。这些实体组合在一起形成了文本的地理参照系。根据相关地理认知理论[22-25]，参照系中的地理实体类型属于基本层次上的地理类别，主要包括各级行政区；参照系中的地理实体按照上下位关系形成了层次结构。本文将此参照系称为文本的地理参照树。其中，它的根结点是与文本内容相关的最上位行政区，通常是国家或省级行政区；从根结点开始，每层行政区比它的直接上层低一个级别；每个行政区结点的权重是其在文本中出现的次数；参照树中最具凸显作用（即最重要）的实体就是焦点地理实体，它代表了文本的地理位置。基于以上观点，地理求焦的基本方法是构建文本的地理参照树，计算参照树中每个地理实体的重要程度，从而确定文本的焦点，得到文本的地理位置。

#### 2.3.2 算法

根据地名解歧得到的地理实体序列，确定文本的地理位置经过三个处理步骤：构建地理参照树；计算地理实体的重要性；确定焦点地理实体。

（1）构建地理参照树

首先，选择文本中出现的行政区，然后，根据上下位关系构建文本的地理参照树。在这个过程中，需要解决参照树中的某些行政区没有出现在文本中和文本中的某些行政区与地理参照作用无关的问题。为了建立一个完整的地理参照树，在构建过程中如果文本中出现若干个同一层次的行政区，并且它们具有相同的直接上位行政区，但是这个上位行政区没有出现在文本中，那么在参照树上增加这个行政区；如果文本中出现的某个行政区同其它行政区在空间上相距很远，则这个行政区出现在文本中更有可能是因为其它原因而不是地理参照作用，所以将其排除在外。

（2）计算重要性

焦点地理实体是文本地理参照树中最重要的行政区，它在文本中出现的次数以及它同其它地理实体的空间关联是影响其重要程度的因素，它在文本中出现的次数越多则越重要，它同其它地理实体的空间关联越高则越重要。基于此，本文定义了一个度量函数 $GRank(e)$ 用于计算参照树中每个行政区 $e$ 的重要程度，如公式

$$GRank(e) = \sum_{e_0 \in super(e)} p(e_0) \cdot c(e_0) \cdot d^{s(e,e_0)} + \sum_{e_1 \in sub(e)} p(e_1) \cdot c(e_1) \cdot u^{s(e,e_1)} + p(e) \cdot c(e) \quad (2)$$

所示。其中，$super(e)$ 是指 $e$ 在参照树中的所有上级行政区的集合；$sub(e)$ 是指 $e$ 在参照树中的所有下级行政区的集合；$c(e)$ 是 $e$ 在文本中出现的次数；$p(e)$ 表示在地名解歧时，地名指称 $e$ 的可信度（即 Rel 值）；$s(e,e_0)$ 表示在树中 $e$ 到 $e_0$ 的路径距离；$d$ 和 $u$ 分别表示在树中上行和下行时的递减因子。

（3）确定焦点地理实体

文本地理参照树的层次性也决定了文本地理位置的层次性。焦点地理实体通常出现在参照树根结点之外的某个层次上。从焦点所在层到根结点的每个层次上都存在文本的地理位置，它们代表了不同精度的地理参照。其中，焦点地理实体的表达精度最高。例如，如果海淀区是焦点，则北京市和中国也是文本的地理位置，它们分别表示文本在地级行政区、省级行政区和国家尺度上的地理参照。

本文提出的地理求焦算法采用自上而下的策略，从根结点开始依次得到每层上最重要的地理实体直到到达焦点所在层。假设 $\gamma$ 是重要性阈值，地理求焦的一般流程如下：对参照树中每个地理实体的 *GRank* 值进行归一化处理，如果所有地理实体的 *GRank* 值都小于阈值 $\gamma$，则不存在焦点地理实体，如果有一个或者多个地理实体的 *GRank* 值大于等于阈值 $\gamma$，则首先计算 *GRank* 值大于等于阈值 $\gamma$ 的地理实体同根结点之间的最大距离 $d$，然后，从根结点开始自上而下依次确定地理参照树中每一层上的焦点（为了简化名称，本文将地理参照树每层中最重要的实体也称作焦点）。判断方法如下：

- 如果层中只存在一个地理实体，则这个实体就是此层的焦点。
- 如果层中存在多个地理实体，则选择 *GRank* 值最大的实体作为此层的焦点。

当确定了一个层次的焦点之后，需要判断是否继续提取下一个层次的焦点。如果满足（1）下一个层次的层高大于 $d$ 值；（2）下一个层次中的地理实体中包含了它们上位行政区的大部分下位行政区，并且它们的 *GRank* 值分布均匀两个条件中的任意一个，则焦点提取处理完毕。本文通过信息熵的方法衡量这个指标，称为 *Spread* 值，计算公式为

$$Spread(e) = tcount(e) \cdot \left(-\sum_{e_0 \in sub(e)} \left(\frac{GRank(e_0)}{\sum_{e_1 \in sub(e)} GRank(e_1)} \cdot \log \frac{GRank(e_0)}{\sum_{e_1 \in sub(e)} GRank(e_1)}\right)\right) / (rcount(e) \cdot \log(rcount(e))) \quad (3)$$

其中,$e$ 是一个地理实体,$rcount(e)$ 是 $e$ 在现实世界中的所有直接下位行政区的数目;$tcount(e)$ 是 $e$ 在参照树中出现的直接下位行政区的数目。如果上位行政区的 *Spread* 值大于预先设定的阈值,则表示它的下位行政区的数量足够多且分布较均匀[26]。

通过以上步骤,得到了文本的地理位置,表征为一个地理实体的层次结构,分别代表文本不同尺度的地理参照。

## 3　评　估

评估实验的目的是评价提取算法的准确度。信息检索系统通常基于公共测试集进行评价。但是,GIR 仍处于发展的初级阶段,直到 2005 年才建立了第一个支持欧洲多语言的测试集 GeoCLEF①。对于中文环境,目前还缺乏一个公认的标准测试集。因此,首先需要准备一套测试文档。文档素材来自搜狐网的新闻数据,内容涉及国内、国际、社会、文化、旅游等 18 个主题②。在使用 GeoSearcher 进行地理标注从而得到每个文档的多尺度地理参照后,从这些空间化的文档中随机抽取 195 个样本作为测试集。然后,通过人工判读的方式标注每个测试文档不同尺度下的地理位置。为了提高结果的可信度,采用了多被试标注法。对于每个被试,对测试集的所有文档进行标注,文档出现的次序是随机的。排除那些在人工标注下只存在国家尺度地理参照或者根据内容也无法判断其地理位置的文档后,最终得到 140 个有效样本。最后,对比自动提取和人工标注的结果显示在省级行政区和地级行政区尺度上文本地理参照的准确度分别达到 90.71% 和 81.97%。

## 4　结　论

基于自然语言处理技术实现文本理解是互联网发展的趋势。本文介绍了我们关于文本地理语义处理的研究工作,取得的主要成果有:(1)基于地理知识库,实现了一个三阶段自动提取文本地理位置的方法。(2)提出了基于证据理论的地名解歧算法。与现有方法相比,它的主要特点是:重语义轻语法,因此是语言无关的;解歧框架同时支持地理和非地理语义;采用了基于证据理论的证据合成方法,具有更加严格的数学基础。(3)地理求焦算法集成了相关认知理论,提取结果更加符合人的认知。

本文提出的算法在 GIR 原型系统 GeoSearcher 中得到实现,评估结果显示文本地理参照的准确度在地级行政区尺度上达到了 81.97%。进一步的研究工作为:首先,为了能够比较不同技术,建立一个中文标准测试集是我们未来的一项重要工作。其次,一个完整的地理语义提取模块是由相互影响的三个独立子模块组成,将来需要分别对这三个部分进行评估。

**致谢**:搜狗研发中心提供了测试文档素材,谨致谢意。

**参考文献**

[1] Hill L L. Georeferencing: The geographic associations of information. Cambridge, MA: MIT Press, 2006. 5-6

[2] Sanderson M, Kohler J. Analyzing geographical queries. In: Proceedings of the SIGIR Workshop on Geographic Information Retrieval, Sheffield, UK, 2004

[3] Jones C R, Purves R S. Geographical information retrieval. *International Journal geographical Information Science*, 2008, 22(3): 219-228

[4] Li H F, Srihari R K, Niu C, et al. InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In: Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, Edmonton, Canada, 2003. 39-44

[5] Purves R S, Clough P, Jones C B, et al. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 2007, 21(7): 717-745

[6] Rauch E, Bukatin M, Baker K. A confidence-based framework for disambiguating geographic terms. In: Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, Edmonton, Canada, 2003. 50-54

[7] Amitay E, HarEl N, Sivan R, et al. Web-a-Where: Geotagging Web content. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval , Sheffield, UK, 2004. 273-280

[8] Pouliquen B, Steinberger R, Ignat C, et al. Geographical information recognition and visualisation in texts written in

① http://ir.shef.ac.uk/geoclef/

② http://www.sogou.com/labs/dl/cs.html

various languages. In: Proceedings of the 2004 ACM Symposium on Applied Computing, Nicosia, cyprus, 2004. 1051-1058

[9] Silva M J, Martins B, Chaves M, et al. Adding geographic scopes to Web resources. *Computers, Environment and Urban Systems*, 2006, 30(4): 378-399

[10] Smith D A, Crane G. Disambiguating geographic names in a historical digital library. In: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, Darmstadt, Germany, 2001. 127-136

[11] Clough P. Extracting metadata for spatially-aware information retrieval on the Internet. In: Proceedings of the 2005 Workshop on Geographic Information Retrieval, Bremen, Germany, 2005. 25-30

[12] Volz R, Kleb J, Mueller W. Towards ontology based disambiguation of geographical identifiers. In: Proceedings of the 16th International World Wide Web Conference, Banff, Canada. 2007

[13] Hauptmann A G, Olligschlaeger A M. Using location information from speech recognition of television news broadcasts. In: Proceedings of the ESCA ETRW Workshop on Accessing Information in Spoken Audio, Cambridge, England, 1999. 102-106

[14] Woodruff A G, Plaunt C. GIPSY: Georeferenced information processing system. *Journal of American Society for Information Science*, 1994, 45(9): 645-655

[15] Zhang Y, Gao Y, Xue L, et al. A common sense geographic knowledge base for GIR. *Science in China Series E: Technological Sciences*, 2008, 51(Supp. I): 26-37

[16] Leidner J L. Toponym Resolution: A First Large-Scale Comparative Evaluation: [Research Report], EDI-INF-RR-0839, School of Informatics, University of Edinburgh, Edinburgh, UK, 2006. https://www.icsa.inf.ed.ac.uk/publications/online/0839.pdf

[17] Tobler W. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 1970, 46(2): 234-240

[18] Wang X, Zhang Y, Chen M, et al. An evidence-based approach for toponym disambiguation. In: Proceedings of the 18th International Conference on Geoinfomatics 2010, Beijing, China, 2010

[19] Hecht B, Raubal M. GeoSR: Geographically explore semantic relations in world knowledge. In: Proceedings of the 11th AGILE International Conference on GIScience 2008, Girona, Spain, 2008. 95-113

[20] Dempster A P. A generalization of the Bayesian inference. *Journal of Royal Statistical Society*, 1968, 30: 205 - 447

[21] Shafer G. A Mathematical Theory of Evidence. Princeton NJ: Princeton University Press, 1976

[22] Lloyd R, Patton D. Basic-level geographic categories. *The Professional Geographer*, 1996, 48(2): 181-194

[23] Tversky B, Hemenway K. Objects, parts and categories. *Journal of Experimental Psychology: General*, 1984, 113: 169-193

[24] Mark D M, Smith B, Tversky B. Ontology and geographic objects: An empirical study of cognitive categorization. In: Spatial Information Theory: A Theoretical Basis for GIS, Berlin: Springer-Verlag, 1999. 283-298

[25] Mark D M. Toward a theoretical framework for geographic entity types. In: Spatial Information Theory: A Theoretical Basis for GIS, Berlin: Springer-Verlag, 1993. 270-283

[26] Chen M, Lin X, Zhang Y, et al. Assigning geographical focus to documents. In: Proceedings of the 18th International Conference on Geoinfomatics, 2010, Beijing, China, 2010

# A semantics-based method for extracting geographic scopes of texts

Zhang Yi, Wang Xingguang, Chen Min, Liu Yu

(Institute of Remote Sensing and Geographic Information Systems, Peking University, Beijing 100871)

**Abstract**

To process geographic information in Web pages, this paper presents a novel method for extracting the geographic scopes of documents. It assigns the multi-scale geographic scope to a document through a three-stage process for dealing with geographic semantics. Firstly, the toponyms in a document are recognized under the support of the geographic knowledge base. Secondly, the ambiguous toponyms are disambiguated based on geographic and non-geographic semantics, and the evidences for disambiguation are combined by the evidence theory. Lastly, a geo-referenced tree is constructed based on a cognitive theory and the geographic focuses are obtained according to sematic relationships. The geographic location of a document is therefore determined. The above method was implemented in GeoSearcher, a prototype system for geographic information retrieval. The evaluation results show that the proposed method can reach the higher accuracy.

**Key words**: geographic information retrieval (GIR), geographic scope of texts, evidence theory