北京大学学报(自然科学版) 第 53 卷 第 2 期 2017 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 53, No. 2 (Mar. 2017)
doi: 10.13209/j.0479-8023.2016.090

基于地理关联度和证据理论的地名消歧方法研究

王星光 张瑞洁 张毅†

北京大学遥感与地理信息系统研究所, 北京 100871; † 通信作者, E-mail: zy@pku.edu.cn

摘要 针对目前地名消歧方法普遍缺乏理论基础和统一形式化方法的现状, 以地理学第一定律为理论基础, 使用地理关联度形式化地理实体之间的邻近性。在此基础上, 提出基于证据理论的地名消歧计算模型, 用于表示与合成上下文中共现的地名证据。该模型模拟人类阅读和理解文本中时空语义的认知过程, 并为地名消歧处理提供一个统一的易扩展的形式化框架。最后, 给出本文地名消歧方法的实现算法及其实验评估。结果显示, 算法综合性能指标 F1 达到 89.60%, 取得较好的实验效果。

关键词 地理信息检索(GIR); 地名消歧; 地理关联度; 证据理论

中图分类号 P208; TP311

Toponym Resolution Based on Geo-relevance and D-S Theory

WANG Xingguang, ZHANG Ruijie, ZHANG Yi†

Institute of Remote Sensing and Geographical Information Systems, Peking University, Beijing 100871;
† Corresponding author, E-mail: zy@pku.edu.cn

Abstract Aiming at the situation that previous toponym resolution researches largely lack theoretical basis and a general formal way, a concept of geo-relevance based on Tobler's Frist Law is proposed to formalize vicinity among geographic entities. Then a toponym resolution computing model based on dempster-shafer (D-S) theory is proposed to represent and combine co-occurring toponym evidences in context. The cognitive process of human reading and understanding spatiotemporal semantics in text are simulated by D-S theory, while a general and scalable formal framework for toponym resolution is provided. Finally, an experiment evaluation is given with a good result of F1 value (89.60%).

Key words geographic information retrieval; toponym resolution; geo-relevance; dempster-shafer theory

在日常交流中, 人们使用定性表达的地名陈述表示空间位置信息。不同于定量地理参照系统, 地名通常只在一定的地理范围内具有唯一性[1]。例如, 全世界共有 30 多个称为“伦敦”的城市, 但是在加拿大安大略省只有一个城市叫做“伦敦”。这种同时指称多个地理实体的地名叫做歧义地名。在蕴含丰富空间信息的文本中, 地名歧义现象比较普遍。根据 Smith 等[2]的研究, Perseus 数字图书馆项目(http://www.perseus.tufts.edu/hopper/)中历史文本语料的地名歧义高达 92%。Amitay 等[3]发现, web 页面中 37%的地名存在一词多义现象, 平均每个地名有两个不同含义。基于新闻语料库的调查发现, 新闻文本中约有 68%的地名指称不同的地理实体[4]。文本中的地名歧义现象制约了对文本时空语义的理解, 该问题的解决有助于消除文本与 GIS 之间的鸿沟, 促进文本空间信息智能处理的发展。

地名消歧(toponym resolution, TR)是一项根据语境消除地名歧义以确定地名所指的技术[5], 最初源于数字图书馆文档自动空间化的需要[2,6−7], 目前在地理信息检索(geographic information retrieval, GIR)技术的推动下有了较大的发展[8]。地名消歧的方法大致有两类: 基于规则的方法和数据驱动的方

国家自然科学基金(41271385)资助
收稿日期: 2015−10−12; 修回日期: 2016−01−06; 网络出版日期: 2016−08−18

法[9−10]。常见的数据驱动方法有基于地名共现统计的方法[11−12]和基于机器学习分类的方法[13−15]。由于缺乏足够的训练集, 数据驱动的方法在地名消歧领域中应用较少。基于规则的方法往往利用先验知识或者文本上下文的规则线索来消除地名歧义, 实际上与人们阅读文本和理解文本时空语义的策略一致[8]。目前基于规则的方法是地名消歧领域的主流方法。

文献[16]对已有地名消歧规则进行梳理, 将它们分为三类: 语用规则、语法规则和语义规则。在实际应用中, 主要依据共现地名的语义实现地名消歧。最简单的语义规则是使用缺省地理实体作为歧义地名的实际指称物。缺省地理实体是歧义地名所有指称中最重要的地理实体。衡量地理实体重要性的因素有人口[3,17−18]、类型[2,18−22]、面积[23]、出现词频[21,24−26]和网页链入度[27]等。另一类语义规则利用共现地名之间的语义关系消歧, 如包含关系规则[13,17−18,28−31]、空间相近规则[2,18−19,23,32−34]等, 这些规则假设在文本中相近出现的地名具有地理空间上的邻近性。缺省规则简单、易实现, 因此应用比较广泛, 但是准确度较低[3,18,35]。邻近规则的消歧结果普遍好于缺省规则[2,24−25,36], 但是也存在若干问题: 1) 缺乏科学的符合认知的邻近性形式化方法, 例如对不同拓扑关系进行专家打分或者采用定量地理距离[2,19,32]; 2) 对于多个证据缺乏科学合理的证据合成方式, 例如采用简单的算术累加[3,18]。

针对以上问题, 本文提出地理语义关联度的概念, 用于形式化地理实体之间的邻近性,并以此为基础, 发展一个基于证据理论的地名消歧计算模型。该模型模拟人类阅读理解文本空间语义的认知过程, 并且易于扩展, 从而使得消歧结果更准确, 也更符合人类的认知结果。

1　地名消歧与地理关联度

1.1　一个实例

“鼓楼区, 南京市中心城区之一, 江苏省党政军首脑机关所在地。它西越秦淮河, 直抵长江之滨”。全国总共有 4 个鼓楼区, 分别属于南京、福州、徐州和开封。因此, “鼓楼区”是一个歧义地名。

1.2　消歧原理

用于消除地名歧义的线索通常来自歧义地名所处的语境。例如, 在南京市说到鼓楼区, 通常指南京市下辖的鼓楼区。除空间、时间和情景等语境因素外, 在词义消歧领域最常用的是上下文, 即文本中歧义词所处位置前后的一组特征词, 一般是上下文中共现的地名序列[37]。

在基于规则的方法中, 文本中共现地名之间的语义关联是重要的消歧线索。根据文献[38], 导致地名共现的语义关联主要有 4 类: 1) 包含关系, 例如, “鼓楼区是南京市中心城区之一”; 2) 邻近关系, 例如, “环渤海经济区以北京和天津为中心”; 3) 空间交互, 例如, “新开通的京广高铁北起北京, 南到广州”; 4) 类属关系, 例如, “北京和伦敦都是首都, 是本国的政治经济文化中心”。根据是否受距离影响, 又可以将它们分为两大类: 地理语义关联和非地理语义关联。由于非地理语义关联的含义过于宽泛和模糊, 因此 GIR 领域中主要使用地理关联作为消歧规则。本文的消歧证据也限定在地理关联范围之内。

地理包含和相近规则是采用最多的地理语义规则。研究者普遍认为, 地名出现在同一文本中的主要原因是它们所指的地理实体之间存在着空间包含关系或者地理邻近关系。它的理论基础来自 Tobler[39]的地理学第一定律, 即“所有事物都是相关的, 并且事物在空间上越相近其相关性就越大”。

1.3　地理关联度的概念

本文提出地理关联度的概念, 用于形式化地理学第一定律。我们认为任意两个地理实体之间存在着地理关联, 关联的程度由两个实体之间的地理距离决定。

用于衡量两个实体地理关联程度的指标是地理关联度, 它可以用一个介于 0 与 1 之间的二元函数表示:

$$G\mathrm{Rel}(x, y) \in (0, 1]\ (x, y \in G),$$

G 是论域中所有地理实体的集合。如果 $G\mathrm{Rel}(x, y)=0$, 则表示 x 与 y 没有关联, 这就违背了地理学第一定律, 因此 $G\mathrm{Rel}(x, y)>0$。$G\mathrm{Rel}(x, y)$越趋近 1, 关联程度越高。如果 $G\mathrm{Rel}(x, y)=1$, 则表示 x 与 y 相等或者包含。地理语义关联度的大小由地理实体之间的距离决定, 距离越近, 关联度越大; 反之, 关联度越小。根据地名消歧原理, 如果一个歧义地名的某个可能所指与上下文中其他地名所指实体的地理关联程度越高, 则该可能所指是歧义地名实际所指的可能性就越大。

1.4 地理关联度的计算

地名库中不仅记录了地理实体的名称、类型、地理覆盖等属性信息, 还记录了地理实体之间的定性空间关系[40]。另外, 地名库的数据规模通常达到百万级别以上。因此, 基于规则的地名消歧方法主要基于地名库实现。其中, 地理实体的名称用于地名识别; 地理实体之间的空间关系用于消歧规则的实现。但是, 地名库中通常只存储空间包含关系, 而实体之间的相邻关系和空间距离则需要通过实体的地理覆盖属性动态计算得到。地理覆盖是描述地理实体位置和形状的属性, 在地名库中抽象为简单点或者外包矩形。因此, 基于此类坐标数据计算实体间的定性空间关系(如相邻关系)和定量距离存在较大误差[19]。鉴于以上情况, 我们提出一个基于认知的定性地理信息系统 CSGKB, 模拟和存储人脑中的地理知识[41]。

与定量 GIS 的绝对空间观不同, CSGKB 基于相对空间观对外部地理世界进行建模, 记录地理实体以及地理实体之间的定性空间关系。在 GIS 领域中, 主要有拓扑、方向和度量 3 类空间关系[42]。CSGKB 主要通过包含、相邻和相交这 3 类拓扑空间关系在地理实体之间建立空间联系; 方位关系可以作为拓扑关系的属性存储在 CSGKB 中, 其中包含关系记录内方位关系, 相邻关系记录外方位关系; 度量关系则是基于地理实体之间的拓扑关系动态计算得到的拓扑距离。

CSGKB 通过定性拓扑关系建立地理实体的结构。其中, 地理参照系统是组织地理知识的基本架构。它有两个基本要求: 1) 应用区域全覆盖; 2) 用于地理定位[43]。常见的地理参照系统有地名系统、邮政系统、地理坐标系、投影坐标系和格网系统等。自然语言文本主要采用定性的地名参照系统。在 CSGKB 中, 地理参照系统是由行政区实体通过相邻和包含关系形成的一个多层次定性地理参照系统, 非行政区实体则通过与行政区的包含或相交关系实现其地理定位。采用行政区作为地理参照系统的主要原因有: 1) 行政区是国家进行分级管理而实行的区域划分, 覆盖全球主要陆地, 并具有互不相交且联合完备的层次结构; 2) 认知心理学认为, 人脑中的类别知识主要按照基本层次(basic level)进行组织, 而行政区作为基本层次, 用于存储与组织地理知识[44]; 3) 现实中的地图集主要按照行政区来组织和展现地理知识。

本文提出通过基于拓扑关系的定性距离计算地理关联度, 是基于 CSGKB 中的地理参照结构来实现。测量定性距离的空间关系有包含关系、相邻关系、相交关系以及基于拓扑度量的相离关系。

假设任意两个地理实体 a 和 b, $d_S(a, b)$表示 a 与 b 之间的拓扑距离(其中 S 表示地理尺度, CSGKB 支持省级、市级和县级 3 类地理尺度)。

1) 如果 a 包含 b, 或 a 与 b 相交, 那么 $d_S(a, b)=0$。例如, 永定河与河北省之间的拓扑距离是 0。

2) 如果 a 与 b 相邻, 那么 $d_S(a, b)=1$。例如, 保定市与石家庄市之间的拓扑距离是 1。

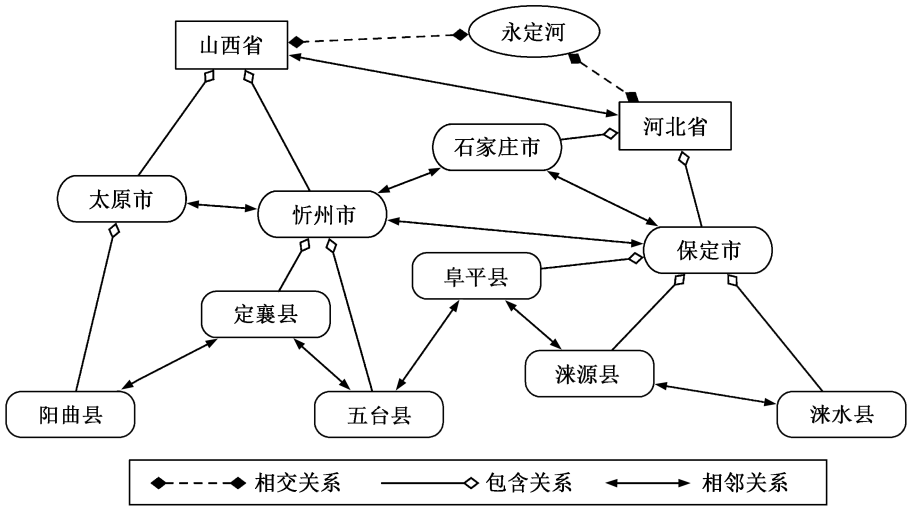

图 1 CSGKB 中的地理实体及其空间关系

Fig. 1 Geographic entities and their spatial relations in CSGKB

3) 如果 a 与 b 相离, 那么 $d_S(a, b)=n$ (n 表示 a 与 b 在地理参照系统中构成最短路径的相邻关系数目)。例如, 如图 1 所示, 山西省阳曲县与河北省涞源县之间的拓扑距离, 在省级尺度下是 1, 在市级尺度下是 2, 在县级尺度下则是 4。

4) 如果 a 或 b 是非行政区实体, 那么需要先通过地理参照结构, 将非行政区映射到行政区, 再计算 a 与 b 之间的拓扑距离。

基于以上定性空间距离的概念, 定义地理实体 a 与 b 之间的地理语义关联度计算公式如下:

$$G\mathrm{Rel}(a, b) = e^{-d_S(a, b)} \text{。} \tag{1}$$

2　基于 D-S 理论的多证据融合

2.1　D-S 理论基础

本文使用 D-S 证据理论作为地名消歧的形式化框架。该理论最早由 Dempster[45]提出, Shafer[46]做了进一步推广和发展。D-S 理论与概率的区别是, 它使用一个概率范围而不是单个概率值表示不确定性, 并且提供证据合成的方法。在地名消歧的语境下, 本节介绍 D-S 理论的基本概念和合成规则。

2.1.1　识别框架

假设存在一个变量 x, 其相互独立的所有可能值构成一个有限集合 Ω, 称为关于 x 的识别框架。Ω 中所有子集的集合为幂集 2^{Ω}。对于 $\forall A \in 2^{\Omega}$, A 都对应一个关于 x 的命题, 则称该命题为“x 的值在 A 中”。

在地名消歧问题中, x 代表歧义地名, Ω 是 x 所有可能所指的集合, A 表示 x 指称特定地理实体的论断。例如本文实例中, x = 鼓楼区, 则 Ω = {南京鼓楼区, 徐州鼓楼区, 开封鼓楼区, 福州鼓楼区}; 如果 A = {徐州鼓楼区}, 则表示 x 指称徐州鼓楼区; 如果 A = {徐州鼓楼区, 开封鼓楼区}, 则表示 x 指称徐州鼓楼区或开封鼓楼区。

2.1.2　基本概率分配函数

如果集函数 $m: 2^{\Omega} \to [0, 1]$ 满足: 1) $\mathrm{m}(\phi)=0$, 2) $\sum_{A \in 2^{\Omega}} m(A) = 1$, 则称 m 为 Ω 上的基本概率分配函数或 mass 函数。$m(A)$表示证据支持命题 A 的信任程度, 称为 A 的基本可信度。每个证据源确定一个 mass 函数, 不同证据源得到的 mass 函数不同。

在地名消歧问题中, 证据源是歧义地名所在上下文中出现的其他地理实体。例如, 实例中歧义地名“鼓楼区”具有 4 个消歧证据源: 江苏省、南京市、秦淮河和长江, 每个证据源分别对应 Ω 上的一个 mass 函数, 实例中的 4 个证据源就分别确定 4 个 mass 函数。

当 $A \sqsubseteq \Omega$ 且 $|A| > 1$ 时, $m(A)$表示对歧义地名指称多个地理实体的信任程度, 无法唯一地确定歧义地名的指称实体, 所以对地名消歧是无意义的, 因此本文规定: 当 $A \sqsubseteq \Omega$ 且 $|A| = 1$ 时, $m(A)>0$; 否则 $m(A)=0$。

为了满足 mass 函数的两个条件, 我们规定: $m(\Omega) = 1 - \sum_{A \in 2^{\Omega} \,\&\&\, A \neq \Omega} m(A)$, 表示没有分配的证据, 称为未分配信任度。在地名消歧时, $m(\Omega)$表示证据源消除地名歧义的不确定性。

2.1.3　信任函数和似然函数

根据基本可信度 $m(A)$, 可以定义另外两个证据度量函数。如果函数 $\mathrm{Bel}: 2^{\Omega} \to [0, 1]$ 满足

$$\mathrm{Bel}(A) = \sum_{B \sqsubseteq A} m(B), \forall A \sqsubseteq \Omega \text{,} \tag{2}$$

则称 Bel 为信任函数, 它表示对命题的总信任程度。如果函数 $\mathrm{Bel}: 2^{\Omega} \to [0, 1]$ 满足

$$\mathrm{Pl(A)} = \sum_{A \cap B \neq \phi} m(B), \forall A, B \sqsubseteq \Omega \text{,} \tag{3}$$

则称 Pl 为似然函数, 表示不否定命题 A 的程度。因此, $\mathrm{Pl}(A)$是比 $\mathrm{Bel}(A)$更加宽松的一种信任估计。

在地名消歧问题中, 除未分配信任度外, 其余信任全部分配给单元素的核, 因此 $\mathrm{Bel}(A)= m(A)$ $(\forall A \subset \Omega)$, 而 $\mathrm{Pl}(A) = m(A) + m(\Omega)(\forall A \subset \Omega)$。

2.1.4　证据合成规则

Dempster 合成规则是最早提出, 也是使用最广泛的证据合成公式。设 $m_1, m_2, \ldots, m_n$ 是识别框架 Ω 上的 n 个 mass 函数, 则合成后的 mass 函数 $m = m_1 \oplus m_2 \ldots \oplus m_n$ 满足

$$m(A) = \begin{cases} 0, & A = \phi, \\ (1-K)^{-1} \times \sum_{\cap A_j = A} \prod_{i=1}^{n} m_i(A_j), & A \neq \phi, \end{cases} \tag{4}$$

其中, $K = \sum_{\cap A_j = \phi} \prod_{i=1}^{n} m_i(A_j)$。$K$ 为冲突因子, 反映证据冲突的程度; K 值越大, 说明证据冲突程度越大。

2.2　地名消歧的证据表示

假设实体 a 是歧义地名 x 的任意一个可能所指, 如果实体 e 是 x 的一个消歧证据源, 那么由该证据源得到的 mass 函数反映其对歧义地名 x 指称

实体 a 的可信程度。根据地名消歧原理, e 支持 x 指称 a 的可信程度是由实体 e 与 a 之间的地理关联度决定的。因此, 本文将 $G\mathrm{Rel}(e, a)$作为信度赋给 $m(A=\{a\})$。由于 $G\mathrm{Rel}$ 是一个介于 0 与 1 之间的度量函数, 所以每个证据源确定的 $m(A)$的总信度等于 N (其中 N 是关于 x 的识别框架中元素的数目)。因此, 为了满足基本概率分配函数的两个条件, 我们定义证据源 e 决定的 mass 函数如下:

$$m(A)=\begin{cases}0, & A=0,\\ \dfrac{G\mathrm{Rel}(e,a)}{N}, & A\neq 0, A\neq \Omega,\\ 1-\sum_{A=2^{\Omega}\text{且}A\neq\Omega} m(A), & A=\Omega\text{。}\end{cases} \quad (5)$$

2.3 地名消歧的证据合成和消歧步骤

对于一个歧义地名, 通常在其上下文中存在多个消歧证据。例如, 实例中“鼓楼区”就有 4 个消歧证据源, 每个证据源得到不同的 mass 函数。通过证据合成规则, 可以计算出一个 mass 值, 它可以作为该组证据联合作用下的 mass 函数。其中, mass 值最高的候选地理实体就是歧义地名的最可能所指。综合所述, 本文地名消歧计算步骤定义如下。

1) 对于文本中出现的歧义地名, 首先获得其上下文中出现的其他地理实体作为消歧证据。例如, 实例中“鼓楼区”有 4 个消歧证据源: 南京市、江苏省、秦淮河和长江。

2) 计算歧义地名的每个可能所指与每个证据源之间的定性地理距离。依据 CSGKB 中显式记录的地理实体之间的定性拓扑距离, 得到表 1 所示的结果。

3) 根据式(1), 计算歧义地名的每个可能所指与每个证据源之间的地理关联度。例如, “南京市鼓楼区”与“南京市”的地理关联度 $G\mathrm{Rel}(\{\text{南京鼓楼区}\}, \{\text{南京市}\}) = e^{-0} = 1$。结果如表 2 所示。

4) 根据式(5), 计算每个证据源的 mass 函数。例如, 证据“南京市”作用下“南京市鼓楼区”的 mass 值 $m(\{\text{南京鼓楼区}\}) = 1/4 = 0.25$, 未分配信度 $m(\{\text{南京鼓楼区, 徐州鼓楼区, 开封鼓楼区, 福州鼓楼区}\}) = 1 - 0.25 - 0.0006 - 1.536\times10^{-6} - 1.536\times10^{-6} = 0.7479$。每个候选指称地理实体在不同证据下的 mass 数据见表 3。

5) 根据式(4), 计算歧义地名每个可能所指的证据合成 mass 值(表 3), 并且选取 mass 值最大的可能所指, 赋予歧义地名的实际指称地理对象。本文实例中, “南京市鼓楼区”的证据合成 mass 值最高

表 1 歧义地名所有可能所指与证据源之间的定性距离 (县级尺度)

Table 1 Qualitative distance among candidates belonging to ambiguous toponym and evidences (county level)

地理实体	定性地理距离			
	南京市	江苏省	秦淮河	长江
南京鼓楼区	0	0	0	0
徐州鼓楼区	6	0	6	6
开封鼓楼区	12	6	12	11
福州鼓楼区	12	12	13	9

表 2 歧义地名所有可能所指与证据源之间的地理关联度

Table 2 Geo-relevance among candidates belonging to ambiguous toponym and evidences

地理实体	地理关联度			
	南京市	江苏省	秦淮河	长江
南京鼓楼区	1	1	1	1
徐州鼓楼区	0.0025	1	0.0025	0.0025
开封鼓楼区	6.144×10^{-6}	0.0025	6.144×10^{-6}	1.670×10^{-5}
福州鼓楼区	6.144×10^{-6}	6.144×10^{-6}	2.260×10^{-6}	0.0001

表 3 证据源和证据合成的 mass 值

Table 3 Mass value combined by evidences

地理实体	Mass 值				
	南京市	江苏省	秦淮河	长江	证据合成
南京鼓楼区	0.2500	0.2500	0.2500	0.2500	0.6296
徐州鼓楼区	0.0006	0.2500	0.0006	0.0006	0.1241
开封鼓楼区	1.536×10^{-6}	0.0006	1.536×10^{-6}	4.175×10^{-6}	0.0003
福州鼓楼区	1.536×10^{-6}	1.536×10^{-6}	5.651×10^{-7}	3.085×10^{-5}	1.158×10^{-5}
未分配信度	0.7494	0.4994	0.7494	0.7493	0.2460

(0.6296), 因此是歧义地名“鼓楼区”的实际所指。另外, 未分配信度表示消歧结果的不确定性, 实例结果是 0.2460。一般来说, 合理证据越多, 不确定性越低。

3　实验与评估

地名消歧研究往往借用信息检索领域的评价指标进行消歧算法性能评估: 准确率 $P = T_C/(T_C+T_I)$, 召回率 $R = T_C/T_N$, $F1 = 2PR/(P+R)$。其中, T_C 是正确消歧得到的地名数目, T_I 是错误消歧得到的地名数目, T_N 是文档集中歧义地名数目。

为评估本文提出的地名消歧算法的实用性能, 从搜狗 2012 年全网新闻数据(http://www.sogou.com/labs/dl/ca.html)随机采集 1063 篇文本作为测试集。其中, 地名出现 13105 次, 歧义地名 2970 次, 平均每篇文本含地名 12.33 次, 地名的歧义比例达到 22.66%。

我们设计两组实验, 以考察本文方法在不同因素作用下的消歧效果。

实验 1　考察歧义地名上下文中证据数目及消歧窗口大小的影响。

图 2 显示上下文中证据数目对消歧效果的影响, 划分为 3 个部分: 1) 证据数目 1~5, 消歧效果显著上升, F1 值最高达到 84.91%; 2) 证据数目 6~41, 消歧效果有水平波动; 3) 证据数目大于 41, 消歧效果缓慢下降。这说明, 文本中歧义地名附近的证据对消歧效果有关键影响, 上下文中远距离的证据与歧义地名的相关性不大, 随着证据数目的过度增多反而会对消歧效果产生负面影响。在指标性能上, 不确定度随着证据数目的增加而递减, 说明充分的共现证据有助于减少消歧理解过程的认知分歧。召回率普遍比准确率低, 是限制整体性能(即 F1 值)的瓶颈因素。

图 3 显示限定证据窗口内(证据数目不大于某一特定数值)的地名消歧结果。准确率在窗口大小为 5 时达到最大值 91.49%, 召回率延后到在窗口大小为 19 时达到最大值 74.89%, 导致 F1 同样延后到窗口大小为 19 时达到最大值 81.92%。准确率的结果普遍较高, 说明基于定性距离的地理关联度与证据合成的策略处理地名歧义问题是可行的。然而, 召回率偏低, 存在两方面原因: 首先, 无证据时, 无法开展证据推理; 其次, 上下文证据少时, 证据合成效果不佳。

实验 2　比较不同消歧实现算法的性能差异。

由于缺乏统一的面向地名消歧任务的中文语料, 难以横向比较本文算法与其他算法的优劣。为了做纵向比较, 设计两个比较算法, 描述如下。

1) 词频缺省值法(简称 Baseline 算法): 将歧义地名所指中词频最高的地理实体赋予歧义地名进行消歧。

2) 组合算法(简称 RE_B 算法): 当歧义地名上下文无证据, 或实验 1 中算法无法实现召回时, 使用词频缺省值对其进行地名消歧, 否则执行实验 1 的算法进行消歧。

图 4 显示组合算法在不同证据窗口大小内的地名消歧效果。F1 在窗口大小为 5 时达到最高值 89.60%。这是因为组合算法消除了实验 1 中召回率偏低的负面影响。

现将 Baseline 算法、实验 1 算法中窗口大小分别为 5 和 19 的具体实现(命名为 RE-5 和 RE-19)、组合算法在窗口大小为 5 时的实现(命名为 RE_B-5)进行性能比较, 结果列于表 4。对比发现, RE-5, RE-19 和 RE_B-5 比 Baseline 的 F1 值均有显著提升(高出 10%以上), 说明基于定性距离的地理关联度和共现地名证据推理方法来动态地推断歧义地名所指的策略是可行的。并且, 基于静态词频的缺省值法可移植性较差, 本文提出的算法具有一定程度上的语料独立性。与 RE-5/19 相比, RE_B-5 准确率略有降低, 但是召回率大幅度提升(提高 15%左右), F1 值为 89.60%, 达到实用性能。RE_B-5 的结果显示, 离歧义地名最近的 5 个共现地名(左 3 右 2)对消歧最有效, 在消歧失败或上下文没有共现证据时再采用缺省值法处理歧义问题, 这与人们阅读文本时理解歧义词汇语义的认知模式是一致的。

另外, 通过分析实验数据, 发现产生消歧错误的主要原因是, 一些歧义地名的多个候选所指具有包含关系, 算法无法区分这些候选所指与共现证据地名之间的拓扑距离, 从而导致消歧失败。例如, “聂拉木”在中国行政单位有两种含义: 西藏日喀则地区的聂拉木县和西藏日喀则地区聂拉木县的聂拉木镇。若上下文中出现地名证据“日喀则地区的南木林县”, 该证据与聂拉木县和聂拉木镇的县级定性拓扑距离相同(均为 3), 无法区分地理关联度的差异, 导致消歧发生错误。若在关联度计算中引入非地理语义关联, 可能是一种降低消歧错误的解决思路。

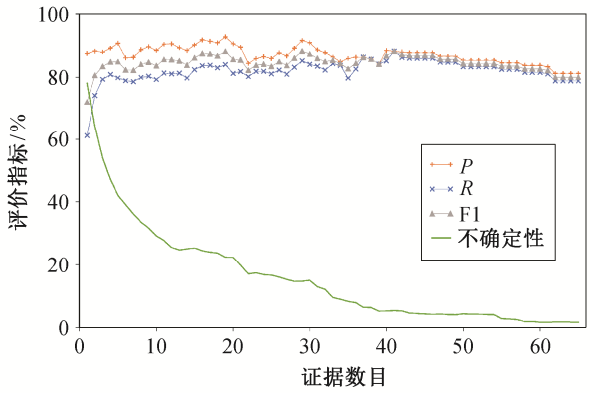

图 2 不同证据数目下的地名消歧效果

Fig. 2 Resolution effect by different number of evidences

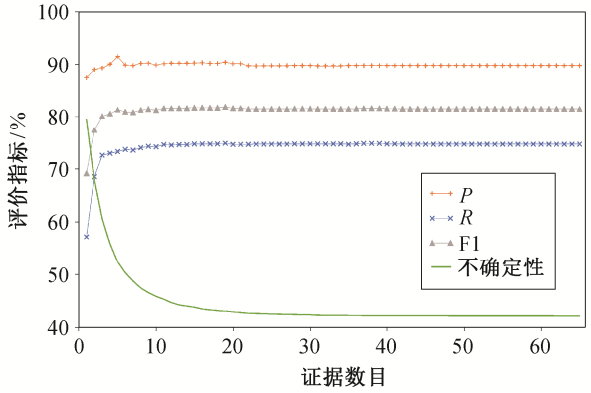

图 3 限定窗口大小内的地名消歧效果

Fig. 3 Resolution effect within limited window of evidences

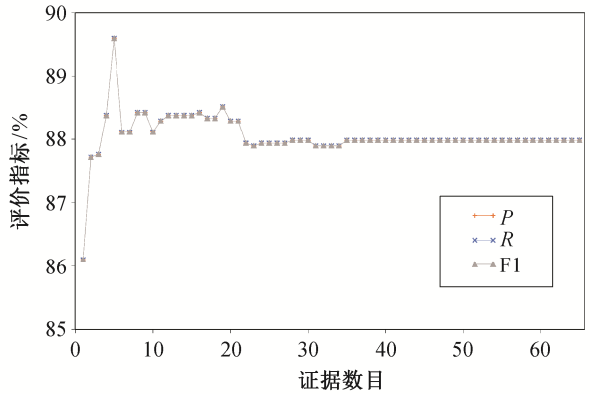

图 4 组合算法的地名消歧效果

Fig. 4 Resolution effect by composition method

4 结论和展望

地名消歧处理是时空文本语义处理的一项核心技术，具有广阔的应用前景，例如 GIR、时空行为提取和时空数据挖掘等。本文通过分析现有地名消歧规则，认为文本中共现地名之间的地理语义关联是地名消歧的重要线索，因而提出基于地理关联度的消歧方法和基于 DS 理论的消歧计算形式化框架。该方法有以下主要特点：1) 以地理学第一定律为理论基础，基于地理语义实现地名消歧，符合人类阅读理解文本的认知过程；2) 通过地理关联度形式化地理学第一定律，基于 D-S 理论，实现消歧证据表示和证据合成，为地名消歧提供一个统一的易扩展的形式化模型。

表 4 不同地名消歧算法的性能比较

Table 4 Performance comparison of different toponym resolution methods

算法	*P*/%	*R*/%	F1/%
Baseline	70.90	70.90	70.90
RE-5	91.49	73.32	81.40
RE-19	90.42	74.89	81.92
RE_B-5	89.60	89.60	89.60

测试结果显示：1) 提供消歧线索的上下文窗口大小选择为 5 较合适；2) 与缺省规则相比，基于地理关联语义的消歧方法将 F1 值从 70%左右提高到 81%；3) 综合使用这两种语义规则，可以使召回率显著提高，F1 值达到 89.60%。

下一步的工作为：首先，影响地理关联度的因素除地理距离外，还需要考虑认知距离，使得关联度的计算更加符合人类的认知过程；其次，考察非地理语义关联对地名消歧的影响，并将其纳入消歧框架中，使得消歧结果更加准确；最后，测试算法在非中文语言文本中的性能，评估其扩展能力。

参考文献

[1] Longley P A, Goodchild M F, Rhind D W. Geographic information systems and science. London: John Wiley & Sons, 2005

[2] Smith D A, Crane G. Disambiguating geographic names in a historical digital library // Proceedings of ECDL. Darmstadt, 2001: 127–136

[3] Amitay E, Har'El N, Sivan R, et al. Web-a-where: geotagging web content // Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2004: 273–280

[4] Garbin E, Mani I. Disambiguating toponyms in news

// Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, BC, 2005: 363–370

[5] Leidner J L. Toponym resolution in text: annotation, evaluation and applications of spatial grounding of place names [D]. Edinburgh: University of Edinburgh, 2007

[6] Olligschlaeger A M, Hauptmann A G. Multimodal information systems and GIS: the informedia digital video library // Proceedings of the 1999 ESRI User Conference. San Diego, 1999: 102–106

[7] Densham I, Reid J. A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service // Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References. Edmonton, 2003: 79–80

[8] Jones C B, Purves R S. Geographical information retrieval. International Journal of Geographical Information Science, 2008, 22(3): 219–228

[9] 张毅, 王星光, 陈敏, 等. 基于语义的文本地理范围提取方法. 高技术通讯, 2012, 22(2): 165–170

[10] Buscaldi D, Rosso P. A conceptual density-based approach for the disambiguation of toponyms. International Journal of Geographical Information Science, 2008, 22(3): 301–313

[11] Overell S, Rüger S. Using co-occurrence models for placename disambiguation. International Journal of Geographical Information Science, 2008, 22(3): 265–287

[12] Batista D S, Silva M J, Couto F M, et al. Geographic signatures for semantic retrieval // Proceedings of the 6th Workshop on Geographic Information Retrieval. New York: ACM, 2010: 93–100

[13] Hu Youheng, Ge Linlin. A supervised machine learning approach to toponym disambiguation // The Geospatial Web. London: Springer, 2007: 117–128

[14] Lieberman M D, Samet H. Adaptive context features for toponym resolution in streaming news // Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2012: 731–740

[15] Santos J, Anastácio I, Martins B. Using machine learning methods for disambiguating place references in textual documents. Geojournal, 2015, 80(3): 375–392

[16] Wang Xingguang, Zhang Yi, Chen Min, et al. An evidence-based approach for toponym disambiguation // Proceedings of the Eighteenth International Conference on Geoinformatics. Beijing: IEEE, 2010: 1–7

[17] Zhang Wei, Gelernter J. Geocoding location expressions in Twitter messages: a preference learning method. Journal of Spatial Information Science, 2014 (9): 37–70

[18] Li Yi, Moffat A, Stokes N, et al. Exploring probabilistic toponym resolution for geographical information retrieval // Proceedings of 3rd Workshop on Geographic Information Retrieval. Seattle, 2006: 17–22

[19] Tang Xuri, Chen Xiaohe, Peng Minxuan. Toponym resolution in discourse // Natural Language Processing and Knowledge Engineering. Beijing, 2008: 1–8

[20] Volz R, Kleb J, Mueller W. Towards ontology-based disambiguation of geographical identifiers // WWW 2007 Workshop I3. Banff, 2007: 1–7

[21] 杜萍, 刘勇. 中文地名识别与歧义消除: 以中国县级以上行政区划地名为例. 遥感技术与应用, 2011, 26(6): 868–873

[22] 唐旭日, 陈小荷, 张雪英. 中文文本的地名解析方法研究. 武汉大学学报: 信息科学版, 2010, 35(8): 930–935

[23] Pouliquen B, Steinberger R, Ignat C, et al. Geographical information recognition and visualization in texts written in various languages // Proceedings of the 2004 ACM Symposium on Applied Computing. New York: ACM, 2004: 1051–1058

[24] Li Huifeng, Srihari R K, Niu Cheng, et al. Location normalization for information extraction // Proceedings of the 19th International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2002: 1–7

[25] Li Huifeng, Srihari R K, Niu Cheng, et al. InfoXtract location normalization: a hybrid approach to geographic references in information extraction // Proceedings of the HLT-NAACL 2003 workshop on Analysis of Geographic References-Volume 1. Stroudsburg: Association for Computational Linguistics, 2003: 39–44

[26] Overell S. Geographic information retrieval: classification, disambiguation and modelling [D]. London: Imperial College London, 2009

[27] Hu Yingjie, Janowicz K, Prasad S. Improving Wikipedia-based place name disambiguation in short

texts using structured data from DBpedia [C/OL] // Proceedings of the 8th Workshop on Geographic Information Retrieval. Dallas: ACM. (2014−11−04) [2015−09−20]. http://geog.ucsb.edu/~jano/GIR2014.pdf

[28] 朱少楠, 张雪英, 李明, 等. 基于行政隶属关系树状图的地名消歧方法. 地理与地理信息科学, 2013, 29(3): 39–42

[29] Bensalem I, Kholladi M K. Toponym disambiguation by arborescent relationships. Journal of Computer Science, 2010, 6(6): 653–659

[30] Overell S, Magalhaes J, Rüger S. Place disambiguation with co-occurrence models // CLEF 2006 Workshop, Working Notes. Alicante, 2006: 59−68

[31] Overell S, Rüger S M. Identifying and grounding descriptions of places // Proceedings of the Third ACM Workshop on Geographical Information Retrieval at SIGIR (GIR'06). Seattle, 2006: 14−16

[32] Leidner J L, Sinclair G, Webber B. Grounding spatial named entities for information extraction and question answering // Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References-Volume 1. Stroudsburg: Association for Computational Linguistics, 2003: 31–38

[33] Buscaldi D, Rosso P. Map-based vs. knowledge-based toponym disambiguation // Proceedings of the 2nd International Workshop on Geographic Information Retrieval (GIR'08). New York: ACM, 2008: 19–22

[34] Lieberman M D, Samet H, Sankaranayananan J. Geotagging: Using proximity, sibling, and prominence clues to understand comma groups // Proceedings of the 6th Workshop on Geographic Information Retrieval (GIR'10). Zurich: ACM, 2010: 1−8

[35] Pouliquen B, Kimler M, Steinberger R, et al. Geocoding multilingual texts: Recognition, disambiguation and visualization // Proceedings of LREC-2006. Genoa, 2006: 53−58

[36] Buscaldi D, Magnini B. Grounding toponyms in an Italian local news corpus // Proceedings of the 6th Workshop on Geographic Information Retrieval. Zurich, 2010: 1−5

[37] Buscaldi D. Approaches to disambiguating toponyms. Sigspatial Special, 2011, 3(2): 16–19

[38] Liu Yu, Wang Fahui, Kang Chaogui, et al. Analyzing relatedness by toponym co-occurrences on web pages. Transactions in GIS, 2014, 18(1): 89–107

[39] Tobler W R. A computer movie simulating urban growth in the Detroit region. Economic geography, 1970, 46: 234–240

[40] Hill L. Core elements of digital gazetteers: placenames, categories, and footprints // Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries. Berlin: Springer, 2000: 280–290

[41] Zhang Yi, Gao Yong, Xue Lulu, et al. A common sense geographic knowledge base for GIR. Science in China Series E: Technological Sciences, 2008, 51(1): 26–37

[42] 邬伦, 刘瑜, 张晶, 等. 地理信息系统: 原理方法和应用. 北京: 科学出版社, 2005

[43] Longley P A, Goodchild M F, Maguire D J, et al. Geographical information systems and science. Chichester: John Wiley & Sons, 2005

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

[45] Dempster A P. Upper and lower probabilities induced by a multivalued mapping. The Annals of Mathematical Statistics, 1967, 38(2): 325–339

[46] Shafer G. A mathematical theory of evidence. Princeton: Princeton University Press, 1976