计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191000161

基于 Bootstrapping 的水利空间关系词提取

相 颖 冯 钧 夏珮珮 陆佳民
河海大学计算机与信息学院 南京 211100
(1290496228@qq.com)

摘 要 目前,在利用水利领域数据库构建知识图谱的过程中发现水利空间关系词的提取存在以下问题:数据库中水利对象空间关系词较少,难以满足查询需要;水利对象间的关系类型复杂,依靠人工构建太过费力。为了解决上述问题,文中首先从专业性强的高质量水利公文文本中提取空间关系词形成种子集;然后,通过外部词典进行空间关系词的扩展,并结合语料提取面向水利空间关系词的句法模式;最后,通过泛化后的句法模式,对大规模水利文本数据进行空间关系词提取,生成空间关系元组,再将其作为种子集重复上述步骤。该方法使用少量的人工操作便可从语料中获得大量空间语义句法模式以及空间关系元组,逐步扩展构建并最终形成水利空间关系词词典,成为扩充水利对象知识图谱、提升智能检索的准确率的重要支撑。

关键词:关系抽取;水利领域;空间关系;知识图谱
中图法分类号 TP391.1

Extraction of Water Conservancy Spatial Relationship Words Based on Bootstrapping

XIANG Ying,FENG Jun,XIA Pei-pei and LU Jia-min
College of Computer and Information,Hohai University,Nanjing 211100,China

Abstract At present,the following problems are found in the extraction of water conservancy spatial relational words in the process of using water conservancy domain database to construct knowledge map. First,there are few water conservancy object spatial relational words in the database,which is difficult to meet the needs of query. Second,the relationship between water conservancy objects is complex and it is too laborious to rely on manual construction. In order to solve the above problems,firstly,this paper extracts spatial relation words from professional high-quality water conservancy official documents to form seed sets. Then,it expands spatial relationship words through external dictionaries,and combines corpus to extract water-related spatial relationship words Syntactic pattern. Finally,through the generalized syntactic pattern,spatial relation words are extracted from large-scale water conservancy text data,spatial relationship triples are generated,and then used as seed sets. Repeating the above steps can gradually expand and construct water resources. This method can obtain a large number of spatial semantic syntactic patterns and spatial relationship tuples from the corpus with a small amount of manual operations,gradually expand the construction and eventually form a dictionary of water conservancy spatial relationship words. The word dictionary plays an important role in expanding the knowledge map of water conservancy objects and improving the accuracy of intelligent retrieval.

Keywords Relationship extraction,Water conservancy field,Spatial relationship,Knowledge graph

1 引言

水利信息涉及环境、资源、灾害等领域,为解决防汛抗旱、土壤侵蚀等问题,需要及时全面且准确的信息支撑。水利行业一般采用构建水利对象级元数据粒度的资源目录系统[1]来实现信息发布和共享。但是这一系统依据所输入的关键词进行匹配来完成元数据检索,导致检索准确率较低,因此本文试图采用构建面向水利对象的知识图谱,来添加对象间的语义关系,从而提供知识和语义层面的综合查询,为用户的检索提供更加智能、准确、人性化的服务。

针对既有的水利知识图谱,我们还需要补充水利对象间的空间关系。这是因为在现有的水利领域知识图谱中,空间属性,例如经纬度,与数据库中的其他数据一样,只是作为属性值与实体链接[2],缺乏大量空间语义关系,如“流经”“相邻”“位于”“岸别”等,以目前知识图谱的构建方式并不能充分地描述和应用这些数据的空间特性,因此依靠现有的图谱难以

到稿日期:2019-10-24 返修日期:2020-03-20 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划(2018YFC0407901);国家自然科学青年基金项目(61602151);江苏高校文化创意协同创新中心资助项目(XYN1702)
This work was supported by the National Key R&D Program of China(2018YFC0407901),Young Scientists Fund of the National Natural Science Foundation of China(61602151) and Jiangsu Collaborative Innovation Center for Cultural Creativity(XYN1702).
通信作者:冯钧(fengjun@hhu.edu.cn)

进行空间语义上的查询。

由于水利行业的特殊性，从既有水利公文文本中抽取水利空间关系存在一些难点：各类空间关系表述存在专业性、一义多词的特点，如“流经”“贯穿”“分水岭”等都是描述“穿过”这一类关系的词语；现有方法依靠人工构建类型体系太过费力，需要以自动化的方法进行语义标注[3]，但是水利空间关系类型复杂繁多，难以适用于海量数据空间关系自动化标注的要求[4]。为了解决上述两个问题，我们需要使用句子上下文中的词汇来描述实体间的空间关系，建立水利空间关系词典，从而避免预先构建空间关系类型，并选用合适的自动化方法进行面向水利对象的空间语义关系抽取；同时，根据构建的词典进行关系词的扩充，寻找它们的近义词作为空间关系词的候选词，以方便用户进行查询。例如，通过构造少量关系种子或是制定抽取规则，在大规模语料库上半自动或全自动地提取关系描述模式，来构建符合水利业务应用需求的水利空间谓词集合，识别空间关系实体对，节省大量的人工操作，也可以解决目前图谱中空间关系匮乏、现有手段需要人工标注、针对水利业务需求进行空间语义查询[5]等问题，从而实现水利空间关系词的识别与提取，以及基于空间语义的水利数据资源查询。

2 相关工作

信息抽取(Information Extraction)是从自然语言文本中抽取结构化信息的有效方法[6]。信息抽取涉及3个关键技术：实体抽取、关系抽取和事件抽取。其中实体抽取的目的是从文本中识别实体信息，如地名、机构名、人名等，是信息抽取的基础。为了进一步理解自然语言文本信息，有必要基于实体识别来提取实体间的语义关系，即关系抽取。实体关系信息可以形式化为关系三元组〈实体1，关系，实体2〉，其中实体1和实体2是指实体类型，关系是描述两个实体间关系的词语。关系抽取就是从语料中抽取出三元组以提取信息。

Schutz等[7]认为，关系抽取是将实体和描述这对实体的关系进行自动识别，获得相关三元组。结构化与半结构化的关系抽取通常是根据网页结构制定专门的包装器进行抽取，不同的网页需要修改以形成对应的包装器。非结构化的文本关系抽取主要分为传统关系抽取以及开放式关系抽取。

(1)传统关系抽取：主要任务是从语料库中抽取目标关系对。对于这类任务，需要预先制定关系类别和高质量的已标注语料[8]，还需要有公认的评价方法。常用的评测集有MUC，ACE[9]，KBP，SemEval。传统机器学习方法主要使用统计学知识，将三元组转换为高维空间的特征向量，或者通过离散结构将实例直接转换为特征向量，将标注语料训练成分类器以识别实体关系。基于机器学习的方法通常使用不同的机器学习算法来将关系提取问题转化为分类问题，选择代表特征，最后使用分类器来识别实体对间是否具有语义关系[10]。

(2)开放式关系抽取：开放式关系抽取不限制关系类或语料。因此其难点在于如何获得训练语料、如何获得实体的关系类别，以及如何提取各种类型的目标文本关系。Xu等[11]利用依存句法结构，实现了领域自适应关系抽取，并通过自下而上的方法创建规则模板，利用启发式规则消除错误和无效模板，大大改善了关系抽取的效率。

信息抽取的方法一般分两种：统计方法和基于规则的方法。统计方法虽然能获取文本中的规律，但失去了一些语言特征；基于规则的方法更适合水利领域文本信息的专业性、规律性的特点，能得到更好的抽取效果。自由文本信息抽取通常使用模式匹配法，即使用上述规则在模式的指导下完成关系的识别与抽取，需要采用人工或自动的方式获取模式[12]。

半监督关系抽取方法主要有BootStrapping[13]方法，一般人为地建立一些关系实例作为初始种子，然后训练模型或利用模式学习的方法迭代循环搜索，得到关系实例。该方法可以大大减少人为参与和对标注语料的依赖。但基于BootStrapping的方法在找到一定量的关系实例后，很难再继续挖掘，需要进一步选取可靠的关系实例。无监督的关系抽取方法完全依赖文本，通过发现其规律来识别并抽取文本中可能存在的关系。该方法不需要任何手动标注，因此可以适应不断增长的网络文本，在开放领域的研究中应用广泛。然而，由于开放领域关系抽取具有复杂性和多样性，与传统的关系抽取相比，其没有统一的关系抽取标准和评价标准，因此有待研究者进一步的探讨。

本文提出的关系抽取方法将统计和句法模式识别相融合，并结合了词法特征和句法特征。

3 方法

3.1 总体架构设计

空间关系抽取的目的是识别自由文本中水利实体间的空间关系。

本文旨在研究并设计一种面向水利行业的空间关系抽取系统，针对水利公文文本抽取出空间关系三元组，以实现对既有水利知识图谱空间关系缺乏的补全。该系统主要包括：预处理模块、获取种子集模块、空间关系词扩展模块、原始句法模式获取模块、句法模式泛化模块、空间关系抽取模块。本文方法的整体架构如图1所示。

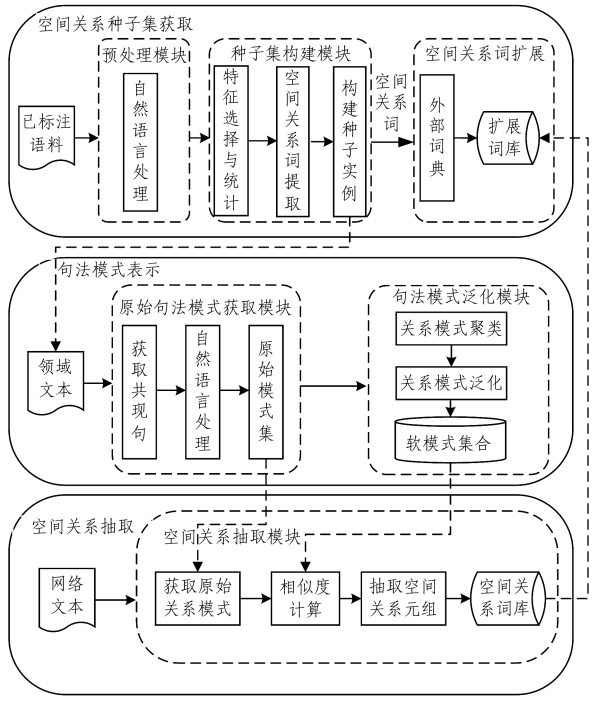

图1 本文方法框架

Fig. 1 Framework of proposed method

首先，本文方法对获取的文本数据进行自然语言处理，生成带自然语言标记的句子集合；标注其中的空间关系词，对句子集合进行特征统计，形成定量化表达，自动挖掘空间关系词，构建种子实例，并根据外部词典构建空间关系同义词集合。

然后，以种子集为输入，结合语料获取共现句，提取面向水利空间关系词的句法模式，对模式进行泛化。

最后，通过泛化后的句法模式，对大规模水利文本数据进行空间关系词提取，生成空间关系元组，并将其作为种子集，重复上述步骤，可逐步扩展种子集并最终形成水利空间关系词库。

3.2　空间关系种子集的抽取

关系种子集的获取过程包括：特征选择与统计、关系词重要度计算、关系词扩展。

3.2.1　特征选择与统计

本文借鉴空间关系文本在词性、词序、距离等方面的规律[14]，为获取水利实体空间关系种子集提供先验知识。通过统计以下 7 个特征来获取空间关系词在句中的分布规律，作为后续提取空间关系词的重要依据。

(1)词性 (POS)。句中的空间关系词是名词、动词、方位词还是其他。例如名词(词性标注为 n)有“干流”“支流”等；动词(词性标注为 v)有“汇入”“流经”等；方位词(词性标注为 f)有“南部”“上游”等。

(2)空间关系词对于水利对象实体的位置 Location (LOC)，即在第一个实体的左边还是第二个实体的右边，或是两个实体中间。

(3)空间关系词左边有连词或者介词时的位置 Location of Conjunction or Preposition (LCCP)，即两个实体左右或是中间。

(4)空间关系词到实体 1 的距离 DIS_1。

(5)空间关系词到句尾的距离 DIS_2。

(6)空间关系词的长度 LEN(以字为单位)。

(7)两个实体间的距离 $DIS(e_1,e_2)$(以词语为单位)。

考虑到非结构化文本数量繁多且构建语料成本高，利用 BootStrapping 技术对小样本随机重采样，来推测总体的统计量，当多次抽样后的总样本规模够大时，统计结果可以代表原始样本的总体水平。首先，利用等间距抽样的方式从文本中随机选择 100 个句子；其次，手工标注语料中的关系词；最后，从已标注样本中随机采样形成规模相同的新样本。重复该过程 10000 次可以得到一个大规模的标注语料，每个标注语料样本大小是 100 条，统计上述 7 个特征值，计算出每个特征的比例及均值。

上述 7 个特征主要用于辅助数据预处理，与传统方法将词法和句法的统计规律直接用于模式匹配不同[15]，这些结果是通过统计学方法将规律定量化表达，用于后续的关系词提取，指导种子集的构建。这些定律不是文本转换的某种形式，而是对实际数据的直观反映。

3.2.2　关系词重要度的计算

根据上述方法得到的统计结果，考虑到词语的词性、位置和距离的重要性，通过计算关系词重要度来获取空间关系词，具体公式如下：

$$wgt(POS_i)=\begin{cases}0.60, & POS_i=n\\ 0.31, & POS_i=v\\ 0.02, & POS_i=f\\ 0.01, & POS_i=\text{other}\end{cases} \tag{1}$$

$$wgt(LOC_i)=\begin{cases}0.71, & I_i>I_{e2}\wedge I_i>I_{ccp}\\ 0.07, & I_i>I_{e2}\\ 0.21, & I_{e1}<I_i<I_{e2}\wedge I_i>I_{ccp}\\ 0.79, & I_{e1}<I_i<I_{e2}\\ 0.14, & I_i<I_{e1}\end{cases} \tag{2}$$

$$wgt(DIS_i)=\begin{cases}0.43, & (I_i>I_{e2})\cap(Len-I_i=1)\\ 0.02, & (I_i>I_{e2})\cap(0<Len-I_i-1<5)\\ 0.06, & (I_i>I_{e2})\cap(Len-I_i-1\geqslant 5)\\ 0.35, & (I_i<I_{e2})\cap(I_i-I_{e2}=1)\\ 0.17, & (I_i<I_{e2})\cap(1<I_i-I_{e1}<5)\\ 0, & \text{other}\end{cases} \tag{3}$$

$$wgt_i=(wgt(POS_i)+wgt(LOC_i)+wgt(DIS_i))/3 \tag{4}$$

其中，wgt_i 表示词语 i 在词性、位置、距离方面的重要性。POS_i，LOC_i，DIS_i 分别表示对应的词性、位置、距离的重要性。I_i，I_{e1}，I_{e2}，I_{ccp} 分别表示词语 i、水利对象 e_1 和 e_2、介词或者连词在句中的索引。Len 表示句子的长度。对于每对水利对象实体(e_1，e_2)，使用式(4)从上下文中选择一个重要度最高的词语作为水利实体间的空间关系词。公式中的数字来源于上述特征的统计结果，如式(1)中的 0.60 表示标注语料中有约 60%的名词为关系词。这些数字随着文本体裁、风格的变化而变化，而非固定的值。

将水利实体对(e_1，e_2) 和从句中抽取的重要程度最高的空间关系词 rel 组合，形成空间关系实例(e_1，rel，e_2)。其中，默认关系元组中，e_1 是首要实体，e_2 是次要实体，如果关系词位于 e_2 的右边，则将 e_1 设置为次要实体，e_2 设置为首要实体。

3.2.3　关系词的扩展

同一个实体对之间的空间关系可以有许多种不同或相近的表达方式，如描述河流之间的上下游关系有下游、中下游，或是流入、汇入等。于是，利用文本信息的冗余性，尽可能多地抽取反映同一类空间关系的词或短语，来对实体对进行标注。如上所述，同一类描述空间关系的词或短语可能是相同或相似的，为了把这些相同或相似的空间关系指示词汇聚在一起，可以使用外部词典对抽取到的关系词进行扩展。将得到的空间关系词汇作为处理对象，并对空间关系词进行扩展，形成一个同义词表，便于后续抽取更多表达同类关系而关系词不同的数据，初步形成一个关系类型体系。

我们借助《同义词词林(扩展版)》[16](下文简称《词林》)的分层结构，以种子中的关系词作为统一描述词，同义同类词作为候选词，建立水利空间关系体系，从而一词多义现象(即一个词具有多个不同层级的语义代码)得到了初步解决。该词库可以通过关系抽取的迭代过程进行不断的扩充。

3.2.4　种子集构建算法

通过提取空间关系词来构造空间关系种子实例的伪代码如算法 1 所示。

算法 1 种子实例构建算法

输入:预处理后的句子 Sentences

输出:种子空间关系三元组集合 Triples

```
1. for each sentence ∈ Sentences do
2.    termList = HanLP(A Processed Sentence which contains entity1 and entity2) //获得词语序列
3.    sequence = deleteStopWord(termList)
4.    i=0
5.    rel = calculateWeight(sequence[i])
6.    for each word ∈ sequence do
7.       i=i+1
8.       if rel < calculateWeight(sequence[i]) then
9.          rel = calculateWeight(sequence[i]) //从语序列中选择一个重要程度最高的词
10.   end for
11.   Instance.setEntity1(entity1) //构建种子实例
12.   Instance.setEntity2(entity2)
13.   Instance.setRelation(rel)
14.   add Instance inTriples
15. end for
16. return Triples
```

3.3 句法模式的获取与泛化

本节主要介绍句法模式表示、词语权重计算、句法模式的获取与泛化。

3.3.1 句法模式表示

本文借鉴关系词、语义、词序等信息[17],使用了以下句法模式表示方法:

$$P=\langle e_1, e_2, r, C, Pos_{e1}, Pos_{e2}, Pos_r\rangle$$

其中,P 表示句法模式;r 表示空间关系词;C 表示句中除去实体以外的词语集合$\{w_1, w_2, w_3, \cdots, w_n\}$,包含其在《词林》中的语义代码以及权重;$e_1$ 和 e_2 分别为两个水利实体类型的标签。在水利领域中,关系类型对实体的对象类型较为敏感,例如:在相交关系中,黄河流域流经青海,"流经"的首要实体类型为流域,次要实体为行政区划,颠倒水利对象的位置则关系矛盾,这在大部分其他空间关系中也是如此。因此在规定敏感位置的关系中,e_1 为首要实体,e_2 为次要实体;Pos_{e1},Pos_{e2},Pos_r 分别表示实体 1、实体 2 和空间关系词的位置信息。

3.3.2 词语权重计算

权重是指空间关系词的上下文内容相对于关系词和水利实体的重要性分析。在获取种子集时使用的是特征词与关系词之间的词语数作为平面距离,没有考虑到语法结构上的距离,因此本文通过计算特征词的权重来体现每个词语对关系词以及实体的重要程度。

通过上下文的信息可以计算出命名实体间的语义关系,那么,具有相同或相似的上下文模式的句子中就可能存在相同的空间关系。研究发现[17-19],句中的词语对实例的描述能力与词语和关系词的相对位置有关,即上下文中距离关系词和实体越近的词语越有价值。传统的相对距离计算需要先进行词法分析获得词性标注序列,再计算两个词的相对距离。但这只考虑了句子的平面结构,没有考虑到句法结构的修饰关系,缺乏语法和语义表达能力。本文通过句法分析后得到的短语结构句法分析树来计算词语的相对位置,考虑到了词语的语义表达能力,能够明显提高空间关系抽取的性能。句法分析包括确定句子的句法结构和词语在句中的依存关系。本文使用的是句法结构分析,以树形结构输出。这种结构可以体现句子的语法关系和词语之间的层次。

相对距离的计算公式如下:

$$D(w_i, w_j)=\frac{Distance(w_i, w_j)}{Depth_{\text{father}}(w_i, w_j)} \tag{5}$$

其中,$Distance(w_i, w_j)$表示短语结构句法树中两个词语间的有向路径长度;$Depth_{\text{father}}(w_i, w_j)$为词语 i 和 j 在树中公共父节点的深度。

文献[20]中关于上下文与关系词间的权重计算方法,是利用句中短语结构与关系词间的结点距离来衡量各短语的权重大小。本文在其基础上提出了面向空间关系句法模式表示方法的词语权重计算方法,计算公式如式(6)所示:

$$D(w_i, C, P)=\frac{1}{n-1}\left(1-\frac{\sum_{core\in P.CORE} D(w_i, core)}{\sum_{w\in C}\sum_{core\in P.CORE} D(w, core)}\right) \tag{6}$$

其中,n 表示 $Context$ 中词的个数,$core$ 表示水利实体或空间关系词,$P.CORE$ 表示关系词和两个实体的集合,$D(w, core)$用于计算上下文词语 w 与两个实体或空间关系词间的语义距离。此处再引入式(4)中词语在词性、位置、距离上的重要性 wgt_i,得到最终词语权重计算公式如下:

$$weight(w_i)=\frac{1}{2}\left(\frac{wgt_i}{\sum_{j\in C} wgt_j}+D(w_i, C, P)\right) \tag{7}$$

通过式(7)计算词语在句中的权重可以用来衡量该词对空间关系描述能力的强弱。

3.3.3 句法模式的获取

首先根据种子集获取对应的共现句,并且对共现句进行分词、去除停用词、词性标注、命名实体识别、生成短语结构句法树等预处理操作。在进行抽取时,句中的上下文信息通常被认为反映了水利对象间的关系,这为空间关系抽取提供了依据。本文使用的是 HanLP 自然语言处理工具,句法树分析利用了 NLTK 的 Stanford CoreNLP。Stanford CoreNLP 是一个自然语言处理工具包,它可以给出词语的基本形式——词性(如公司名、人名等,规范化的日期、时间、数字等),根据短语和语法依赖来标记句子的结构,发现实体之间的关系、情感以及人们所说的话等。Stanford CoreNLP 生成句法树后,可以获得句子语义结构和修饰关系。

在处理完共现句之后,句子变成词语序列,此时对该序列进行句法模式表示,具体方法如下:

(1)利用 2.2 节的种子元组作为输入在语料中获取共现句,对句子进行预处理;

(2)利用自然语言处理工具[21-23]进行词法分析,再使用 Stanford CoreNLP 工具进行句法分析,得到句法树,使用式(5)计算两个词语间的相对距离;

(3)保留该词语序列中的动词、名词、形容词等有效词汇,过滤数词、代词等无意义词语;

(4)对保留的词语序列进行权重计算,并标识每个词在《词林》中的语义代码;

(5)标识两个实体和关系词在句中的位置;

(6)将得到的信息作为句法模式表示并存储。

原始空间关系句法模式获取的伪代码如算法 2 所示。

算法 2　原始句法模式获取算法

输入：种子实例 instance=(e_1，rel，e_2)的共现句 sentence

输出：原始空间关系句法模式 pattern

```
1. termList=HanLP(sentence)
2. pattern.setEntity1(instance.e1)
3. pattern.setEntity2(instance.e2)
4. pattern.setRelation(instance.rel)
5. SegSenten=SenSplitBySpace(termList)
6. StanfordXMLFile=stanfordCoreNLP(SegSenten)
7. WordsList=ValidWords(termList)
8. WeightWordsList=weightAndSem(StanfordXMLFile,Words-
   List)//计算词语序列中每个词的权重以及语义代码
9. pattern.setContext(WeightWordsList)
10. pos1=termList.indexOf(instance.e1)//获得水利实体及空间关
    系词词序
11. pos2=termList.indexOf(instance.e2)
12. posr=termList.indexOf(instance.rel)
13. pattern.setPos1(pos1)
14. pattern.setPos2(pos2)
15. pattern.setPosr(posr)
16. return pattern
```

3.3.4　句法模式的泛化

对每个共现句进行原始句法模式表示后，由于抽象程度不佳，不能直接用于空间关系抽取；并且共现句数量大，句法模式数量也随之增大，泛化程度低，直接进行抽取无法正确地抽取空间语义信息。因此，通过将多个表达同类空间关系的原始句法模式泛化为一个模式，可以减少模式数量并提高抽象程度，有助于后续的空间关系抽取。

首先，需要将表达不同空间关系的句法模式进行划分，这里通过模式相似度计算来对句法模式进行聚类，将每一类中的句法模式抽象生成一个泛化模式。例如，针对 $P_i=\langle e_1, e_2, r, C, Pos_{e1}, Pos_{e2}, Pos_r\rangle$ 和 $P_j=\langle e_1, e_2, r, C, Pos_{e1}, Pos_{e2}, Pos_r\rangle$ 两个模式进行泛化的具体步骤如下。

(1)句法模式聚类。采用层次聚类方法，相似度的计算公式如下：

$$sim(p_i,p_j)=\{sim(C_i,C_j), e_{1i}=e_{1j}\&\&e_{2i}=e_{2j}\&\&C_i\cap C_j\neq\varnothing\&\&ord(Pos_{1i},Pos_{2i},Pos_{ri})=ord(Pos_{1j},Pos_{2j},Pos_{rj})\} \tag{8}$$

其中，$sim(p_i,p_j)$ 指 C_i 和 C_j 两个向量间的余弦距离；$ord(Pos_{1i},Pos_{2i},Pos_{ri})=ord(Pos_{1j},Pos_{2j},Pos_{rj})$ 表示首要实体、次要实体以及关系词之间的相对位置。当两个句法模式间实体与关系词的相对位置相同且实体类型相同，以及上下文中有相同的有效词语时才计算相似度，否则直接认为不相似。

(2)句法模式泛化。经过聚类会形成多个簇，簇内是相似度较高的模式。句法模式泛化是指将每个簇内多个模式泛化成一个抽象模式，以便进行空间关系抽取。具体步骤为：

1)将模式中的词语序列整合成一个序列。

2)更新 pos_1，pos_2，pos_r 的值。抽象模式中的词为簇中模式间的相同词汇，且这些词汇出现的频率大于簇中一半的词语，词的权重取簇内模式中词语权重的平均值，pos_1，pos_2，pos_r 的值取任意一个模式的位置信息。

泛化模块的伪代码如算法 3 所示。

算法 3　句法模式泛化算法

输入：原始空间关系句法模式集合 srcPatternList

输出：泛化后的句法模式集合 AbstractPatternList

```
1. DividedPatternList=divByWordOrd(srcPatternList)
2. for each PatternList ∈ DividedPatternList do//对划分后的每个集
   合的模式进行聚类
3.   for each p1 ∈PatternList do
4.     for each p2 ∈ PatternList do
5.       if p1 !=p2 then
6.         if calculateSimilarityofPattern(p1,p2)>β then
7.           PatternClusterList=cluSrcPattern(srcPatternList)//如
             果模式相似度大于阈值则将句法模式聚类
8.           AbstractClusterPattern=abstractSrcPattern(srcPattern-
             ClusterList)
9.         add AbstractClusterPattern in AbstractPatternList
10.      end for
11.    end for
12. end for
13. return AbstractPatternList
```

3.4　空间关系提取

3.3 节使用句法模式的泛化来获得多个抽象模式，这些抽象模式都可以用来提取空间关系元组。本节重点介绍如何在语料库中利用抽象模式来获取更多的空间关系元组，图 2 为空间关系的提取流程。

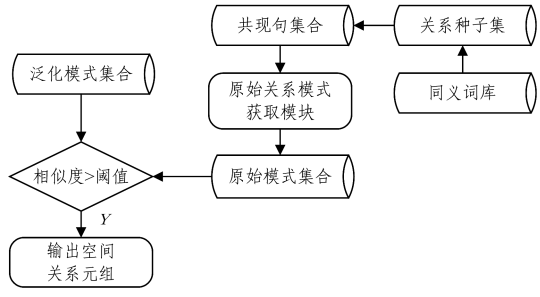

图 2　空间关系的提取流程

Fig. 2　Flowchart of spatial relation extraction

首先，根据空间关系词词表从大规模语料中检索出包含该词的共现句，通过分词、词性标注、去除停用词等预处理，以及上述原始句法模式获取方法得到这些句子的句法模式；然后将生成的每个模式与抽象模式集合内的模式匹配，当原始句法模式与抽象模式中的实体类型相同、关系词的词序相同，并且原始句法模式与某个抽象模式的相似度大于阈值 β(设置为 0.8)时，则抽取出对应的空间关系三元组。

此时已经抽取了部分的实例和相应的空间关系句法模式，由于这个过程是全自动的，我们将使用该过程取代人工标注空间关系词。然后进行特征统计，以提升空间关系词提取的准确性。根据从互联网中爬取的水利文本或其他水利业务公文，进行空间关系种子实例的抽取，对种子共现句进行句法模式表示，通过迭代的过程可以不断地获取到更多的空间关系描述词以及空间关系句法模式。当对外来的句子进行抽取时，就可以直接通过模式库中的抽象模式进行抽取。若句子

的原始句法模式与抽象库中的模式相匹配，则可以认为句中的水利实体对存在与该抽象模式对应的空间关系。

4 案例分析

4.1 种子集的构建

首先对水资源公报、水量调度执行情况公告等水利部调研文件内容进行预处理，然后针对选择的各项特征对共现句进行统计，结果如表1所列。

表1 特征的比例及均值

Table 1 Proportion and means of features

特征	比例及均值					
POS	名词	动词	方位词	其他	无关系	
	0.5996	0.3070	0.0217	0.0117	0.06	
LOC	左		中		右	
	0.1413		0.7935		0.0652	
LCCP (*CCP*)	左		中		右	
	0.0714		0.2143		0.7143	
DIS_1	0	1	2	3	4	>4
	0	0.5	0.1087	0.0870	0.0326	0.2717
DIS_2	0	1	2	3	4	>4
	0.4348	0.0761	0.1196	0.0652	0.2935	0.0109
LEN	名词		动词	方位词	其他	
	[2,3]		2	2	2	
$DIS(e_1,e_2)$	[1.67,13.25]					

通过表1的统计结果可以得到以下7个规律：

(1)语料中6%的句子不存在水利实体间的空间关系，且大部分的空间关系词为名词，余下的为动词和方位词；

(2)极少数空间关系词位于实体1左边或实体2右边，大多数都位于两个实体的中间；

(3)当空间关系词左边有介词或者连词时，有71.43%的空间关系词位于实体2的右边；

(4)半数的空间关系词到实体1的距离为1个词，即空间关系词与实体1最为接近；

(5)空间关系词远离实体2的比重最大，位于句尾的比例也最大；

(6)大部分空间关系词的长度为2个字，最多为3个字；

(7)当存在空间关系时，两个水利对象实体间的距离不超过13个词语。

然后，将统计的词性、位置、距离重要性引入关系词提取，从词语序列中选择一个重要度最高的词作为空间关系词rel。例如，对于句子“包头市南** 河段对应右岸为鄂尔多斯市达拉特旗大* * 河段”进行空间关系词提取。在进行分词、去除停用词、命名实体识别后，该句的水利实体对为〈南** 河段，大** 河段〉，对于剩余的词，计算每个词在词性、位置、距离的综合重要程度，结果如表2所列。

表2 上下文词的重要程度

Table 2 Importance of context words

上下文词语	重要度
包头市	0.36
对应	0.47
右岸	0.52
鄂尔多斯市	0.52
达拉特旗	0.52

由于统计结果中空间关系词的长度一般为2～3个字，因此最终选取“右岸”为〈南** 河段，大** 河段〉对应的空间关系词。

将抽到的结果与实体对(e_1,e_2)组成关系实例(e_1,rel,e_2)作为种子集，如表3所列。

表3 部分种子类型及关系

Table 3 Seed types and relations

关系类型	关系种子
下游关系	〈黄河，下游，小浪底〉
下游关系	〈黄河，下游，花园口〉
流经关系	〈黄河流域，流经，青海〉
流经关系	〈黄河流域，流经，甘肃〉
位于关系	〈* 化漏斗，位于，关中平原〉

空间关系词扩展利用外部词典《词林》进行同义词的扩展，构建空间关系同义词集合。

例如，种子关系中的空间关系词“流经”在《词林》中的原子词群是走过、流过、流经、几经、纵穿、横过、穿行、横穿、横贯，从而可以获得“流经”的扩展词。

4.2 原始句法模式的获取与泛化

原始句法模式是从处理共现句而得来的，因此我们需要先通过种子集找到语料中对应的共现句。例如：对于种子〈* 桥漏斗，位于，陕西省* * 县〉，获得共现句“位于陕西省* * 县的* 桥漏斗，由于上年末观测井抽水，因此不能反映漏斗中心埋深的年增减情况”，根据模式$P_i=\langle e_1,e_2,r,C,Pos_{e1},Pos_{e2},Pos_r\rangle$，$e_1$对应的水利实体为“* 桥”，$e_2$对应的水利实体为“陕西省** 县”，空间关系词$r$为“位于”。

使用自然语言处理工具对句子进行预处理，保留该词语序列中的动词、名词、形容词等有效词汇，过滤数词、代词等无意义词语，得到该词语序列：“位于 观测 井 抽水 漏斗 中心 埋 深 增减 情况”；计算序列中各个词语的权重，并对每个词标识出其在《同义词词林(扩展版)》中的语义代码：

位于：code=‘Jd02A02=’，weight=0.118

观测：code=‘Hj26A01=’，weight=0.082

井：code=‘Bn15B01=’，weight=0.107

抽水：code=‘Ih03C03=’，weight=0.082

漏斗：code=‘Bo08A06=’，weight=0.108

中心：code=‘Cb04C01=’，weight=0.108

埋：code=‘Fa11B01=’，weight=0.082

深：code=‘Ca09A01=’，weight=0.050

增减：code=‘Ih05A10@’，weight=0.077

情况：code=‘Da02A01=’，weight=0.103

根据词语权重计算公式，由于“位于”在句中词性、位置、距离上的重要性最高，且与实体的相对距离最近，则权重最大，因此“位于”为该模式的空间关系词r。

获取首要实体、次要实体和空间关系词的位置信息：

pos_1,pos_2,pos_r：2,1,0

原始句法模式如下：

e_1,e_2：GWHS(地下水水源地)，AD(行政区划)

r：位于

C：[位于：code=‘Jd02A02=’，weight=0.118

观测：code=‘Hj26A01=’，weight=0.082

井：code＝‘Bn15B01＝’，weight＝0.107

抽水：code＝‘Ih03C03＝’，weight＝0.082

漏斗：code＝‘Bo08A06＝’，weight＝0.108

中心：code＝‘Cb04C01＝’，weight＝0.108

埋：code＝‘Fa11B01＝’，weight＝0.082

深：code＝‘Ca09A01＝’，weight＝0.050

增减：code＝‘Ih05A10@’，weight＝0.077

情况：code＝‘Da02A01＝’，weight＝0.103]

Pos_{e1}，Pos_{e2}，Pos_r：2，1，0

尽管通过原始句法模式获取模块可以获得许多模式，但这些句法模式不够抽象，若直接用于关系提取则效率不高。因此，我们需要通过模式泛化方法将其抽象为更通用的模式。

例如，分别对共现句“位于陕西省＊＊县的＊桥漏斗，由于上年末观测井抽水，因此不能反映漏斗中心埋深的年增减情况”和“位于山西＊＊盆地的＊城漏斗，由于年末代表漏斗中心的监测井由位于储运公司监测井调整为＊城水文局院内监测井，上年末漏斗中心埋深有变”进行原始句法模式获取。后一句的原始句法模式如下：

e_1，e_2：GWHS(地下水水源地)，AD(行政区划)

r：位于

C：[位于：code＝‘Jd02A02＝’，weight＝0.109

代表：code＝‘Aj13B01＝’，weight＝0.010

漏斗：code＝‘Bo08A06＝’，weight＝0.099

中心：code＝‘Cb04C01＝’，weight＝0.099

监测：code＝‘Hd14C04＃’，weight＝0.076

井：code＝‘Bn15B01＝’，weight＝0.099

储运：code＝‘Hf05B12＃’，weight＝0.075

公司：code＝‘Dm04A01＝’，weight＝0.099

调整：code＝‘Hc03C01＝’，weight＝0.076

埋：code＝‘Fa11B01＝’，weight＝0.075

深：code＝‘Ca09A01＝’，weight＝0.050]

Pos_{e1}，Pos_{e2}，Pos_r：2，1，0

后一句与前一句的模式间的相似性可以通过模式相似度计算获得，为 0.926。并且关系词“位于”相同，即使不同，可以靠语义代码来辨别，例如“坐落”的语义代码与“位于”相同，都是“Jd02A02＝”，那么在模式相似度计算时，可以将这两个词看作同一个词语。由于以上两个模式的相似度大于阈值 0.8，则聚为一类，再对其泛化(泛化规则不再详述)，则泛化后的软模式如下：

e_1，e_2：GWHS(地下水水源地)，AD(行政区划)

r：位于

C：[位于：code＝‘Jd02A02＝’，weight＝0.114

井：code＝‘Bn15B01＝’，weight＝0.103

漏斗：code＝‘Bo08A06＝’，weight＝0.104

中心：code＝‘Cb04C01＝’，weight＝0.104

埋：code＝‘Fa11B01＝’，weight＝0.079

深：code＝‘Ca09A01＝’，weight＝0.050]

Pos_{e1}，Pos_{e2}，Pos_r：2，1，0

利用该模式可以抽取其他实体类型为地下水水源地和行政区划之间的“位于”关系。

4.3　空间关系的提取

泛化后的软模式可用于提取多个空间关系元组，本节将介绍如何使用获取的模式从水利领域语料中获得更多的关系元组。

对于“位于”关系，通过上述原始句法模式的获取和泛化方法，可以获得多个软模式。我们能够通过这些软模式抽取更多的“位于”关系，具体步骤如下：

(1)根据构建的同义词表不断从文本中检索含有该词的句子，例如，根据“位于”的同义词“坐落”检索得到句子：“坐落＊＊盆地(山西的＊原漏斗(承压水)漏斗中心地下水埋深减少＊m，其他地下水降落漏斗中心地下水埋深相对稳定”。

(2)利用 HanLP 工具对句子进行预处理，使用 Stanford CoreNLP 对共现句进行句法分析，计算词语序列中的词语权重以及标识语义代码。

(3)使用句法模式获取方法对共现句进行表示，获得该句子的原始句法模式。

(4)将该句法模式与软模式集中的模式一一匹配，如果相同词语超过半数并且句法模式间的相似性大于阈值 β，则认为该模式包括所提取的“位于”关系。

结束语　本文提出了包括空间关系种子集获取和基于句法模式表示的空间关系抽取两个阶段的空间关系词自动识别与提取框架。

首先，该方法通过基于 BootStrapping 的方法来统计样本中空间关系词在词性、位置、距离方面的特征，并将这些特征的重要性引入空间关系词提取中。其次，对种子集中的空间关系词进行扩展。最后，将种子集作为输入，对种子共现句进行预处理，获取空间关系原始句法模式、句法模式聚类及泛化，得到抽象程度高的软模式。使用同义词库中空间关系词的候选词进行检索，获取可能包含该关系的共现句，再次进行预处理，获取句法模式，并与模式库中的软模式进行比较，若符合相似条件，则进行相应的空间关系抽取。该方法通过半自动化手段，实现了从语料中获得大量空间语义的句法模式以及空间关系元组的目标，能够持续扩充水利空间关系词库，丰富了水利空间关系词体系。

然而，本文研究还存在一些不足之处，今后的工作包括以下几点：

首先，本文的数据不够丰富，因此下一步的工作是获取更多高质量的数据源，如查找其他水利对象的调研数据，对网页数据筛选出合适的数据源，从各类百科数据中的摘要部分获取更多空间信息。

其次，本文构建同义词库的方法不够智能，很多水利专用词汇并不在《词林》中，且许多同义词在《词林》中是同义，但在水利领域的空间层次上并不是同义，从而引入了一些噪声数据。因此后续的研究可利用机器学习的方法进行同义词库的构建。

最后，自然语言处理工具的错误也会给空间关系抽取带来误差，水利对象名称的识别需要自己构建词典来解决水利对象命名实体识别的问题。下一步的工作是利用水利领域的语料来训练实体识别的模型。

参 考 文 献

[1] CHENG J G, FENG J, YANG P, et al. Research on key technologies of water resources data directory service [J]. Water Resources Informationization, 2014(6): 18-21.

[2] FENG J, TANG Z X, ZHU Y L, et al. Study on metadata definition of water resources information catalog service [J]. Water Resources Informationization, 2011(S1): 19-22.

[3] ZHAO J, LIU K, ZHOU G Y, et al. Open Text Information Extraction [J]. Journal of Chinese Information Processing, 2011, 25(6): 98-110.

[4] LIU Y. Construction of Jilin Regional Knowledge Map Based on Geographic Ontology [D]. Beijing: Beijing Jiaotong University, 2017.

[5] HU C X, FU Y Q, ZHONG M Y. Extension of Semantic Query Based on Domain Ontology [J]. Journal of Computer Systems, 2012, 21(7): 83-89.

[6] JURAFSKY D, MARTINJ H. Speech and Language Processing [OL]. http://web.stanford.edu/~jurafsky/slp3/.

[7] SCHUTZ A, BUITELAAR P. RelExt: a tool for relation extraction from text in ontology extension[C]// International Conference on the Semantic Web. 2005.

[8] RINK B, HARABAGIU S. Utd: Classifying semantic relations by combining lexical and semantic resources[C]// Proceedings of the 5th International Workshop on Semantic Evaluation. 2010: 256-259.

[9] DODDINGTON G R, MITCHELL A, PRZYBOCKI M A, et al. The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation[C]// Language Resources and Evaluation. 2004.

[10] HUANG X, YOU H L, YU Y. A Summary of Research on Relationship Extraction Technology [J]. Modern Library and Information Technology, 2013, 29(11): 3039.

[11] XU F Y, USZKOREIT H, KRAUSE S, et al. Boosting Relation Extraction with Limited ClosedWorld Knowledge[C]// 23rd International Conference on Computational Linguistics (COLING 2010). Beijing: Association for Computational Linguistics, 2010.

[12] LI R J, ZHANG J, ZHANG X M, et al. Web information extraction in health field[J]. Journal of Computer Applications, 2016, 36(1): 163-170.

[13] ABNEYSP. Bootstrapping[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2002: 360-367.

[14] YU L, LU F, LIU X L. Bootstrapping method for extracting open geographic entity relations [J]. Journal of Surveying and Mapping, 2016, 45(5): 616-622.

[15] DENG M, XU R, LI Z L, et al. Research on the Transformation Method of Natural Language Spatial Relations and Metric Spatial Relations in Spatial Queries: Taking Area Targets as Examples [J]. Journal of Surveying and Mapping, 2009, 38(6): 527-531.

[16] MEI J J, ZHU Y M, GAO Y Q. Synonym Ci Lin (Second Edition) [M]. Shanghai: Shanghai Dictionary Publishing House, 1996.

[17] LI H G. Research on Chinese named entity relationship extraction based on location and semantic features [D]. Hefei: Hefei University of Technology, 2011.

[18] CHEN C. Research on Internet-based binary entity relation extraction [D]. Shanghai: East China Normal University, 2013.

[19] LU S, BAI S. Quantitative description of the effective range of word context in natural language processing [J]. Chinese Journal of Computers, 2001, 24(7): 742-747.

[20] BUNKYOKU H, MATSUO Y, ISHIZUKA M. Relation Extraction from Wikipedia Using Subtree Mining Dat P. T. Nguyen [C]// National Conference on Artificial Intelligence. 2013.

[21] SURDEANU M, TIBSHIRANI J, NALLAPATI R, et al. Multi-instance Multi-label Learning for Relation Extraction[C]// Joint Conference on Empirical Methods in Natural Language Processing & Computational Natural Language Learning. 2012.

[22] CHE W, LI Z, LIU T. LTP: A Chinese Language Technology Platform[C]// 23rd International Conference on Computational Linguistics, Demonstrations (COLING 2010). Beijing, China, 2010.

[23] KLEIN D. Accurate Unlexicalized Parsing[C]// Proceedings of the 41st Meeting of the Association for Computational Linguistics. Sapporo, Japan, 2003.

XIANG Ying, born in 1995, postgraduate, is a member of China Computer Federation. Her main research interests include relation extraction and so on.

FENG Jun, born in 1969, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include spatiotemporal data management, intelligent data processing, data mining and water conservancy informatization.