文章编号: 1671-8860(2006)12-1101-04 文献标志码: A

Web GIS中文查询语句的词义理解算法研究

徐爱萍[1,2] 边馥苓[2] 曹 杰[1]

(1 武汉大学计算机学院, 武汉市珞喻路129号, 430079)
(2 武汉大学空间信息与数字工程研究中心, 武汉市珞喻路129号, 430079)

摘 要: 针对Web GIS中文查询语句中词义的多义性给查询语句的正确理解带来的困难, 在系统语料库的基础上设计了能解决该问题的算法, 提出了句子链栈的存储结构。测试结果证明了该算法能达到95%的正确率, 为语义分析子系统能正确地提取查询目标、查询条件、分组信息和排序信息提供了有效的词义信息。

关键词: 语料库; 多义性; 句子链栈; 词义链; 表链
中图法分类号: P208

一个完整的中文查询语句理解一般包括切分词、文法分析、语义分析和SQL语句生成等子系统。切分词子系统包括基于语用环境语料库的建立、词的切分、切分词义处理3个阶段。切分词的算法要能过滤查询句中的未登录词, 并将未登录词通过人机交互加入到语料库中; 切分词的语义处理是为了处理词的多义性, 为语义分析子系统提供确定的词义信息和一个用词类符号表示的句子字符串[1,2]。文法分析子系统包括文法规则的制定、文法规则库的管理。语义分析子系统根据文法规则从句子字符串中抽取GIS中文查询语句的查询目标、查询条件、分组信息和排序信息, 形成中间语言。SQL语句生成子系统依据中间语言到SQL语句的格式转换规则, 把中间语言转换为当前GIS平台能够执行的SQL语句。本文将对切分子系统中切分词的词义处理进行研究。

1 语料库

在中文查询语句理解中, 语料库是中文分词、语义理解的基础, 为便于实现通用、可靠的分词系统, 把要提取的词条分为通用词和专用词, 语料库的数据结构如下:

```
Struct corps
{
  Char * Word; // 有一定语义的词
  Char Wordtype; // 词类
  Char * Describe; // 词对应的语义信息
}
```

属于领域无关词类的词存储在系统的通用词库中[3], 在系统移植时这些词一般不需要修改。设*A*-限定词、*B*-连词、*G*-动词、*D*-疑问词、*Z*-助词、*O*-关系词、*L*-逻辑词、*F*-函数词、*N*-数量词、*W*-单位词、*S*-空间关系、*G*-分组词、*I*-排序词、*E*-表名、*P*-属性、*V*-属性值、*H*-代词。

专用语料库是空间数据库查询系统所特有的语料库, 其语料库的设计与应用领域紧密联系。设在"国家-城市-河流"GIS中有Country.TAB、City.TAB和City.TAB3个表, 在Country.TAB中有Name、Continent、Pop和GPD属性; 在City.TAB中有Name、Country、Pop和Cptial属性; 在River.TAB中有Name、Origin和Length属性, 每个表都有相应的空间对象, 则专用语料库示例如表1所示。

由此可见, 语料库中词的多义性严重, 一个完整的地理信息系统语料库更为复杂。词的多义性将直接影响查询语句的理解, 本文将针对这一问题进行研究。

收稿日期: 2006-08-28。
项目来源: 武汉市青年科技晨光计划资助项目(20045006071-16); 国家自然科学基金资助项目(60573183)。

表 **1** 部分专用语料库

Tab. 1 Part of Special Corpus

Word	Wordtype	Describe
国家	*E*	Country
国家	*P*	Country. Name
国家	*P*	City. Country
城市	*E*	City
城市	*P*	City. Name
河流	*E*	River
河流	*P*	River. Name
名称	*P*	Country. Name
名称	*P*	City. Name
名称	*P*	River. Name
人口	*P*	Country. Pop
人口	*P*	City. Pop
中国	*V*	Country. Name= '中国'
中国	*V*	City. Country= '中国'
中国	*V*	River. Origin= '中国'

2 切分语义处理算法

切分词是根据系统语料库, 找出切分所得各词的词类和形式描述等信息。由于词的多义性, 词义理解算法的设计思想如下。

1) 将切分到的词加入到句子链栈, 新建一个词义链结点。

① 当切分到的词只有一个词义时, 如果其词类是 *P* 或 *V*, 则取出其前面的表名, 如果表名不在表链中则加入到表链。

② 当切分到的词有多个词义时, 如果其词类是 *E* 且其表名不在表链中则要加入到表链, 其他词类为 *P* 或 *V* 的词, 如果其词义所带的表名在表链中, 则将其加入到词义链, 否则, 将其过滤掉; 如果其词类全部是属性 *P* 或属性值 *V*, 并且其词义所带的表名不在表链, 则将不能确定其明确的语义, 在第二次扫描时再确定。

2) 第一遍扫描后表链中就有了表名, 通过查找表链就可以来确定第一遍扫描时不能确定的词义, 将其词义所带的表名在表链中的词义留在词义链, 其他的词义过滤掉。

2.1 在分词结果中提取语料库中词的词义

分词后通过以下算法提取语料库中较确定的词义信息: 算法的输入是分词后得到的多个词义信息; 算法的输出是句子字符串、表链和无多义性的句子链栈和词义链。

由于查询语句的目标一般在句末, 所以句子语义信息的存储结构需要有后进先出的特点, 一

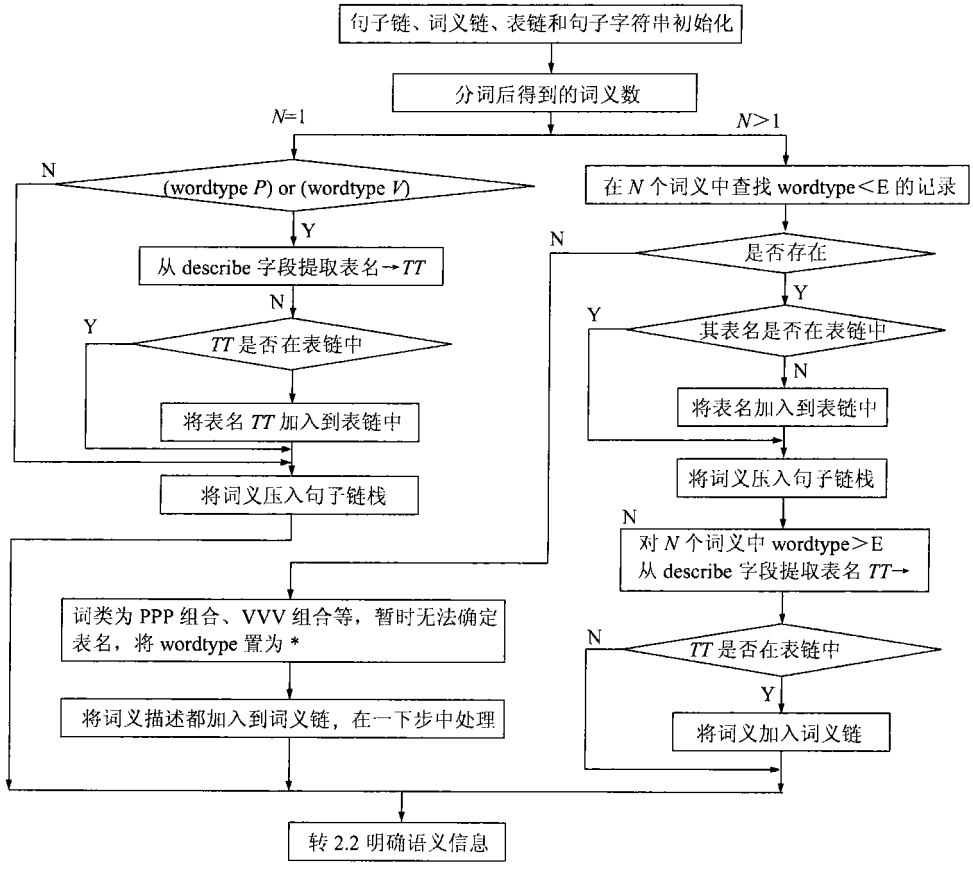

图 1 解决词的多义性算法流程

Fig. 1 Algorithmic to Resolve Word Meaning Ambiguity

般采用栈的存储结构; 而查询语句的长度和每个词的词义均不为定值, 因此, 在此建立了两级数据链栈结构。

第一级链为句子链栈, 由 LiStack 类型的结点组成, 结构如下:

```
typedef struct Word { char * word_name;
char wordtype;
struct Word * nextword;
struct Mean * nextmean;
} LiStack;
```

第二级链称之为词义链, 由 Mean 类型的结点组成, 结构如下:

```
typedef struct Mean {
char * discribe;
struct Mean * next;
}LiMean;
```

表链用来存放算法分析中出现的表名, 结构如下:

```
typedef struct Table{
char * name;
struct Table * next;
}LiTable;
```

解决词的多义性算法流程描述如图 1 所示。

2.2　明确语义信息

前面不能确定的信息在句子栈链的 LiStack-> wordtype 设置了标记' * ', 下面对此词义进行处理。其算法描述流程如图 2 所示。

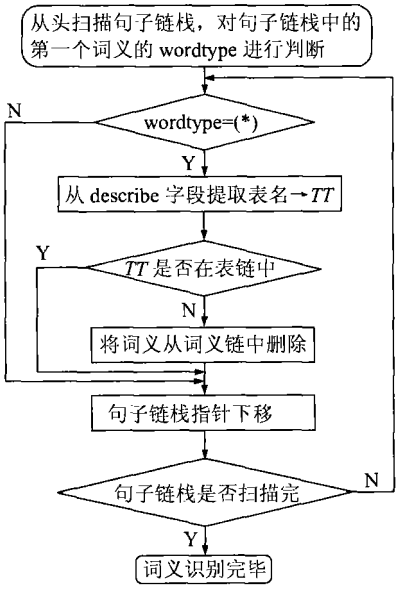

图 2　明确词义信息的算法描述流程

Fig. 2　Algorithmic to Identify Word Meaning

3　算法的性能分析

算法的可用性取决于算法的正确率, 但其性能的好坏取决于时间、空间的复杂度。由于空间复杂度是本算法的特点, 在此通过实验对算法的正确率和空间复杂度进行测试分析。算法的测试环境是上面给出的"国家-城市-河流"地理信息系统, 本文对 40 条查询语句进行了测试。

3.1　算法的正确率

该算法的正确率与系统语料库、分词算法、未登录词的处理有关, 系统语料库建立的原则在前面已讨论过, 但要保证切分词的惟一性和完备性是一项复杂而细致的工作, 所以需要通过学习来不断完善语料库; 而分词算法要求能处理未登录词(该算法将另文介绍)。由于是测试系统, 建立了一个完善的系统语料库, 设计的分词算法也能处理未登录词, 在此前提下对此算法进行了正确性测试。如图 3 所示是查询语句"哪些河流流经湖北"。经过以上的多义性处理后可得到句子栈链、词义链和表链。

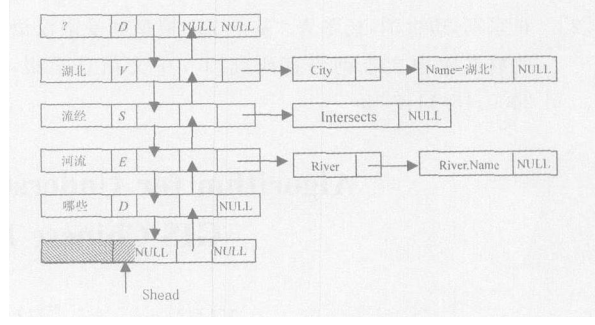

图 3　句子栈链、词义链和表链

Fig. 3　Sentence Link Stack, Word Meaning Link and Table Link

经过反复调试, 在上面设定的条件下其正确率可以达到 95%, 少数不能正确分析的原因主要在于词义不是独立的, 有时要靠修饰关系来决定, 而修饰关系是在语义处理阶段来分析的。比如"列出格兰德河发源地国家的首都和人口", 这里"人口"是指国家人口, 但"列出格兰德河发源地国家的首都和城市人口", 这里"人口"是指首都人口, 所以需要在语义分析阶段作进一步判断。

3.2　算法的空间复杂度

该算法的特点是采用了链栈式存储结构, 相对采用栈结构[2,4] 的存储形式, 在空间分配上有更大的灵活性。设栈结构方式中申请的句子栈的大小为 n 个单元, 每个词的词义栈的大小为 m 个

单元, 表栈的大小为 p 个单元; 查询语句的分词数为 x, 每个词的最多词义数为 y, 查询语句涉及到的表的数目为 z。栈方式与链栈方式的比较结果如表 2 所示。

表 2 固定栈与链栈方式的比较

Tab. 2 Comparison of Way of Fixed Stack and Way of Link Stack

比较项目	固定栈方式	链栈方式
先进后出的特点	有	有
栈的大小	n、m 和 p 固定	不固定
存储空间的浪费	随着 x、y 和 z 的减少而增大	根据 x、y 和 z 的实际大小申请结点空间
单词间和单词的词义间的关系	不好对应, 处理麻烦	根据链指针便可方便得到对应关系(如图 1)

4 结 语

本文的前期工作是基于系统语料库进行分词, 但限于篇幅在此没有给出其算法, 只是在其分词的基础上进行了词的多义性处理, 设计了处理算法, 提出了句子链栈的存储思想, 该思想较栈结构[2,4] 的存储方式能节省存储空间, 也不失先进后出的特点, 有较好的研究价值。

参 考 文 献

[1] 马林兵, 龚健雅. 面向自然语言的空间数据库查询研究[J]. 计算机工程与应用, 2003(22): 16-18

[2] 崔宗军, 唐世渭, 杨冬青. 基于 ER 模型和受限汉语的数据库中文查询语言研究[J]. 中文信息学报, 2000, 15(4): 7-13

[3] 周炎坤, 李满春. 基于中文的 GIS 查询界面的初步研究[J]. 科技通报, 2001, 17(1): 35-39

[4] 许龙飞, 杨晓昀, 唐世渭. 基于受限汉语的数据库自然语言接口技术研究[J]. 软件学报, 2002, 3(4): 537-543

[5] 张连蓬, 刘国林, 江涛, 等. 受限自然语言查询在 GIS 中的应用[J]. 测绘学院学报, 2002, 19(4): 283-284

[6] 赵伟, 戴新宇, 尹存燕, 等. 一种规则与统计相结合的汉语分词方法[J]. 计算机应用研究, 2003(3): 23-25

[7] 姚天顺, 张桂平, 吴映明. 基于规则的汉语自动分词系统[J]. 中文信息学报, 1999, 4(1): 37-43

第一作者简介: 徐爱萍, 博士生, 副教授。现从事网上信息管理、Web GIS、自然语言理解等研究。

E-mail: xap1464@126.com

Algorithm for Understanding Word Meaning in GIS Chinese Inquiry Sentences

XU Aiping[1,2] *BIAN Fuling*[2] *CAO Jie*[1]

(1 School of Computer Sceience, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

(2 Reserch Center of Spatial Information & Digital Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

Abstract: In GIS Chinese inquiry system, the word meaning ambiguity has brought many troubles in understanding inquiry sentences correctly. The problem is analyzed carefully and a new algorithm is presented based on corpus which can solve ambiguity successfully. As an innovation, A new data structure, called "sentence link stack", is adopted in algorithm. The experimental results show that the algorithm can achieve accuracy of 95% and can offer valid information for the semantic analysis subsystem to draw inquiry targets, inquiry condition, group message and sort message.

Key words: corpus; ambiguity; sentence link stack; word meaning link; table link

About the first author: XU Aiping, Ph. D candidate, supervisor of postgraduate, associate professor. She is engaged in management information system on Web, WebGIS and natural language understanding.

E-mail: xap1464@126.com