地理实体概念及其位置关系的获取和验证*)

姜 琳[1,2] 李 宇[1,2] 卢 汉[1] 曹存根[1]

(中国科学院计算技术研究所 中科院智能信息处理重点实验室 北京 100080)[1]

(中国科学院研究生院 北京 100080)[2]

摘 要 文本知识获取(Knowledge acquisition from text,简称KAT)是知识工程中的一个重要研究课题。重点研究如何从大规模Web网页文本中获取地理实体概念及其位置关系知识,本文首先介绍了如何自动和半自动地获取这些地理实体概念及其位置关系的文法模式,建立文法模式库;然后基于文法模式库获取例句来抽取候选概念并进行概念验证;最后利用基于图论的方法构造位置关系图,利用地理领域特定规则进行分析验证。作为统一概念图管理下概念空间的一个重要组成部分,地理实体概念及其位置关系本身不仅是知识库的一个重要部分,还可为知识库中其它领域的知识提供支持。

关键词 文本知识获取,地理实体概念获取,位置关系获取,知识验证

Acquiring Geographical Entities and their Relations from Texts

JIAGN Lin[1,2] LI Yu[1,2] LU Han[1] CAO Cun-Gen[1]

(Institute of Computing Technology, Chinese Academy of Sciences, Key Laboratory of Intelligent Information Processing, Beijing 100080)[1]

(Graduate School of the Chinese Academy of Sciences, Beijing 100080)[2]

Abstract There are various sorts of knowledge contained in text, and knowledge acquisition from text (KAT) has become one of the important research problems in knowledge engineering. How to acquire Geographical entities and their relations from large-scale Webpages is the research topic in this paper. Firstly, grammatical patterns are acquired automatically and semi-automatically to construct the Extraction Pattern Base. Then quasi-concept are extracted and verified locally with those resulting sentences generated from extraction patterns. At last, a Geographical Graph is defined based on graph theory, with the geographical rules to verify geographical entities and their relations. As an important part of the Conceptual Space managed by the Concept Graph, geographical entities and their relations are not only the important part of the Knowledge Base, but also provide support for other domain or domain independent knowledge.

Keywords Knowledge acquisition from text, Geographical concept acquisition, Geographical relation acquisition, Knowledge verification

1 前言

随着互联网的迅猛发展,庞大的网络信息为文本知识获取提出了迫切需求和新的挑战。文本知识获取是指将自然语言描述的文本知识转变为计算机可理解的形式,且具有查询推理的功能。由于自然语言处理的多歧义性和非规范性等,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[1]。

互联网上的文本中存在大量地理知识,这些知识虽然是领域限制的,但又是很普遍、常用的知识。并且作为概念空间和知识库的重要组成部分,地理知识本身除了可以用于自然语言处理系统(如机器翻译、文本理解)、信息抽取、信息检索等领域外,还可为其它领域知识提供基础性支持。

本文要获取的是地理实体概念以及它们之间的位置关系。地理实体概念包括:自然地理实体(如河流、山脉等)和人文地理实体(如机构、设施等)。除了命名实体(Named Entity, NE)识别中的中文、外文翻译的地名、机构名外,还包括与区域、方位等结合紧密的固定或半固定的词或短语,如中国东北、长江下游等。

NE的识别方法主要有两种:基于规则的方法和基于统计的方法。基于规则的方法对概念构成模式要求严格,会提高抽取结果的准确率,但却使查全率下降很多。基于统计的方法,由于它不考虑句法、语义上的信息,因此实现起来相对比较简单,并且这种方法不局限于某一专门领域,也不依赖任何外部资源,但是不可避免地对一些低频术语的获取和邻接高频词引入的噪声上存在一些问题[2]。

命名实体识别是整个信息抽取系统的基础[3]。近年来,受到很多会议的关注,如 MUC-6, MUC-7, Coling2002, Coling2003 等[4]。按照 MUC 会议的定义,命名实体识别任务主要是对文本中的专有名称,包括人名、地名、机构名以及时间表达式和数字表达式进行识别。

文本知识中关系获取的研究主要集中在上下位关系和部分-整体关系的获取上。Hearst [5]应用词汇-句法模式从文本语料中获取上下位关系。Pantel [6]自动获取词汇-词性模式,

*)本文工作得到自然科学基金的资助(#60273019、60573064、60573063 和 #60496326)和国家重点基础研究发展计划 2003CB317008 和 G1999032701 的资助)。**姜 琳** 硕士研究生,主要研究方向为关系获取、文本挖掘;**李 宇** 硕士研究生,主要研究方向为文本挖掘、古地名获取;**卢 汉** 博士研究生,主要研究方向为知识获取与文本挖掘;**曹存根** 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、智能教学。

以此为基础从T级别的文本语料中获取上下位关系。Cederberg[7]应用隐含语义分析和并列模式同时提高上下位关系获取的查准率和查全率。

关系获取的方法主要有两种：基于语句分析的方法和基于模式的方法[8]。理论上说，文本知识的获取需要在彻底理解的基础上进行。因此，句法分析和语义分析技术很自然地使用于这一领域，但它存在脆弱性和多歧义问题。基于模式的方法可以避免对语句进行彻底分析，效较高，但同一信息的不同表达形式会使模式数量大为膨胀[9,10]。

对Web语料进行模式匹配，获得包含地理实体位置关系知识的例句，即受限语料库；然后，从中提取出候选字符串集合，在通用领域进行概念获取，得到候选的概念集合，在此基础上，提出地理实体概念的验证算法，得到地理实体概念库；之后，再对受限语料库进行句法、结构以及语义的分析，结合地理实体概念库，得到候选的地理实体之间的位置关系；最后用构造地理实体位置关系图，验证位置关系知识，构建位置关系知识库。

2 地理实体及其位置关系获取的基本框架

本文研究的基本框架为(图1)：首先设计文法模式库，并

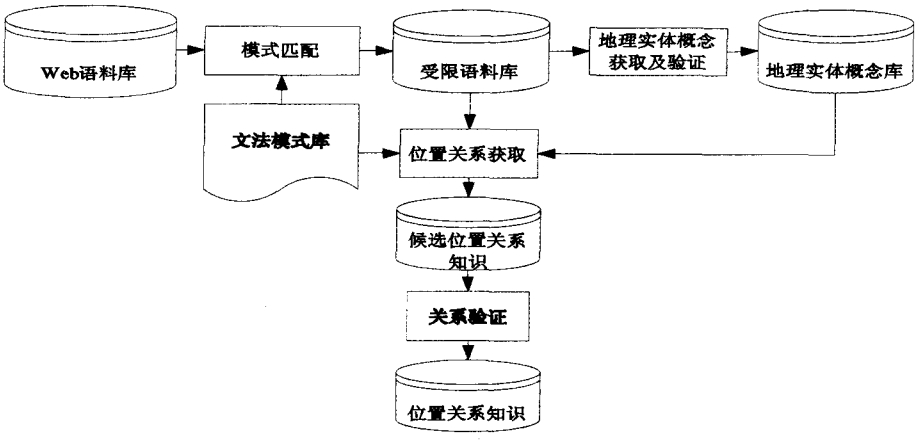


图1 地理实体及其位置关系获取的基本框架

3 文法模式匹配

3.1 句型的设计

由于我们面对的是不受限制的语料库，因此第一步就是先构造一个文法模式库。在对中文的处理中，通常称文法模式为句型。通过该模式库，对原始语料匹配，可以缩小语料库的范围，限定在某一领域内。

不过区别于一般句型的是，这里的句型除了遣词造句之外，还有逻辑表达式组成的规则加以限制，是用于从海量语料库中挑选出满足特定条件的规则的集合。只有设计出好的句型，才能正确和高效地获取知识。在图2中，给出了一个句型表示的例子。

其中"!"表示定义一个常量，"?"表示定义一个变量，"|"表示或者关系，"〈? C1〉"，"〈? C2〉"都表示任意字符串，基本规则是一组规则逻辑表达式，由一些函数以及函数之间的符号"∧"组成，每个函数代表一条规则，符号"∧"表示这些规则之间的逻辑"与"关系。Lengeq(〈? C1〉,4)表示句子中的〈? C1〉变量部分的长度要大于等于4个字节，即两个汉字。

```
defconstant 常数
{
      方位词: 东|西|南|北|东南|东北|西南|西北
      ……
}
defpattern 句型 001
{
      句型:  <?C1>位于<?C2>的<!方位词>部<?C3>
      基本规则: lengeq(<?C1>, 4)
}
……
```

图2

基本规则是由一些函数组成的逻辑表达式，目的是用来限制变量、关键词之间的关系，消除一些切分歧异，使得句型要表达的意思更为精准。

若一个句子S满足句型和基本规则，并且〈? C1〉中存在地理实体概念c1，〈? C2〉中存在地理实体概念c2，使得位置关系"位于(c1,c2)"成立，则s为含有位置关系的句子，位置关系记作：位于(c1,c2)。例如：

{{济南高新区}c1是国家级高新区，/ 位于/{济南市}c2/的东部/，分建成区和东部新建区两部分。}

{{昆士兰州}c1/ 位于/{澳洲大陆}/c2 的北部/。}

{{彭州} c1/ 位于/{成都市}c2/ 的西北部/，是兰花的故乡，号称"川兰第一市"。}

3.2 句型的扩充

如果句型不全，就必然丢失很多包含所需知识的句子，因此尽可能全面地找到合适的句型非常重要。最初寻找句型是通过浏览大量的自然语言文本实例(主要来源于网页、地理书籍等)，总结地理关系关键词，通过这种方法我们只能得到一部分句型。因此还需要进一步扩充句型，具体用了以下几种办法：

(1)通过Internet上的搜索引擎和句型匹配程序，用已知的地理实体名词去寻找关键词。例如在google输入"中国"，那么就可以将出现中国的网页搜寻出来。在句型匹配程序中输入句型：〈? C1〉天津〈? C2〉北京〈? C3〉，将会把出现天津和北京的句子都找出来，如：

航母入驻/天津/港后的许多日子里，港口附近不断光临/北京/、河北等地牌照的汽车。

我们就可发现新的句型：〈? C1〉〈入驻|光临〉〈? C2〉，〈?

C2〉中可能含有我们要获取的地理实体概念。

(1)通过一些同义词的补充和常识的扩充，如“飞往”可扩充为“飞往|飞至|飞抵”等。

(2)关键词本身的词性的扩充。除了动词、动词短语外，一部分名词、形容词也可作为关键词。例如：“锦州历史悠久，自然风光秀丽，名胜古迹众多。”“历史悠久”、“风光秀丽”经常和地理实体一起出现。

(3)有一类名词前往往会伴随出现机构名称或地名，比如“美国总统”、“北京市长”、“联想集团总裁”等。

通过以上的方法寻找出了大量的关键词，再将这些关键词写到句型中去。我们一共总结了170多个句型，最后经过试验整理归并，从中得到82个效果较好的句型。

模式匹配后，对符合位置关系句型的句子主要进行两步分析：(1)地理实体概念的获取和验证；(2)地理实体位置关系的获取和验证。

4 地理实体概念的获取验证

对于一个任意的候选字符串〈? C1〉，因为通用领域概念获取是无领域限制的获取概念，所以必然会引入大量非地理实体概念的干扰。例如，字符串“世界海拔最高的城市广场——布达拉宫广场”在通用领域概念获取后得到的结果为：“世界海拔”、“城市广场”、“布达拉宫广场”[2]。三个概念中，“布达拉宫广场”是一个地理实体，而其它两个都是干扰概念。另外，通用领域概念获取本身的查准率也会直接影响后面地理实体概念的验证。

如何从候选概念集合中识别出地理实体概念，是本文要解决的一个问题之一。领域概念往往含有本领域显著的构词特征规则，如后缀、前缀等。但是这些规则往往过于依赖于具体语言、文本格式，容易产生错误。而基于统计的方法利用统计特性的假设在实际语言现象中难以成立，且受限于训练语料库的规模。如何更好地结合这两种方法，扬长避短，成为概念提取的关键[11]。

鉴于这种情况，本文提出了一种将后缀、构词特征、标志位信息三者相结合的验证方法。

4.1 验证候选概念的后缀

相当一部分地理实体概念的尾部都是某个后缀词根，比如“北京市”、“阿尔卑斯山”“常州市安安家具厂”的尾部分别是“市”(行政区划单位后缀)、“山”(自然地理实体后缀)和“厂”(机构名后缀)这些后缀词词根。因此，总结出这些词根，利用地理名词的构词法，就能获取到一部分地理实体。

在设计词根表的时候，首先参照了国家标准GB/T13923－92《国土基础信息数据分类与代码》和《世界人文地理手册》[12]，再根据实际语料不断补充，得到了388个词根，基本上覆盖了所有的可能出现的词根。词根表是可扩充的，在以后的使用中可以进一步增加。

4.2 构词组合特征验证

还有一部分地理实体概念是一个复合概念，如：“黑龙江省中西部”、“云南省东南部”，这部分概念是由：(地理名词＋方位词)或者(地理名词＋方位词＋部位词)这样的组合模式出现，作为一个整体概念。

针对某种特定类型的概念做获取研究，根据边界获取概念，可以保证获取的查准性比较高。地理名词后缀和方位词都是识别地理实体概念右边界的非常重要的信息。

4.3 句型和通用领域概念获取后的标志位信息结合验证

通过上面两种方法，还有一部分地理实体无法识别出来，主要是一些外文译名。这些概念没有后缀，也不符合地理实体概念的构词组合特征，仅仅通过规则无法将它们获取出来，如日本、俄罗斯等。

这些概念出现的频率并不少，开始考虑用频率结合上下文信息获取这些词。即领域概念往往是在某个语境中经常出现，因此把相应的语境找出来。如果一个概念是在某个频率范围内在这个语境中经常出现，那么这个概念是属于这个领域的领域概念。这种方法首先要获得领域语境向量，然后将待验证的候选概念放入语料库中验证，代价较大。而且从实验的结果来看，虽然从一定程度上提升了整个结果的查全率，但是查确率却下降了。分析其原因，是验证的时候将一部分经常出现的、常用的普通概念也错误地获取出来了。

为了能以比较高的查确率获得这部地理实体概念，我们设计了一种结合句型和通用领域概念获取的标志位信息的地理实体概念验证方法。

首先给出句型查准率的定义：

$$\text{句型查准率}=\frac{\text{得到的含有所需要知识的句子数}}{\text{得到的句子总数}}\times 100\%$$

算法步骤如下：

Step1：计算句型查准率

句型的查准率有高有低。经过实验统计，有些句型查准率比较高，生成的文件中大部分句子都是有用的。

从定义可以看出，查准率较高的句型正是那些越可能含有地理实体概念的句型。

Step2：标记位信息选取

在通用领域概念获取时，输入一个字符串。概念获取后，我们可以得到四种类型的标记信息，分别是a：全词，即这个字符串本身就是一个概念；p：部分词，字符串中的一部分是一个概念；m：多词，从这个字符串中抽取出多个概念；n：无，这个字符串中没有概念。从实验结果分析，标记为“a”的那部分查准率较高[9]。

Step3：句型查准率和标记位信息相结合判断

在验证一个概念是否是地理实体概念的时候，我们可以考察它出自哪个句型以及在通用领域概念获取时的信息位情况来判定它是一个地理实体概念的概率。通过实验，我们发现，当一个概念的句型属于较高查准率句型，并且它的信息位标记为“a”的时候，它是一个地理实体概念的概率较大。

实验情况：

随机抽取200个概念，将后缀和构词特征作为一次验证，统计查准率和查全率。在一次验证的基础上，加入标志位信息做二次验证，再统计查准率和查全率。

两次验证、实验数据如表1。

表1

	查准率	查全率
一次验证	92.3%(84/91)	84%(84/100)
二次验证	93%(93/100)	93%(93/100)

200个候选概念，其中实际含有的地理实体概念为100个。一次验证后，得到91个地理实体概念，其中84个是正确的。二次验证后，得到100个地理实体概念，其中93个是正确的。通过二次验证，查准率由原来的92.3%提高到93%，查全率由原来的84%提高到93%。

5 位置关系的获取和验证

我们所获得的知识是从大量的语料库中提取出来的，而语料库的选取是通过搜集大量的网页。网页上的内容是否全部是正确的？显然是不可能的。再考虑到汉语的表达的复杂性，模式匹配的结果并不是完全精确的，所以我们必须对获得的知识进行验证。

知识验证，即对获得的待验证断言进行分析，检查其可能存在的异常现象，如知识的不一致性、冗余性、非完备性等，以确保构成正确完备的知识库[13]。

5.1 位置关系知识谓词形式的获取

候选位置关系谓词获取的过程如下。

Step 1：对 Web 语料库预处理和模式匹配，得到有标记的受限语料；

Step 2：在受限语料的基础上对每个例句提取谓词，转换成谓词三元组 $R(c1,c2)$ 的形式；

Step 3：对 $R(c1,c2)$ 中的字符串分别进行地理实体概念的提取、验证；

Step 4：谓词形式化简。

例如，语料中有这样一句："…我们下榻的斯堪的纳维亚饭店位于哥本哈根的西南部…"，位置关系获取的过程如图 3。

Step1：我们下榻的斯堪的纳维亚饭店/位于/哥本哈根/的西南部/
Step2：位于西南部([1]我们下榻的斯堪的纳维亚饭店/ /[1]哥本哈根)
Step3：位于西南部(纳维亚饭店/ /哥本哈根)
Step4：位于_部位(纳维亚饭店；哥本哈根；西南)

图 3 位置关系获取的过程

5.2 基于位置关系图(G 图)的知识验证

5.2.1 G 图的定义

在知识库的建设中，知识分析是非常重要的一环，如知识的一致性和完整性分析就是两种最基本的知识分析。为了验证获取的待验证地理位置关系知识，我们首先定义了一个带标记、带权的有向图——概念图。

定义 概念图是一个有向图 $G=(\sum,C,P,E,Q,A,\alpha)$，其中

(1) $\sum=\{c_i\}$ 为概念集合；

(2) CN 为概念节点集合，CN 递归定义如下：

① $\sum\subseteq CN$

② $2^{CN}\subseteq CN$

③ $CN^*\subseteq CN$

④没有其他元素属于 CN

(3) $P=\{pr_i\}$ 为概念关系模式集合，$P^*=\{r_{ij}\mid r_{ij}$ 为 pr_i 所产生的关系实例$\}$；

(4) $E\subseteq CN\times P^*\times CN$ 为边集合，其中 E 中的元素也记为 $r_i(cn,cn')$；

(5) $Q=\{qr_i\}$ 为关系约束模式集合：$qr_i:E\to CN$；

(6) A 为 $\sum$ 的属性集合，$A=\{a_i\}$；

(7) $\alpha:\sum\to 2^A$，$\alpha(c_i)$ 为概念 c_i 的属性集合。

包含地理实体概念及其它们之间位置关系的图我们称为地理位置关系图，简称 G 图。G 图中一个节点 c_i 表示一个地理实体概念，r_i 为一条有向边，表示它所连接的两个实体之间的位置关系。

目前，概念图有以下几种类型的关系模式：上下位关系模式[14]、同指关系模式[15]、整体部分关系模式、地理位置关系模式。统一的概念图可以整合这些看似相互孤立的知识，构建统一的概念空间。各领域的知识可以相互验证，相互提供支持。

5.2.2 G 图上的验证

利用已获得的位置关系构造 G 图的元图(即没有验证的初始图)，需要验证以下两个问题：

(1)冗余性

冗余性包括等同冗余和等价冗余。

等同冗余：若新加入关系对 $r_i(c1,c2)$，图中存在关系对 $r_j(c3,c4)$，并且 $r_i=r_j$，$c1=c3$ $c2=c4$，即要放入图中的新知识是图中现有的某条知识的重复出现。

等价冗余：$r_i=r_j$，$c1$ 与 $c3$、$c2$ 与 $c4$ 相同或同义(同指)。

如："位于东南沿海(晋江市，福建省)"与"位于东南沿海(晋江，福建)"，两条知识的谓词关系相同，都是"位于东南沿海"，其中"晋江市"与"晋江"同指，"福建省"与"福建"同指，两条知识是等价冗余。

(2)不一致性问题

不一致性是指新知识与现有知识产生冲突。即新加入关系对 $r_i(c1,c2)$，图中存在关系对 $r_j(c1',c2')$，使得 r_i 与 r_j 不相容。

如"北邻(科威特，伊拉克)"与"南临(科威特，伊拉克)"冲突，是两条矛盾的知识。

(3) G 图上的验证算法

输入：已经获取的候选位置关系知识集合

输出：部分位置关系知识以及每条位置关系知识的置信度 θ_i(从候选集合中删除已经被验证为错误的知识)

Step1：每条知识 $r_i(c1,c2)$ 有一个 θ_i，初始值为 I；

Step2：1)if $r_i(c1,c2)$ 中 $c1==c2$，θ_i--；

2)if strlen($c1$)>50 或者 strlen($c2$)>50，confi−−；

3)if $r_i(c1,c2)$ 中 $c1$ 出度 $OD>4$，入度 $ID=0$，且 $c1$ 无相似概念(与同指词根组合无匹配)，θ_i-；

Step3：查找重复出现的知识，更新置信度，$\theta+n*0.5+k*0.5$

其中，N 代表相同谓词的个数，k 代表不同谓词表示但是谓词含义相同的谓词的个数特殊情况说明：if $n>3$，且 $k=0$，θ_i--；

Step4：读入一条知识 $r(c,c1)$，检查已有知识库中 $r(c,x)$ 的情况(x 为与 $c1$ 不同的概念)

1)若没有这样的知识，执行 step5；

2)若有这样的知识 $r(c,c2)$ …… $r(c,cn)$(如图 4(1)所示)，执行 step7；

3)执行 step5；

Step5：检查知识库中 $r(x,c1)$ 的情况(x 为与 c 不同的概念)

1)若没有这样的知识，，执行 step6；

2)若有这样的知识 $r(c2,c1)$ …… $r(cn,c1)$(如图 4(2)所示)，执行 step7；

3)执行 step6；

Step6：检查知识库中 $r'(c,c1)/r'(c1,c)$ 的情况(r' 为与 r 不同意义的关系)

1)若没有这样的知识，结束；

2)若有这样的知识 $r1(c,c1),\cdots.rn(c,c1)$(如图 4(3)所示),执行 step8;

3)执行 step9;

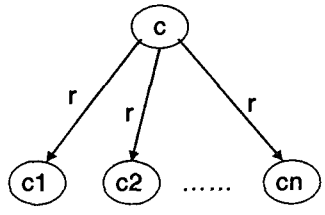
4)结束。

Step7:判断多个概念的相似度(两两判断)

前缀:概念 $c1$ 是概念 $c2$ 的子集,并且是 $c2$ 的前缀,两条知识 θ 分别+0.5;

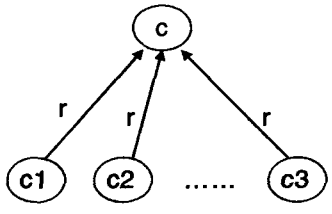
后缀:概念 $c1$ 是概念 $c2$ 的子集,并且是 $c2$ 的后缀,两条知识 θ 分别+0.5;

交集部分:概念 $c1$ 和概念 $c2$ 有交集,并且交集部分的长度>=4(连续两个汉字),两条知识的 θ 分别+0.5;



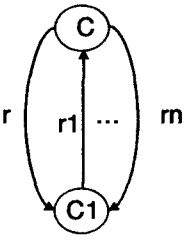
r(c,c1),r(c,c2) ..... r(c,cn)

(1)



r(c1,c),r(c2,c) ..... r(cn,c)

(2)



r(c,c1),r1(c,c1),...rn(c,c1)

(3)

图 4

Step8:判断 $r,r1,\cdots,rn$ 的关系

1)若存在矛盾的地方,θ_i − −;

如:东(c,c1)〈−〉西(c,c1),东(c,c1)〈−〉东(c1, c),接壤(c,c1)〈−〉位于内部(c,c1)

2)若存在相容的地方,θ_i+0.5;

如:东(c,c1)〈−〉东南(c,c1)

3)若存在关系等价的地方,θi ++;

如:东邻(c,c1)〈−〉东临(c,c1),东临(c,c1)〈−〉西临(c1, c)

Step9:若存在 $r(c',c1')$,c 和 c' 同指、$c1$ 和 $c1'$ 同指,两条知识的置信度分别 θ ++。

通过上述方法验证后,每条知识的置信度 θ 发生变化,设置置信度阈值,小于阈值的关系是错误的知识,置信度 θ 越小,错误可能性越大;大于等于阈值的关系对是正确的知识,θ 越大,正确可能性越大。

5.3 试验分析

5.3.1 试验数据

从定义的关系模式中抽取出 18 个位置关系模式,以 2.6G 的中文开放语料(经过处理的网页)为测试语料获取位置关系。

基本过程是先模式匹配提取候选例句,然后对候选例句进行谓词抽取和概念过滤得到候选位置关系,再对这些候选关系进行分析验证,最终得到位置关系集合。

5.3.2 试验结果分析

通过地理位置关系获取得到的 2167 个候选位置关系,用地理实体概念库初步验证过滤后,得到关系对 894 对,查准率为 84.22%。在此基础上,构建 G 图。

在 G 图上,孤立的关系对有 571 对,这部分知识中,一部分可以通过非自反、字符串长度、实体同指等方法验证,另一部分由于语料的限制,在现有的 G 图内还没有足够的证据验证,今后可采用更大语料库和增加句型来解决。

除了关系对的出现频率影响外,主要是以下几种情况:

(1)实体同指关系识别,如图 5。

图中,“上海”与“上海市”指的是同一个地理实体,而对于关系对“位于腹地(新疆,亚欧)”与“位于腹地(新疆,亚欧大陆)”,“亚欧”与“亚欧大陆”是同一个实体,这是由于在自然语言描述中,一个地理实体常常有多种表述方式。

(2)描述精度问题,如图 6。

图中,“位于北部(杨浦区,上海市)”与“位于北部(杨浦区,上海市中心城区)”都是正确的,但是后者描述更严格、更精确。

以上两种情况,我们通过同指概念验证和概念相似度这两种方法来解决。

在第一种情况中,概念 $c1$ 通过同指后缀词匹配后,使得 $c1=c2$,那么可以认为 $c1$ 和 $c2$ 是同一个概念。

第二种情况,若有 $r1(c1,c2)$,$r2(c3,c4)$,且 $r1=r2,c1=c3$,而 $c2\approx c4$,即 $c2$ 和 $c4$ 相似,则 $r1(c1,c2)$和 $r2(c3,c4)$是等价的。

(3)入度为 0、出度较大的节点所代表的概念往往不是地理实体。而入度较大或出入度都较大的节点,往往代表一个比较有名、覆盖面积(或流域)比较大的地理实体。如长江、中国。表 2 中列出了一些概念的出度、入度的统计情况:

表 2

概念	出度	入度	出入度和
简况	10	0	10
位置	4	0	4
项目	4	0	4
新华网北京	6	0	6
震中	4	0	4
阿富汗	3	0	3
埃及	2	0	2
太行山	0	5	5
澳洲大陆	0	2	2
长江	4	33	37
中国	11	15	26
太平洋	2	8	10

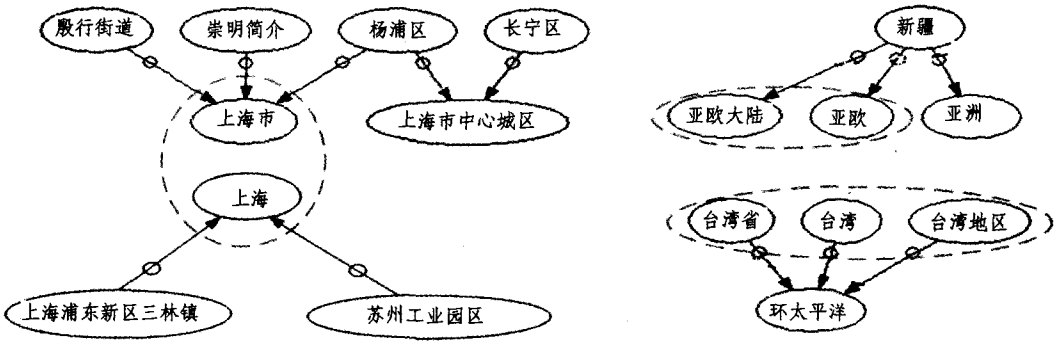


图 5

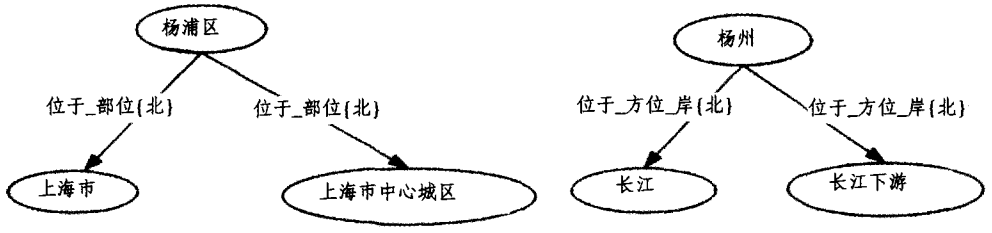


图 6

经过统计分析，当入度为 0、出度大于 4 的时候，这些概念基本上都是一个错误概念或者是非地理实体概念。考虑到现在的语料库有限，对此验证策略做了改进：一个概念入度为 0、出度大于 3 的时候，不以地理后缀结尾，并且与同指词根组合后在概念库中无匹配，则此概念为错误概念。图 7 给出了概念图实例中的一部分，“简况”、“位置”都不是地理实体概念。

(4)位置关系中不满足自反关系，因此 $r(c1, c2)$ 中，$c1 \neq c2$。$c1=c2$ 的为错误关系，如：位于北岸(罗布泊，罗布泊)位于西部(重庆，重庆)东连(长江，长江)。

通过 G 图验证后，我们得到正确的知识 851 对，其中正确的 756 对，正确率为 88.84%。

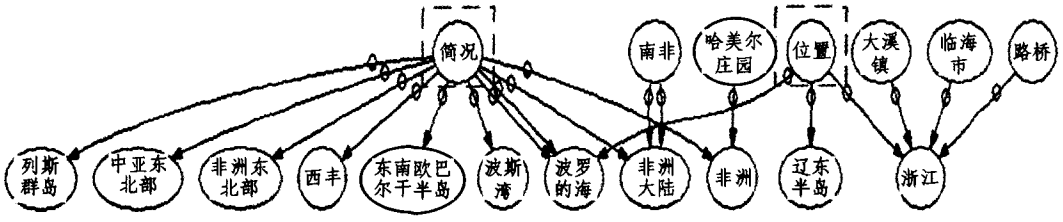


图 7

图 8

6 存在的问题与进一步工作

本文讨论了如何从大规模中文语料中获取地理实体概念及其位置关系，包含三个基本阶段：首先句型设计和模式匹配过滤语料库，从自由文本语料中获得可能包含地理实体位置关系知识的例句；然后在从这些例句中获取和验证地理实体概念；最后，按句法结构从抽取候选位置关系，建立地理位置关系图(G 图)，在图上验证候选的位置关系，最后位置关系知识库。

实验证明，这种从待获取的目标词及关系标志词的特征入手，获取目标概念和关系的方法得到的结果令人满意，这种方法具有一定的通用性，还可以用于其它类型概念及关系的获取。在这个过程中，还存在一些需要解决的问题，例如模式匹配中的切分歧义处理，关系验证中证据不足问题。下一步我们首先用更大的语料库和更多的模式解决孤立点边问题，

(下转第 174 页)

利用上节给定的算法1进行约简和算法2来获取强特征，得到如下规则：

规则1 （a1，晴）∧（a3，高）→（d，N）。

规则2 （a1，多云）→（d，P）。

规则3 （a1，雨）∧（a4，否）→（d，P）。

规则4 （a1，雨）∧（a4，真）→（d，N）。

规则5 （a1，晴）∧（a3，正常）→（d，P）。

对于本文给出的CARMA与C4.5方法得到的性能对比如图3所示。可见，用CARMA进行分类的效率高。硬件平台为P4 2.8G Hz. 256M. 在SQL SERVER2000上实现。

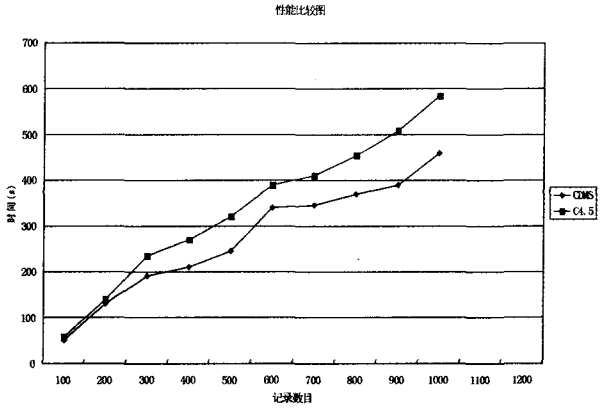


图3 CARMA与C4.5性能比较图

本文的属性约简算法1的空间复杂度是$O(|C|)$，而传统的算法用分辨矩阵计算约减属性的核，空间复杂度是$O(|C|^2)$，由此可见此算法大大降低了空间复杂度。

分类精确度定义为能够正确分类的记录数目和测试集中记录总数目的比值。实验对同样的数据集使用基于正区域的属性约简算法[5]、基于区分矩阵的属性约简算法[6]、基于信息熵的属性约简算法[7]和基于本文的属性约简算法，它们的分类精度实验结果分别是83.4、77.5、84.2、88.6。可见，用本文中的属性约简后分类的精确度比用其他几种属性约简方法更高。

结论 为了有效地、更快地对海量数据进行分类，本文提出了将属性约简和分类关联规则挖掘相结合的分类挖掘系统的算法。它运用粗糙集理论把关系数据库按属性值分成若干等价类、约简冗余属性及依赖属性，然后对数据约简后的目标关系表求取分类支持度大于阈值的强类和特征置信度大于阈值的强特征，从而有效获取强类中的强特征的决策规则。实验结果表明，CARMA对于数据的分类是有效的，目前比文中提到的其它方法具有更高的分类精度和效率。它能够有效地克服ID3系列算法的冗余性、复杂性和对大数据量的不适应性，而且在原始数据增加的情况下，可以通过约简来压缩数据规模，使之只与属性值有关系，而与原始的数据量无关，因此对增量数据能够达到较好的分类效果。算法的执行和处理是简单易行的，产生的规则也是准确的。领域决策者就可以直接运行分类器进一步对同类问题进行分类和决策。在各种评估工作、医疗诊断、交通信息、案件信息、天气预测等大型数据库的数据挖掘中有广泛的应用前景及实用价值。

参考文献

1 Han Jiawei, Kamber M. Data Mining: Concepts and Techniques. Simon Fraser University, 2000

2 Pawlak Z. Rough sets [J]. International Journal of Computer and Information Science, 1982, 11: 341～356

3 Mitra P, Murthy C A, Pal S K. Unsupervised feature selection using feature similarity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301～312

4 张修平，仇国芳. 基于粗糙集的不确定决策[M]. 北京：清华大学出版社，2005. 28～37

5 Guan J W, Bell D A. Rough computational method for information systems [J]. Artificial Intelligence, 1998, 105(1-2): 77～103

6 Skowron A, Rauszer C. The discenibility matrices and fuction in information system [C]. In: Slowinski R, ed. Intelligent Decision Support Handbook of Application and Advance of The Rough Sets Theory, Dordrecht: Kluwer Academic Publishers, 1992. 331～362

7 苗夺谦，胡桂荣. 知识约简的一种启发式算法[J]. 计算机应用与发展，1999，36：681～684

8 Agrawal R, Imielinski T, A Swami. Mining association rules between sets of items in large database [C]. In: Proceeding of the ACM SOGMOD Conference on Management of data, 1993, 5: 207～216

9 Agrawal R, Srikant R. Fast algorithms for mining association rules [C]. In: Proceedings of the 20th VlDB Conference, Santiago, Chile, 1994. 487～499

10 Relue R, Wu Xindong, Huang Hao. Efficient Runtime Generation of Association Rules. In: Proceeding of 2001 ACM CIKM Tenth International on Information Knowledge Management, 2001

11 徐余法. 粗糙集理论及应用. 上海电机学院学报，2005，8(2)：39～43

12 蒋良孝，蔡之华，刘钊. 一种基于粗糙集的决策规则挖掘算法. 微机与应用，2004，3：7～9

13 瞿彬彬，卢炎生. 基于粗糙集的属性约简算法研究. 华中科技大学学报(自然科学版)，2005，33(8)：30～33

14 洪家荣. 归纳学习——算法理论和应用. 北京. 科学出版社，1997

15 潘巍，王阳生，杨宏戟. 粗糙集理论中求取最小决策规则的研究. 计算机科学，2007(5)

16 宋笑雪，解争龙，张文修. 集值决策信息系统的知识约简与规则提取. 计算机科学，2007(4)

17 孙成敏，刘大有，孙舒杨. 含序信息的粗集方法研究. 计算机科学，2006(11)

(上接第156页)

引入关系规则推导挖掘隐含的位置关系，并可利用其它关系知识交互验证位置关系。

参考文献

1 曹存根，张春霞，王海涛. 基于本体的文本知识获取研究[A]. 见：王珏，陆汝钤，等. 智能信息处理系列研讨会[C]. 上海：2003. 7～8

2 余蕾. 从大规模中文语料中获取和验证概念的研究[D]. 北京：中国科学院计算技术研究所，2006

3 于海滨，秦兵，刘挺，等. 命名实体识别和指代消解在文摘系统中的应用. 计算机应用研究，2006，23(4)：180～182，195

4 Guo Honglei, Jiang Jianmin, Hu Gang, et al. Chinese Named Entity Recognition Based on Multilevel Linguistic Features. In: IJCNLP-04, Hailan, China, March 2004. 294～231

5 Hearst M A. Automatic Acquisition of Hyponyms from Large Text Corpora. In: 14th International Conference on Computational Linguistics (COLING 1992), August 1992. 539～545

6 Pantel P, Ravichandran D, Hovy E. Towards Terascale Knowledge Acquisition. In: Proceedings of Conference on Computational Linguistics. (COLING-04), Geneva, Switzerland, 2004. 771～777

7 Cederberg S, Widdows D. Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In: Conference on Natural Language Learning (CoNLL-2003), Edmonton, Canada, 2003. 111～118

8 Hull R, Gomez F. Automatic acquisition of biographic knowledge from encyclopedic texts [J]. Expert Systems with Applications, 1999, 16: 261～270

9 宋柔，许勇. 基于词汇语义的百科辞典知识提取实验[J]. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 37～40

10 宋柔. 汉语词汇语义信息的研究和应用[A]. 见：第三届中文词汇语意学研讨会论文集[C]. 台北：2002. 169～177

11 张春霞. 领域文本知识获取方法研究及其在考古领域中的应用[D]. 北京：中国科学院计算技术研究所，2005

12 毛汉英，刘伉. 世界人文地理手册. 北京：知识出版社，1983

13 顾芳. 多学科领域本体的设计方法研究[D]. 北京：中国科学院计算技术研究所，2004

14 刘磊，曹存根，王海涛，等. 一种基于“是一个”模式的下位概念获取方法. 计算机科学，2006，33(9)：146～151

15 Tian Guogang, Cao Cungen, Liu Lei, et al. MFC: A Method of Co-referent Relation Acquistion from Large-scale Chinese Corpora [C]. In: Fuzzy Systems and Knowledge Discovery(FSKD 2006), Xi'an, China, 2006. 1259～1268