

# 基于语义文法的地理实体位置关系的获取

周琦<sup>1</sup> 陆叶<sup>1</sup> 李婷玉<sup>1</sup> 王亚<sup>2</sup> 张再跃<sup>1</sup> 曹存根<sup>2</sup>

(江苏科技大学计算机科学与工程学院 镇江 212000)<sup>1</sup>

(中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)<sup>2</sup>

**摘要** 地理信息与数据是客观知识世界的重要组成部分。研究如何从大量非结构化的信息中自动抽取地理实体位置关系具有重要意义。提出一种基于语义文法的地理实体位置关系获取方法,该方法可准确地从网页文本中获取多个地理实体之间的复合位置关系。首先,设计一种反映地理实体位置关系的语义文法 GeoRSG。GeoRSG 反映了地理实体位置关系的层次分类关系,并采用基于规则的方式刻画地理实体位置关系在文本中的语言表达方式。然后,实现地理实体位置关系解析器 GeoRSG Parser。该解析器利用 GeoRSG 对文本进行解析,获得谓词表达形式的位置关系知识。实验结果显示,该方法从 1000 条语句中获取了 81 条三元和 816 条二元地理实体位置关系,并且取得了 88.85% 的正确率。

**关键词** 地理实体位置关系,语义文法,知识抽取

**中图分类号** TP391.1      **文献标识码** A      **DOI** 10.11896/j.issn.1002-137X.2016.7.038

## Acquiring Relationships Between Geographical Entities Based on Semantic Grammar

ZHOU Qi<sup>1</sup> LU Ye<sup>1</sup> LI Ting-yu<sup>1</sup> WANG Ya<sup>2</sup> ZHANG Zai-yue<sup>1</sup> CAO Cun-gen<sup>2</sup>

(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212000, China)<sup>1</sup>

(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)<sup>2</sup>

**Abstract** Geographic information and data are important components of the objective knowledge world. Geographic information extraction (GIE) aims to extract various relationships between geographic entities from unstructured geographic text. A novel method for GIE was proposed, which depends on semantic parsing with a geographic grammar. First, GeoRSG (Geographical Relationship Semantic Grammar) was constructed, which reflects geographic relationships in Chinese written language. GeoRSG also reflects a classification of relationships between geographic entities, and uses a rule-based method to depict linguistic expressions of relationships in the text. Then, we implemented a parser, called the GeoRSG Parser, which is used to obtain the geographical knowledge in the form of the predicate with the help of GeoRSG. Experiments indicate that the method can obtain 81 triples relationships and 816 binary relationships between geographic entities from 1000 statements, and has achieved a precision rate of 88.85%.

**Keywords** Relationship between geographic entities, Semantic grammar, Knowledge acquisition from text

## 1 引言

在互联网时代,大量的信息都以非结构化的形式存在。如何从海量的文本信息中智能化地获取知识成为新的挑战。由于自然语言的灵活性、复杂性、歧义性、模糊性等特点,计算机自动文本知识获取被认为是一项非常困难和费时的任务,一直是阻碍智能系统研究和开发的瓶颈问题<sup>[1]</sup>。

地理领域知识是知识库的重要组成部分,是很多智能应用系统的基础,如专家系统、军事经济决策系统、咨询系统、智能教学系统等<sup>[2]</sup>,另外在自然语言理解、地理信息系统等方面也有重要的应用。如果要得到一个正确完备的地理领域知识

知识库,则需要先获取候选地理实体位置关系。在已获取的地理实体位置关系的基础上,通过空间推理规则,挖掘出更复杂、隐含、完全的关系知识,这对于推动整个地理信息科学的发展有重大的意义。地理实体位置关系在空间环境事实中占有较高的比例,规模庞大,并且普遍存在于互联网文本中,如果仅仅依靠人工的方法在海量的信息库中分析、过滤和获取知识,效率会很低,而且正确率也得不到保证,所以很难实现地理空间知识的语义表达、共享与互操作,因此必须研究自动获取地理领域知识的有效方法。

地理实体位置关系反映了某个地理实体内部或者是多个实体之间存在的空间关系,其描述的目的是建立不同地理实

到稿日期:2015-06-11 返修日期:2015-08-11 本文受国家自然科学基金项目(91224006,61173063,61203284),科技部项目(201303107)资助。

周琦(1991-),女,硕士生,主要研究方向为文本挖掘、关系获取,E-mail:1161514987@qq.com;陆叶(1991-),女,硕士生,主要研究方向为属性知识的获取;李婷玉(1991-),女,硕士生,主要研究方向为兼语句知识获取;王亚(1988-),女,硕士生,主要研究方向为自然语言处理;张再跃(1961-),男,教授,主要研究方向为基础数学、智能信息处理;曹存根(1964-),男,研究员,博士生导师,主要研究方向为知识获取与共享、文本挖掘。

体之间的一一映射机制。在现有的地理本体中,还不能完备且详细地说明地理实体位置关系特征,这样的地理本体还不是严格意义上的地理本体<sup>[3]</sup>。目前,GIS软件中地理实体位置关系也十分有限,深入研究空间关系是实现GIS空间推理的基础<sup>[4]</sup>。国内外很多专家学者相继开展了相关工作,取得了一定的研究成果,如三元地理空间本体SWETO-GS、通过扩展OWL来使地理本体描述空间特征<sup>[5,6]</sup>等。其中基于语义的内容识别方法<sup>[7]</sup>是当前研究的热点。另外,我们课题组的成员在地理实体位置关系获取上积累了大量的工作。张德海结合本体论获取城市知识<sup>[2]</sup>;李宇提出了一种基于句型的古地名及其地理位置关系获取的方法<sup>[8]</sup>;姜琳提出了从文本知识获取的角度获取地理实体位置关系<sup>[9,10]</sup>。

但是,根据分析,现有的研究工作存在以下不足:

(1)获取到的地理实体位置关系表达形式通常以词为单位,相对简单,不能表达如“与…以…分界”的复杂关系。

(2)地理实体位置关系研究没有清晰的层次结构,无法充分地表现出类别关系,不利于知识的共享和重用。

(3)需要依赖大量的其它知识资源对获得的地理实体位置关系进行验证,甚至需要人工干预才能确保一定的精度。因此这些方法缺乏可升级性。

针对以上不足,做出以下改进:

(1)丰富了位置关系的表达形式,不仅能够处理词,还可以处理短语形式的复合关系,从而增加了地理实体位置关系的多样性和获取数量。

(2)构建文法模式时,对不同的地理实体位置关系进行分类,形成层次的树状结构。这样,不仅可以清晰地描述地理实体位置关系的类别,还增加了文法的可维护性。

地理实体位置关系知识的获取需要在彻底理解的基础上进行,所以不仅需要从语义的层面,还需要从知识的层面来增强对关系的理解能力。句法分析<sup>[11]</sup>和语义分析<sup>[12]</sup>技术是解决该问题的重要手段。语义文法<sup>[13,14]</sup>能有效地实现语义分析,也能与句法分析紧密结合,准确理解地理信息知识文本,并且无二义性。

本文提出了一种基于语义文法的地理实体位置关系精准分析体系,首先构造地理实体位置关系语义文法(Geographical Relationship Semantic Grammar,GeoRSG),该文法以地理领域文本为主要研究对象,支持获取地理实体位置关系,且具有层次明确的分类结构;然后实现地理实体位置关系语义解析器(GeoRSG Parser),利用GeoRSG对文本进行解析,提取地理实体位置关系,将无结构的地理信息知识转化为有结构的位置关系知识,并以形式化的谓词表示。

其中,GeoRSG是为实现获取地理实体位置关系知识而设计的文法模式,可以满足在地理领域中不同结构的文本处理需求。其主要方法是通过关键词设计构造句型,再根据句型的获取功能的不同分类处理句型,对原始语料进行匹配,用形式化的一阶谓词逻辑描述地理实体位置关系,得到地理实体位置关系知识。

本文第2节给出GeoRSG中语义类以及语义文法的相关定义;第3节详细介绍GeoRSG的构建方法,包括框架的搭建以及语义文法的设计,并介绍位置关系知识谓词形式的获取;第4节介绍GeoRSG Parser的实现;第5节给出实验结果以及结果分析;最后全面讨论本文提出方法的优缺点,总结全

文。为了说明该方法的实用性,本文以状态分类体系中典型的“邻近关系”为例来描述地理实体位置关系的获取方法。

## 2 预备知识

语义类的构建以及文法产生式的设计是GeoRSG的重要内容,下面对文法做简单介绍,并给出语义类以及语义文法的相关定义。

### 2.1 文法简述

所谓语言的分析<sup>[15]</sup>,就是将语言分解成词、短语等很小的部分,然后分析它们的关系,以便了解句子表达的含义。汉语的研究一般分为4层:词法分析、句法分析、语义分析、语用分析。

形式文法是句子结构分析的一个重要手段<sup>[16,17]</sup>,经常用于自然语言处理。形式文法和自然语言的文法很相似,是描述某个字母表中有限长字串集合的一种方法。它的基本思想是从一个初始符号出发,根据产生式规则,生成一个字串的集合,其中产生式规则指定了符号组合的替换规则。根据形式文法,可以将无结构的自然语言转变为结构化的语法分析树。

语义文法(Semantic Grammar,SG)是Burton<sup>[18]</sup>首次提出的概念,它是一种上下文无关文法。语义文法和其他文法的主要区别在于文法中的非终结符被定义了语义内容,语义文法可以包含语义层次上的非终结符,也可以包含句法和语义层次上的非终结符。

如何用语义文法来表述包含地理实体位置关系的文本是本文研究的关键内容,是对地理实体位置关系进行精准语义分析的基础。相比于传统的基于句法文法的解析结果,语义文法可以获取到文本内容的语义信息,而不局限于句子的句法信息。但是语义文法在文法规则上增添了语义信息,导致文法规则的数量较多,而且一般文法之间没有联系,结构缺乏层次。为了使构建的文法通用于地理领域,且避免引入大量非地理知识的干扰,结合本体的思想,提出“地理实体位置关系语义文法”。GeoRSG结构中突出地理领域位置关系的特点,句型中包含地理领域的语义信息。通过GeoRSG,可以确定地得到文本所表达的语义。因此,GeoRSG是对地理领域文本进行精准语义分析的依据。只有设计出结构清晰、表达完全的文法,才能正确和高效地获取知识。

### 2.2 语义类的定义

将所有的语义归为两类范畴:事件范畴和状态范畴<sup>[19]</sup>。事件范畴是指表达不同事件内涵、语言表达、事件发生的前提和后果、事件之间的关系等内容。状态范畴描述了事物的某一属性特征,表示事物的性质和关系的内容,以及成立的前提和后果。状态范畴是一种静态的属性特征,不会造成其他事物的任何变化。

GeoRSG是用于描述地理实体位置关系的文法,属于描述实体间的静态位置关系特征,属于状态范畴。给出以下状态语义类的定义。

定义1 状态语义类命名为“defcategory〈状态〉”,包含唯一的语义描述,其框架由定义、语义类语句、谓词、例句组成。

“定义”,定义出该状态类所描述的地理实体位置关系,表达的含义唯一且无二义性。

“语义类语句”:所有描述“定义”中地理位置关系的所有文法产生式集合。

“谓词”:是对状态的逻辑描述,即知识的结构化描述。

“例句”:表达该类地理实体位置关系的实例。

也就是说,每一个 defcategory 描述唯一的地理实体位置关系,即每一个 defcategory 中的不同文法产生式包含相同的关键词类,表达单一的内涵。

下面,给出一个具体的状态语义类框架例子。

```
defcategory 毗邻类:
{
定义:某个地区或自然物与另一个或多个地区、自然物边界接壤。
语义类语句:
<毗邻类语句>::=<〈当事:地区|自然物〉[〈方位词类〉]>〈毗邻词类〉〈连事:地区|自然物〉
|〈〈当事:地区|自然物〉[〈方位词类〉]>〈并列连词词类〉〈同事:地区|自然物〉[相]>〈毗邻词类〉
谓词:〈方位词类〉毗邻(当事,连事)
谓词:〈方位词类〉毗邻(当事,同事)
例句:香港 毗邻 广州。
例句:庄河市 东 与 东港市 毗邻。
}
```

图1 状态语义类框架例子

## 2.3 文法的定义

GeoRSG 的文法产生式是由语义非终结符、词类非终结符、常量终结符连接组成。

### (1) 语义非终结符

语义非终结符包含事元角色和类型。

将句子根据不同的“语意角色”进行系统分类。句型主要由主语、述语、宾语等组成,一个句型代表一种句子类型。一个句模描述一类“句意”,代表一个句子的模式。将句模分为以下语义单位:概念(concept)、事元(argument)、事件(event)。一个句模包括“中枢事元”和“周边事元”。一个“事件”由一个“中枢事元”和若干个相关的“周边事元”组成。中枢事元和周边事元所充当的语义角色分别叫做“中枢角色”和“周边角色”。中枢角色包括:状态、心理、关系等;周边角色是事件或者事件的参与者所涉及的情境,包括主体、客体、系体、邻体等<sup>[20]</sup>。根据参与者不同的语义事件,定义了30个事元角色:施事、当事、感事、领事、受事、客事、成事、致事、源事、宿事、经事、向事、位事、同事、用事、涉事、属事、类事、连事、缘由、时间、空间、意图、范围、物量、时量、频量、属性、值事、隐施事。

地理实体位置关系文法描述空间实体状态范畴的特征,主要用到表1中几个事元。

表1 GeoRSG 主要事元

事元	定义
当事	非可控事件的主体,或各种关系的主体
源事	事件中作为来源的邻体或时空的起点
宿事	主体运动后到达的地方
经事	主体所越过的空间或经历的过程
位事	主体始终所在的地方
同事	事件中的伴随者、排除者、相关者或作为基准的邻体
连事	关联关系中的相关者或牵连者

类型是指充当事元角色的实体的类型集合。

事元角色定义和事元角色选择是构建文法的关键。文法

表示具有唯一性、一致性、搭配性的特性。

### (2) 词类非终结符

首先看以下例句。

例句:香港 毗邻 广州。

例句:保安族地区 东南 与 临夏县 接壤。

可以发现上面两个句子中的“毗邻”、“接壤”两个词语提示了地理实体概念之间的位置关系,因此给出以下定义,称它们为关键词。

定义2 对于句子中提示了某种特定的地理实体位置关系信息的词语,可称之为关键词。

结合大量的实际语料,可以总结出很多能提示地理实体位置关系的关键词,例如:“位于”、“毗邻”、“发源”、“延伸”等。通过这些关键词,可以从一定程度上判断出该句中包含的地理实体之间的位置关系知识。可见,需要挖掘的知识在句法上通常都具有一定的规律。

为了更全面地覆盖包含地理知识的句子,给出以下“词类”的定义。

定义3 词类表示意义相近、出现位置相似的词语的集合。

词类的补充不需要修改句型,直接修改词类定义即可。这样,可以使一个文法产生式匹配多个句子,提高文法系统的利用率。

一个关键词词类所表达的地理实体位置关系是相似的。

### (3) 常量终结符

常量终结符是由一个或者多个词语组成的字符串,一般不能再扩展或分解。

例如“隔\_为邻”状态的文法产生式“〈当事:自然物|地区〉〈并列连词词类〉〈同事:自然物|地区〉[之间]〈隔词类〉〈客事:自然物|地区〉〈为邻词类〉”,“〈当事:自然物|地区〉”、“〈同事:自然物|地区〉”、“〈客事:自然物|地区〉”为文法中的语义非终结符,其中“〈当事:自然物|地区〉”表示对象的类型是“自然物”或“地区”;“〈隔词类〉”、“〈为邻词类〉”为词类非终结符;“[之间]”为常量终结符。

## 3 GeoRSG 的设计

设计 GeoRSG 时,为了使文法质量更高,总结出以下设计原则以及设计方法。

### 3.1 GeoRSG 的设计原则

为了使文法模式具有更好的可读性、可用性和可扩展性,设计 GeoRSG 时遵循以下原则。

#### 3.1.1 语料采集原则

尽量搜集内容广泛的地理知识语料,包含不同结构、不同含义的句子。只有准备的语料充分,才能够使文法的覆盖广、解析能力强。

#### 3.1.2 框架搭建原则

构建文法时,使文法结构清晰,层次分明。保持文法的可扩展性,也就是说无需大量修改已有的文法,就可以在现有文法的基础上写入新的文法,且保证已有的文法性能不受影响。

#### 3.1.3 文法的设计原则

(1) 保证语义类描述内容的唯一性、无二义性,且不同的语义类表达的含义具有不重叠性,同时符合一般描述地理实体位置关系的表达习惯。例如:“邻近”不仅表示“实体1”毗

邻“实体2”，也可表示“实体1”和“实体2”遥遥相望，所以定义语义类描述地理实体位置关系时，需注意命名表达的含义是否明确。

(2)自然语言具有灵活性、复杂性等特点，合理地将语句中无用的知识作为无关内容。

(3)保持文法的可读性，便于 GeoRSG 的维护。

(4)选择语义非终结符要准确。事元表达该成分在句子中的身份，根据句子中的情景选择合适的事元对文法产生式的理解有很大的帮助作用。

(5)在不影响功能的情况下，减少文法的冗余，增强 GeoRSG 文法的匹配效率。

(6)尽量提高文法的泛化能力，扩大文法的覆盖范围。

### 3.1.4 词类非终结符的设计原则

(1)非终结符命名要规范。非终结符的命名需要符合习惯表达方式，且确保其无二义性，能明确表达该非终结符内容的含义。

(2)尽量避免非终结符的冗余。非终结符需要尽可能保持精简，冗余的非终结符会对文法的质量产生严重的影响。

(3)非终结符中内容的含义要一致。为了更好地维护 GeoRSG 中的非终结符，非终结符内容所表达的含义需保持在一定的精度范围内。

## 3.2 GeoRSG 的设计方法

我们总结出 GeoRSG 的设计步骤，如图 2 所示。

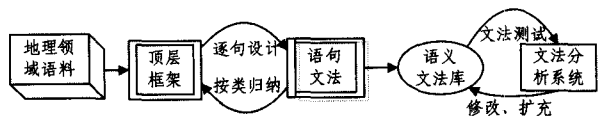


图 2 GeoRSG 设计步骤

步骤 1: 获取地理领域的 Web 语料。

步骤 2: GeoRSG 框架的搭建。分析地理领域自身的特点，对整个领域进行不同语义层次的划分，然后结合地理领域语料进行修改和扩展。在不同的位置关系结构层次下构建相应的地理位置关系语义类，将所有的语义类区分开。

步骤 3: 设计语句文法。即对步骤 1 中的语料以句子为基本单位，按中心词归纳总结相似的句子，设计地理实体位置关系文法，并将文法进行归纳和整合，放入 GeoRSG 框架中相应的分类层次中。

步骤 4: 文法测试。根据实验结果，对未匹配到的语料和匹配错误的语料再次进行分析、修改和扩充 GeoRSG，从而增加 GeoRSG 的准确率以及覆盖率。

### 3.2.1 GeoRSG 框架的搭建

在地理本体中，层次关系在地理实体位置关系树状结构中非常重要。地理实体位置关系的系统划分，不仅可以使方向组合以及拓扑关系模型具有层次性，易于文法的维护，还可以使关系之间的联系一目了然，为地理实体位置关系的研究提供可靠的基础。

GeoRSG 是一个多层次的地理实体位置关系体系文法。

根据上述框架搭建原则，采取从上向下的方式构建语法树，将地理实体位置关系分类。首先通过分析地理实体概念间的语义关系，可以大致对地理实体位置关系进行划分，然后再通过总结大量具有代表性的地理领域语料，根据位置关系的特征对地理实体位置关系框架进行修改和扩充，最后对每

个层次再进行细分划分，得到一个层次分明的地理实体位置关系结构。

GeoRSG 框架具有覆盖率全面、层次清晰分明的特点。

空间位置关系主要是由地理实体的形状和几何特征位置决定的。基本地理实体分为：点、线、面域 3 类。本文根据空间对象的相交、相切和相离等拓扑关系对整个地理领域位置关系进行划分，将空间关系分为以下几类：

(1)表示包含或者提及的顺序关系，例如位于内部、外部等。

(2)表示邻近或者模糊的拓扑关系，例如毗邻、接近等。

(3)表示相交的关系，包括线与线、面与线的相交，例如起始、终止、经过、相交关系等。

(4)两个空间实体的距离关系，用某种度量方式描述两个对象之间的空间关系，如长度、车程等。

空间关系除了拓扑、距离等关系，主要还包括方位关系。方位关系存在于上述 4 个分类中。

将所有的地理实体位置关系分为以下 9 种关系。划分的一级层次关系结果如图 3 所示。

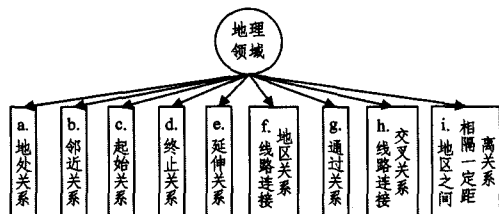


图 3 地理实体位置关系分类框架

图 3 中每个层次还可按照相同的方法继续细分语义类，直到分到位置关系的最细层次，每个层次表达唯一的地理实体位置关系。

下面以邻近关系为例。

邻近有两层含义，分别是表示位置上接近关系和附近关系。“位置上接近关系”包含相切关系，“附近关系”表达两地相离且距离很近。以“毗邻关系”表示“位置上接近关系”，“靠近关系”表示“附近关系”。结合语料可以发现，语料中出现跨某地与另一地邻近的关系出现频繁，且毗邻关系和靠近关系并不能直接地表达，所以我们补充“跨某地与另一地邻近关系”。邻近关系分类如图 4 所示。

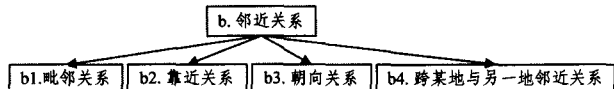


图 4 邻近关系分类

### 3.2.2 文法设计

文法模式是影响文本处理的关键内容，本文整个分类体系，根据中文文本的特点，将语义、语用等信息融入语法模式，做到语义、语法和知识相结合。用户可以根据不同的需求和处理对象的不同来自定义文法模式，使其更有效地处理各种文本、语料，这不仅提升了文法的灵活性，还提升了匹配的准确率。

在设计文法时，通过总结，形成以下文法规则设计方法。

#### 1. 文法产生式的设计

根据文法的设计原则(2)一(6)，文法产生式的设计具体使用了以下几种方法。

(1)在一定的程度上，只需要抓住语言所表达的重点，忽

略一些不重要的修饰成分。人们在理解自然语言文本时,由于个人的行为目的、行为方式导致个体认知差异而形成不同的语义理解<sup>[21,22]</sup>,因此很难逐句逐字地对其进行描述。

(2)文法产生式采用BNF语法,“::=”表示“被定义为”;尖括号(<>)中的内容表示必选项;方括号[ ]中的内容为可选项;竖线(|)表示前后内容的“或”的逻辑关系。

(3)编写文法时,保证文法产生式的可读性。即保证文法结构清晰,并且容易理解。从文法产生式中语义非终结符的组成即可看出句子所表达的地理实体位置关系。

例如,对于句子“我国北部与中国云南接壤。”设计这样的文法产生式:“<当事:自然物|地区><方位位置短语><并列连词><同事:自然物|地区><接壤词类>”。该文法结构明了,便于让人理解,并且很容易看出该句表达的地理实体位置关系。

(4)文法产生式中的语义非终结符严格按照语句中地理实体概念所担当的语义角色来选取。在GeoRSG中,通过某种或几种不同事元的组合,可以初步得到两者的逻辑关系。

(5)编写文法时,可选项的使用可以减少产生式的数量,从而使文法处理的效率更高,同时大大增加文法的利用率,节约空间。使用可选符号时,可以通过不加该非终结符后语句是否通顺来判别。一般情况下,修饰语、状语等非终结符是可选的,而提示某种特定的地理知识的非终结符不可选。

(6)一个状态类中可能会包括多个文法产生式,也就是说每个关键词可能存在于多种结构的句子中,同一个状态类中所有的文法产生式表达状态的内涵是一样的。文法产生式如果不全,就必然会丢失很多包含知识的句子,因此需要总结每一个关键词词类在语义类中可能存在的表达,并尽可能地将它们在句型中表示出来。

通过改变句子结构、联想或者在语料中搜集相关含义的句子的方式来扩充文法。

如“与\_隔\_为邻”状态,有两个基本的文法,如文法1和文法2所示。

文法1:<与\_隔\_为邻类语句>::=<当事:自然物|地区><并列连词词类><同事:自然物|地区>[之间]<隔词类><客事:自然物|地区><为邻词类>

文法2:<与\_隔\_为邻类语句>::=<当事:自然物|地区><隔词类><客事:自然物|地区><并列连词词类><同事:自然物|地区><为邻词类>

(7)文法递归的使用。为了更灵活地匹配语料,合理使用递归可以让文法更有灵活性,而且能提高文法的匹配效率。例如:“山东半岛内陆与河北、河南、安徽、江苏毗邻。”可以描述为“<毗邻类语句>::=<当事:地区>和<同事:地区>[相]<毗邻词类>;<同事>::=<名词性短语>|<名词性短语><并列连词词类><位事:地区>”。

## 2. 词类非终结符的设计

词类非终结符的设计是GeoRSG中一个非常重要的部分,也是评价文法的一个重要的指标。在设计文法的同时,也对词类非终结符进行设计。

在词类非终结符的设计原则的基础上,总结出以下设计方法以及注意事项。

### (1)词类非终结符设计的设计方法

1)利用同义词林或者联想的方法,将每一类词语做相近词的补充和常识的扩充。如关键词“毗邻”可以扩展为“毗邻|

毗|紧邻|濒临|邻接|连接|连通|通连|沟通”,这里所扩充的词语集合即为“词类”,用一个具有代表性语义的词语来命名该词类,即该命名能够直观地反映此非终结字符的含义,例如上述可命名为“<毗邻词类>”。

2)除了关键词组成词类非终结符,还有一些使用率较高的词语,此类非终结符除了动词之外,还可以是名词、形容词性的词类,例如“方位词类”是一个描述地理领域方向词的集合。

3)词类非终结符合理地使用递归。

地理领域中,空间中的方位关系是一种常识空间认知的关系,也是空间关系研究的重要内容之一。按照人类的常识,方位关系可以分为:东(E)、南(S)、西(W)、北(N)、东北(NE)、东南(SE)、西南(SW)和西北(NW)。同时,方位关系还包括更高级的语言表达,包括“偏北”、“靠南面”、“最南方”等。为了更清晰全面地描述语料中对方位的表达,使用递归的方法,例如“<方位位置短语>”的设计。

<方位位置短语>::=[最]<方向词类>[<部位词类>]|<偏+方位短语>[<部位词类>]|<方向词类>[<部位词类>]|<偏+方位短语>|之<方向词类>|以<方向词类>

<偏+方位短语>::=偏<方向词类>[部]|靠<方向词类>

<方向词类>::=东|南|西|北|东南|东北|西南|西北

<部位词类>::=<内部部位词类>|<边缘部位词类>

<内部部位词类>::=部|郊|角|隅|端|麓|坡|区|段|疆|首|大门|沿|端

<边缘部位词类>::=侧|边|口|方|面|方向|翼|角|缘|延|岸|点

### (2)词类非终结符设计的注意事项

1)构建文法产生式时,保证所选择的词类非终结符中任何一个词语都适用于该文法。<毗邻词类>表示对“毗邻|毗|紧邻|濒临|邻接|连接|连通|通连|沟通”的引用。

例如词语“相接”和“毗邻”意思相近,若将“相接”放入“<毗邻词类>”中则不妥,因为文法产生式“<当事><相接><连事>”不通顺,不符合常用表达方式。

2)词类非终结符的合并。

若两个词类非终结符出现的位置相似,并且可以互相替换,那么可以考虑将这两个词类非终结符合并。

如果两个词类非终结符出现交叉的部分,首先考虑交叉部分的词语是否含有二义性,若有二义性,则允许该词语存在于两个含义不同的词类中,表达不同的语义;若无二义性,则考虑两个词类非终结符是否可以合并。

3)词类非终结符的拆分。

有时词类非终结符包含的含义较广泛,为了使其描述的含义更细化,需要对其进行拆分。这样,可以确保描述的含义单一且无异义。

## 3.3 GeoRSG 知识谓词

知识表示的多样性会造成知识库的异常现象,且会给知识一致性的检查造成一定的困难。逻辑谓词解决了自然语言的模糊性、不一致性等问题,可以清楚地描述位置在空间上的拓扑关系。

地理实体位置关系获取的研究主要集中在上下位关系和部分-整体的获取上。

定义4 二元谓词“R(G1,G2)”表示两个地理实体之间的位置关系,其中“G1”和“G2”分别代表两个地理实体,“R”代

表“G1”和“G2”之间具有的位置关系。类似地，三元谓词“R(G1,G2,G3)”表示3个地理实体之间的位置关系。

选用关键词作为谓词形式的名称，可以直接明确地看出句子所表述的状态。

例如：“香港 毗邻 广州。”这条知识的谓词表示为：

谓词1：毗邻(香港,广州)

获取的地理实体之间的位置关系绝大多数都是两个实体之间的关系，也有一小部分是多元实体之间的关系，对于这种情况，将其分为两种情况处理。

(1)仍统一转化为二元谓词形式表示，前提是“G2”部分语义上可以看成是一个实体集合。如：

谓词2：毗邻(山东半岛,{河北,河南,安徽,江苏})

(2)G2和G3语义上不可作为一个实体集合，使用三元谓词。如：

谓词3：经过\_到达(当事,经事,宿事)

到目前为止，设计了261个地理位置相关的GeoRSG谓词。

表2给出部分GeoRSG谓词以及相应的谓词解释。序号1—序号12为常见谓词，序号13—序号16为不常见的谓词。

表2 GeoRSG部分谓词

序号	谓词	解释
1	位于(当事,位事)	“当事”位于“位事”
2	位于_(方向词类)(当事,位事)	“当事”位于“位事”某方位
3	位于_(以内词类)(当事,位事)	“当事”位于“位事”以内
4	毗邻(当事,连事)	“当事”毗邻“连事”
5	(方位短语)毗邻(当事,连事)	“当事”某方位毗邻“连事”
6	与_为邻(当事,同事)	“当事”与“同事”为邻
7	发源(当事,源事)	“当事”发源于“源事”
8	(方向词类)_起于(当事,源事)	“当事”某方位起于“源事”
9	流入(当事,宿事)	“当事”流入“宿事”
10	贯穿(当事,经事)	“当事”贯穿“经事”
11	长度(当事,物量)	“当事”的长度为“物量”
12	方位位置(源事,宿事,<方向词类>)	“源事”位于“宿事”某方位
13	与_隔有(当事,同事,客事)	“当事”与“同事”之间隔有“客事”
14	位于_接壤地带(当事,位事1,位事2)	“当事”位于“位事1”和“位事2”的接壤地带
15	经过_到达(当事,经事,宿事)	“当事”经过“经事”到达“宿事”
16	与_以_分界(当事,同事,客事)	“当事”与“同事”之间以“客事”为界

在GeoRSG中，每一条文法产生式都对应相应的谓词表达形式。但是，自然语言转换成谓词逻辑的方式并不唯一，例如：语句“庄河市东与东港市毗邻。”和“庄河市东毗邻东港市。”在语义上等价，都适用于表2中序号5。因此，不同的文法产生式可能对应相同的谓词。

根据上述谓词，正在设计一个地理位置公理系统。公理是进行知识验证的依据。运用这些公理，可以进行地理实体位置关系知识的推理和验证，同时也可以用来检查位置关系的一致性和完整性。

在谓词逻辑的基础上构造地理实体位置关系公理和自定义推理规则，可以更清晰地表达语义。

将地理实体位置关系公理分为“等同语义关系类”、“矛盾语义关系类”、“拓扑关系类”、“方位关系类”、“拓扑关系 & 方位关系类”。具体的方法和公理将在另文给出。

下面给出上述5类公理分别对应的例子。

公理1 与\_隔有(X,Y,Z)→位于\_之间(Z,X,Y)。

公理2 毗邻(X,Y)→非位于内部(X,Y)。

公理3 位于\_以内(X,Y)∧毗邻(Y,Z)→凑近(X,Z)。

公理4 位于\_东部(X,Y)∧位于\_东部(Y,Z)→位于\_东部(X,Z)。

公理5 南边\_毗邻(X,Y)→位于\_北边(X,Y)∨位于\_西北边(X,Y)∨位于\_东北边(X,Y)。

### 3.4 GeoRSG测试

调用文法分析系统对文法进行测试，并且根据文法匹配的结果对文法进行错误分析以及完善；同时，对未能匹配而且含有知识点的语料重新进行分析、修改或者扩充文法库，使文法库覆盖更多的地理实体位置关系知识。如果文法不全，就必然会导致覆盖率低。

GeoRSG的设计是一个不断补充和完善的迭代过程。

## 4 GeoRSG Parser的实现

本文研究实现了基于语义文法的地理实体位置关系解析器GeoRSG Parser，它是一种面向文本知识处理解析器，为GeoRSG提供了运行平台。

GeoRSG Parser采用多主体<sup>[23]</sup>的思想，完成对地理实体位置关系文本的处理。每个主体可以根据地理领域文本的特征采取不同的处理方式，并且它们功能专一，相互协作，最终得到知识抽取的结果。

GeoRSG Parser首先输入地理领域网络文本以及GeoRSG，然后以GeoRSG为基础，将每一条文本解析成对应的语义语法树，最后根据GeoRSG中每条语义文法所对应的谓词，执行主体定义中的语义动作，获取地理实体位置关系知识元组，从而实现对本体语料的自动处理。

GeoRSG Parser的执行过程如图5所示。



图5 GeoRSG Parser的执行过程

GeoRSG Parser解析前，需要进行以下预处理。

(1)切分语料。根据主体的定义，以句为单元对话料进行切分，得到句子的集合。

(2)指代消解。指代在地理领域语料中普遍出现，为了获取更全面的地理实体位置关系，需要对话句进行指代消解预处理。例如：“保加利亚共和国位于巴尔干半岛东南部，北与罗马尼亚隔多瑙河相望。”处理后分为两个句子，分别为“保加利亚共和国位于巴尔干半岛东南部。”、“保加利亚共和国北与罗马尼亚隔多瑙河相望。”

(3)分词和词性标注。利用条件随机场训练的分词模型，对话句进行分词和词性标注。例如“保加利亚共和国北与罗马尼亚隔多瑙河相望。”经过分词处理后的结果为：“保加利亚/ns 共和国/n 北/f 与/p 罗马尼亚/ns 隔/v 多瑙河/ns 相望/v。 /w”

GeoRSG Parser分为以下2个主要模块。

(1)文法匹配模块。功能：调用合适的主体解析文本片段，并给出解析树。

以上述典型的包含地理实体位置关系的例句来演示结果。

例1：“保加利亚/ns 共和国/n 北/f 与/p 罗马尼亚/ns 隔/v 多瑙河/ns 相望/v。 /w”

该语句表达“与\_隔\_相望”关系，GeoRSG Parser在GeoRSG的基础上对其进行模式匹配，得到一棵完整的语义

解析树,如图6所示。

```

text: /w 保加利亚/ns 共和国/n 北/f 与/p 罗马尼亚/ns 隔/v 多瑙河/ns 相望/v. /
w
<事件类语句>:[PARSED SUCCESS]:
<事件类语句>:[PARSED SUCCESS]:
<状态关系_关联语义类>:[PARSED SUCCESS]:
<状态关系_关联语义类_物体在空间上的关系>:[PARSED SUCCESS]:
<物体在空间上的关系_非接触的实体间的关系>:[PARSED SUCCESS]:
<非接触的实体间的关系_邻近关系>:[PARSED SUCCESS]:
<与*以*分界类语句>:[PARSED SUCCESS]:
<与*隔*相望类语句>:[PARSED SUCCESS]:
<句首修饰短语>:[PARSED SUCCESS]:
<当事:自然物&地区:与*隔*相望>:[PARSED SUCCESS]:
<名词短语>:[PARSED SUCCESS]:
<名词短语_避免递归>:[PARSED SUCCESS]:
<基本名词短语>:[PARSED SUCCESS]:
>>:[PARSED SUCCESS]:保加利亚
的:[PARSED SUCCESS]:
<名词短语_避免递归>:[PARSED SUCCESS]:
<基本名词短语>:[PARSED SUCCESS]:
>>:[PARSED SUCCESS]:共和国
<方位位置短语>:[PARSED SUCCESS]:
<方向词类>:[PARSED SUCCESS]:
北:[PARSED SUCCESS]:北
<并列连词类>:[PARSED SUCCESS]:
与:[PARSED SUCCESS]:与
<同事:地区:与*隔*相望>:[PARSED SUCCESS]:
<名词短语>:[PARSED SUCCESS]:
<名词短语_避免递归>:[PARSED SUCCESS]:
<基本名词短语>:[PARSED SUCCESS]:
>>:[PARSED SUCCESS]:罗马尼亚
<隔中心词类>:[PARSED SUCCESS]:
隔:[PARSED SUCCESS]:隔
<客事:自然物&地区&水体&山岳&道路:与*隔*相望>:[PARSED SUCCESS]:
<名词短语>:[PARSED SUCCESS]:
<名词短语_避免递归>:[PARSED SUCCESS]:
<基本名词短语>:[PARSED SUCCESS]:
>>:[PARSED SUCCESS]:多瑙河
<道相呼应对中心词类>:[PARSED SUCCESS]:
相望:[PARSED SUCCESS]:相望
    
```

图6 语义解析树

(2)知识抽取模块。功能:从解析树结构中抽取知识。

GeoRSG Parser对“例句1”进行模式匹配后,按照预定义的谓词“〈方位位置短语〉与〈隔\_相望(当事,同事,客事)〉”抽取知识,得到地理关系“北与〈隔\_相望(保加利亚共和国,罗马尼亚,多瑙河)〉”。

## 5 实验

本文利用地理语料来验证上述方法的有效性,并对地理实体位置关系的识别准确率进行实验对比。

### 5.1 实验数据

从中科院 NKI 课题组的测试语料中选取了 80000 条真实的地理领域语句,该语料库是从 Web 网页上获取,并经过去除 html 标记、去除重复网页、去除全英文网页等预处理后的综合网页文本。

GeoRSG 提供了获取地理实体概念及其位置关系的句法信息,共有 97 个语义类,文法数量达到 526 条。GeoRSG 框架的部分公布数据如表 3 所列。

表3 部分关系的详细分层

一级关系	二级关系
地处关系	位于、分布、环绕等 4 个二级关系
邻近关系	毗邻、靠近、朝向等 4 个二级关系
起始关系	发源、起点位于等 3 个二级关系
终止关系类	流入、途径、到达等 3 个二级关系
延伸关系	向_延伸、向_扩展 2 个二级关系
线路连接地区关系	接入某地、线路连接不相邻的实体 2 个二级关系
通过关系	穿行、途径 2 个二级关系
线路连接交叉关系	不同线路汇合、交叉、平行 3 个二级关系
地区之间相隔一定距离关系	相距距离、相距车程 2 个二级关系

利用 GeoRSG Parser,直接获得谓词形式的地理实体位置关系。

### 5.2 实验结果

GeoRSG Parser 利用 GeoRSG 对 80000 条地理领域文本进行解析,从中随机抽取 1000 条语句评价实验结果,最终得到 81 条三元和 816 条二元地理实体位置关系对,总数达 897 条,覆盖率为 89.70%,其中正确的关系对达到 797 条,正确率为 88.85%。部分结果如表 4 所列。

表4 实验结果

例句	地理实体位置关系结果
杨浦区位于上海市北部。	位于_北部(杨浦区,上海市)
广东省地处中国大陆最南部。	位于_最南部(广东省,中国大陆)
陕西省地处东经 105°29'~111°15'。	位于东经范围(陕西省,105°29'~111°15')
河南东接安徽,北接河北,西接陕西,南临湖北。	东接(河南,安徽) 北接(河南,河北) 西接(河南,陕西) 南临(河南,湖北)
哈尔滨市东与海门市隔通肯河相望。	东与_隔_相望(哈尔滨市,海门市,通肯河)
昌江黎族自治县东与白沙黎族自治县毗邻。	东毗邻(昌江黎族自治县,白沙黎族自治县)
东京都东部以江户川为界与千叶县连接。	东部以_为界与_连接(东京都,江户川,千叶县)
百度大厦北至上地北路。	北至(百度大厦,上地北路)
西潼高速公路途经西安、临潼、渭南、华县、华阴、潼关 2 市 4 县。	途经(西潼高速公路,(西安、临潼、渭南、华县、华阴、潼关))
伏尔加河通过伏尔加河-波罗的海运河连接波罗的海。	通过_连接(伏尔加河,伏尔加河-波罗的海运河,波罗的海)

### 5.3 实验结果分析

通过对比分析实验结果,容易看出文法分析的有效性和准确性。

#### 5.3.1 实验对比

姜琳<sup>[9]</sup>通过构建句型的方法对受限语料库进行语义分析,得到候选地理实体位置关系,然后再对这些候选位置关系进行分析验证。本文使用相同的地理语料数据集进行实验,具体实验结果对比如表 5 所列。

表5 实验结果对比(%)

	姜琳 <sup>[9]</sup>	基于语义文法的地理实体位置关系获取
覆盖率	84.22	89.70
正确率	88.84	88.85
三元关系占比	0	9.03

通过本文提出的方法可以更全面地获取地理领域知识。姜琳总结出 170 多个句型模式,只可获取二元地理实体位置关系,句型覆盖率相对较低。GeoRSG 覆盖率高,不仅可以获取二元关系,还可获取复杂的多元地理实体位置关系。从实验结果可以看出,复合地理实体位置关系对占所有关系对的 9.03%,如果缺少这类关系,将会导致知识的缺失。通过这些复杂地理实体位置关系,可以分析出多元地理实体之间更加精确的位置关系。

GeoRSG 中包含了空间关系中所涉及的拓扑关系、方位关系以及距离关系,详细、明确地刻画了不同地理位置关系的表达模式,文法可读性强且结构清晰,覆盖范围更加广泛。

通过以上对比可知,本文提出的方法更具有优越性以及实用性。



地理实体位置关系的谓词表示有利于进一步的处理与研究。该方法具有一定的通用性,还可以用于其它领域关系的获取。

### 5.3.2 错误分析

实验结果还存在不足,具体分析如下。

(1)由于语料是从网页文本中获取的,部分语句不符合正常表达方式,且具有随意性,因此在进行模式匹配时会发生错误。

(2)文法匹配错误。其中包括两点:第一,名词短语识别错误导致匹配失败。名词短语的识别经常会受到分词和词性标注的影响,例如“北方/f 光电/b 项目/n ”/w 建设/v 地点/n 南/f 至/v 旭光/nz 里/f 小区/n 。/w”,其中“北方光电项目”建设地点”应该为名词短语,但是“建设”词性标注为动词,导致名词短语识别错误,匹配时将“建设”误认为关键词。第二,部分语句含有多个关键词,属于较复杂的复合句,这对模式匹配造成一定的困难,导致匹配结果不准确。解决方法是分析该类语句结构,增加新的文法模式,并对错误匹配的文法增强语义约束。

(3)知识表示错误。因为地理实体概念表达的多样性,导致事元内容的获取具有不准确性。解决方法是寻找一种事元验证的方法,提高事元获取正确率。

**结束语** 本文基于语义文法获取分析地理实体位置关系,弥补了传统方法的不足。首先构造了地理实体位置关系文法 GeoRSG,不仅得到文本的句法信息,还得到语义信息;然后设计地理实体位置关系解析器 GeoRSG Parser,从海量的文本中获取地理实体位置关系。实验结果表明,该方法对地理实体位置关系的获取是可行的,并且有一定的效果。

本文的创新点和优点包括:1)用语义文法来表述包含地理实体位置关系的文本,适用于自然语言表达方式多样化的特点,这是对地理实体位置关系进行精准语义分析的基础;2) GeoRSG 适用于地理领域,可以提取对该领域有用的知识,更精确地获取地理实体位置关系信息。

但是,该方法也包含缺点:虽然文法的设计使地理实体位置关系的获取的准确度很高,但是文法的构建都是通过手工操作,耗费较大的人力和物力,因此希望寻找机器学习的方式,以提高效率。

今后工作包括:1)在此基础上继续完善语义文法,保证文法的泛化能力。因为知识的不断更新,需要对本体进行维护,使得获取的本体更加完善。2)结合概率上下文无关文法,统计句子分析的可能性及消除句法分析的歧义。概率上下文无关文法能够处理概率较小的“不合法”的句子,使语法具有一定的容错能力,提高了规则开发的效率。3)对获取的地理知识进行分析和验证,识别地理实体概念的同指关系、地理实体位置关系描述精度等问题。由于无法保证网络信息的正确性,模式匹配获取的地理知识并不完全准确。这些错误的存在影响了知识库的推导和共用。4)引入地理位置公理系统推导并且挖掘隐含的位置关系,同时还可以用已知的地理实体位置关系进行交互验证。

本文总结 GeoRSG 的设计方法为以后的文法设计提供基础,具有重要的意义。

- [1] Han J, Kamber M. Data mining: concepts and techniques[J]. San Francisco, 2001, 5(4): 1-18
- [2] Zhang De-hai. Geographic Knowledge Acquisition and Analysis of Country and Region of NKI [D]. Kunming: Yunnan Normal University, 2002(in Chinese)  
张德海. NKI 国家和地区地理知识的获取与分析 [D]. 昆明: 云南师范大学, 2002
- [3] Pan Yan-qun. Research on Representation and Reasoning of Spatial Relations of Geo-ontology [D]. Jiangxi: Jiangxi University of Science and Technology, 2009(in Chinese)  
潘燕群. 地理本体中空间关系的表达和推理研究[D]. 江西: 江西理工大学, 2009
- [4] Ma Lei-lei. Research on Spatial Relation Ontology Description and Reasoning Mechanism [D]. Henan: The PLA Information Engineering University, 2012(in Chinese)  
马雷雷. 空间关系本体描述与推理机制研究[D]. 河南: 解放军信息工程大学, 2012
- [5] Aleman-Meza B, Sheth A, Halaschek C, et al. SWETO: Large-Scale Semantic Web Test-bed[C]// International Conference on Software Engineering & Knowledge Engineering Workshop on Ontology in Action. 2004: 21-24
- [6] Huang Mao-jun, Du Qing-yun, Du Xiao-chu. Research on Formal Expression Mechanism of Spatial Characteristics of Geo-ontology [J]. Geomatics and Information Science of Wuhan University, 2005, 30(4): 337-340(in Chinese)  
黄茂军, 杜清运, 杜晓初. 地理本体空间特征的形式化表达机制研究[J]. 武汉大学学报(信息科学版), 2005, 30(4): 337-340
- [7] Li Xiang-yong, Tian Peng. Research on Ant Colony Optimization Algorithm for the Open Vehicle Routing Problem [J]. Systems Engineering-Theory & Practice, 2008, 28(6): 81-93(in Chinese)  
李相勇, 田澎. 开放式车辆路径问题的蚁群优化算法[J]. 系统工程理论与实践, 2008, 28(6): 81-93
- [8] Li Yu. Research on Chinese Location Names and Relations Acquisition [D]. Beijing: ICT, 2007(in Chinese)  
李宇. 古地名及其地理位置关系获取的研究 [D]. 北京: 中国科学院计算技术研究所, 2007
- [9] Jiang Lin, Li Yu, Cao Cun-gen, et al. Acquiring Geographical Entities and their Relations from Texts [J]. Computer Science, 2008, 34(12): 151-156(in Chinese)  
姜琳, 李宇, 曹存根, 等. 地理实体概念及其位置关系的获取和验证[J]. 计算机科学, 2008, 34(12): 151-156
- [10] Cao C, Wang S, Jiang L. A Practical Approach to Extracting Names of Geographical Entities and Their Relations from the Web [M]// Knowledge Science, Engineering and Management. Springer International Publishing, 2014: 210-221
- [11] Li Yan-hui, Liu Xiang. Modelling and its Ant Colony Algorithm for Multi-depot Open Vehicle Routing Problems with Replenishment on the Way [J]. Computer Integrated Manufacturing Systems, 2008, 14(3): 557-562(in Chinese)  
李延晖, 刘向. 沿途补货的多车场开放式车辆路径问题及蚁群算法[J]. 计算机集成制造系统, 2008, 14(3): 557-562
- [12] Hansen P, Mladenović N, Pérez J A M. Variable neighbourhood search: methods and applications [J]. Annals of Operations Research, 2010, 175(1): 367-407



- [13] Gendreau M, Potvin J Y. Handbook of Metaheuristics [M]. Springer, 2010; 475-513
- [14] Hansen P, Mladenović N. Variable neighborhood search for the p-median[J]. Location Science, 1997, 5(4): 207-226
- [15] Liu Qun. Chinese Lexical Analysis and Syntactic Parsing Technology Overview [C]// sWcL2002 Lecture. Beijing, 2002 (in Chinese)  
刘群. 汉语词法分析和句法分析技术综述[C]// 第1届学生计算语言学研讨会 (sWcL2002) 专题讲座. 北京, 2002
- [16] Salehipour A, Sörensen K, Goos P, et al. Efficient GRASP+ VND and GRASP+ VNS metaheuristics for the traveling repairman problem[J]. 4OR, 2011, 9(2): 189-209
- [17] Villegas J G, Prins C, Prodhon C, et al. GRASP/VND and multi-start evolutionary local search for the single truck and trailer routing problem with satellite depots[J]. Engineering Applications of Artificial Intelligence, 2010, 23(5): 780-794
- [18] Schittkat P, Kinable J, Sörensen K, et al. A metaheuristic for the school bus routing problem with bus stop selection[J]. European Journal of Operational Research, 2013, 229(2): 518-528
- [19] Wang Ya. Research on Method of Acquiring Commonsense Knowledge Based on Semantic Taxonomy [D]. Nanning: Guangxi Normal University, 2015 (in Chinese)  
王亚. 基于语义分类的常识知识获取方法研究[D]. 南宁: 广西师范大学, 2015
- [20] Lu Chuan, Hou Rui-long, Dong Li-Ping. Basic Words of Modern Chinese[J]. Chinese Teaching in the World, 2000, 4(1): 11-24 (in Chinese)  
鲁川, 侯瑞隆, 董丽萍. 现代汉语基本句模[J]. 世界汉语教学, 2000, 4(1): 11-24
- [21] Yao X, Thill J C. How Far Is Too Far? -A Statistical Approach to Context-contingent Proximity Modeling[J]. Transactions in GIS, 2005, 9(2): 157-178
- [22] Yao X, Thill J C. Neurofuzzy modeling of context-contingent proximity relations[J]. Geographical Analysis, 2007, 39(2): 169-194
- [23] Vlassis N. A concise introduction to multiagent systems and distributed artificial intelligence[J]. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2007, 1(1): 1-71

(上接第 207 页)

由图 6 和图 7 可知, DFTB\* 树具有最低的更新时间 and 更新 I/O 次数。由于内存通过内存寻址方式访问数据, 因此内存数据更新速度快于磁盘数据更新速度。对比算法中, TPR\* 树的更新性能最差。随着移动对象数量增加, 移动对象的频繁更新使得 TPR\* 树中 MBR 出现覆盖重叠, 导致存在多个查询路径, 降低了更新性能。与 B\* 树和 B<sup>dual</sup> 树相比, STRIPES 具有较高的更新时间 and 更新 I/O 次数。由于 B\* 树和 B<sup>dual</sup> 树采用 B+ 树作为基础索引, B+ 树更新时, 只需要单路径查询, 因此, B\* 树和 B<sup>dual</sup> 树取得了较好的更新 I/O 性能。

**结束语** 本文提出一种基于分布式内存数据库的移动对象全时态索引, 针对存储器 3 层结构, 优化索引节点大小。同时, 设计历史数据迁移链并持久化, 从而支持移动对象全时态索引。实验表明, 与 B\* 树、B<sup>dual</sup> 树、TPR\* 树和 STRIPES 算法相比, DFTB\* 树具有较高的查询和更新效率。本文主要研究的是不受空间约束环境下的移动对象索引。现实生活中, 人们真实的运动场景往往局限于道路交通网络中。因此, 下一步的研究工作主要考虑基于道路交通的移动对象时空索引和查询技术。

## 参 考 文 献

- [1] Qiu P, Zhang J, Zeng J. Study on the mobile LBS development model[C]// 2012 IEEE International Conference on Computer Science & Service System (CSSS). 2012; 1070-1074
- [2] Rao B, Minakakis L. Evolution of mobile location-based services [J]. Communications of the ACM, 2003, 46(12): 61-65
- [3] Sun L M, Song B Y, Yu Y X, et al. Cache-conscious index mechanism for main-memory databases [J]. Wuhan University Journal of Natural Sciences, 2006, 2(1): 309-312
- [4] Hankins R A. Effect of node size on the performance of cache-conscious B+-tree[J]. ACM SIGMETRICS Performance Evaluation Review, 2003, 31(1): 283-294
- [5] Saltens S, Jensen C S, Leutenegger S T, et al. Indexing the positions of continuously moving objects[C]// Proceedings of ACM SIGMOD International Conference on Management Data. 2000; 331-342
- [6] Tao Y, Papadias D, Sun J. The TPR\* -Tree: An optimized spatio-temporal access method for predictive queries[C]// Proceedings of the 29<sup>th</sup> International Journal on Very Large Data Bases (VLDB). 2003; 790-801
- [7] Jensen C S, Lin D, Ooi B C. Query and update efficient B+-tree based indexing of moving objects[C]// Proceedings of the Thirtieth International Journal on Very Large Data Bases (VLDB). 2004; 768-779
- [8] Man L Y, Tao Y, Mamoulis N. The B dual-tree: indexing moving objects by space filling curves in the dual space[J]. The VLDB Journal, 2008, 17(3): 379-400
- [9] Chen S, Ooi B C, Tan K L, et al. ST<sup>2</sup>B-tree: A self-tunable spatio-temporal B+-tree index for moving objects[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. 2008; 29-42
- [10] Patel J M, Chen Y, Chakka V P. STRIPES: An efficient index for predicted trajectories[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. 2004; 635-646
- [11] Zhao L, Chen L, Jing N, et al. An efficient moving object index that supports concurrent access[J]. Journal of National University of Defense Technology, 2010, 32(3): 53-59 (in Chinese)  
赵亮, 陈萃, 景宁, 等. 一种支持高效并发访问的移动对象索引[J]. 国防科技大学学报, 2010, 32(3): 53-59
- [12] Rao J, Ross K A. Making B+- Trees Cache conscious in main memory[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. 2000; 475-486
- [13] Chen S, Todd G P B. Improving index performance through prefetching[J]. ACM SIGMOD Record, 2001, 30(2): 235-246
- [14] Rao J, Ross K A. Cache conscious indexing for decision-support in main memory[C]// Proceedings of the International Journal on Very Large Data Bases (VLDB). 1999; 78-89
- [15] Chen S, Jensen C S, Lin D. A benchmark for evaluating moving objects indexes[C]// Proceedings of the International Journal on Very Large Data Bases (PVLDB). 2008; 23-28