基于语义文法的网络舆情精准分析方法研究

侯圣峦[1,2] 刘 磊[1] 曹存根[2]
(北京工业大学应用数理学院 北京 100124)[1]
(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)[2]

摘 要 传统的基于关键词统计分析的网络舆情分析方法由于缺少对舆情文本必要的语义处理,往往导致分析结果不准确。提出一种基于语义文法的网络舆情精准分析新方法。该方法包括两部分:首先是可执行的网络舆情精准分析语言 Eipoaal,它可根据实际舆情分析需求设计 Eipoaal 程序,具有一定的通用性;二是网络舆情精准分析系统 Ipoaas,它为 Eipoaal 提供运行平台。目前,已经实现了该系统,并应用到贪腐主题的网络舆情分析中,实验结果证明了方法的有效性。

关键词 网络舆情分析,语义文法,舆情本体,多主体
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.10.048

Research on Accurate Analysis of Internet Public Opinion: A Semantic Grammar-based Method

HOU Sheng-luan[1,2] LIU Lei[1] CAO Cun-gen[2]
(College of Applied Sciences, Beijing University of Technology, Beijing 100124, China)[1]
(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)[2]

Abstract The conventional methods of public opinion analysis based on keywords statistics are inaccurate due to lack of semantic processing which is necessary. A novel semantic grammar-based method for accurate analysis of Internet public opinion was presented. This method has two parts. One is an executable Internet public opinion accurate analysis Language (Eipoaal), which is a general-purpose program language that can be designed according to actual demand; and the other is an Internet public opinion accurate analysis system (Ipoaas), which provides a running platform for Eipoaal. This system has been implemented and tested in the analysis of Internet public opinion about corruption. Experimental results show the validity of the method.

Keywords Internet public opinion analysis, Semantic grammar, Ontology of public opinion, Multi-agent

1 引言

网络舆情[1,2]是指网民对现实生活中的某些焦点热点问题通过互联网表达和传播的各种情绪、态度和观点。随着互联网的普及以及以智能手机为主流的智能移动终端技术的快速发展,当前互联网日趋便携、开放和智能,成为重要的信息发布与传播平台。网民通过微博、博客、BBS 等诸多网络应用在第一时间发布和获取时事、政策、产品和服务等信息以及对信息进行反馈和传播,这些网络信息形成了网络舆情,它一定程度上体现了整个社会的心理、态度和情绪等。

我国网络发展迅速,但缺少有效监管以及相关法律法规,近年来,由于网络舆论失控而导致群体性事件频频爆发,网络舆情对政治经济生活秩序和社会稳定的影响与日俱增。因此,如何准确分析互联网上实时涌现的舆情信息,及时发现其中的问题已经显得十分迫切。从政府层面讲,了解和掌握舆情发展动态成为政府掌握民意、及时应对非常规突发事件的重要手段;对企业来说,实时掌握客户对产品及服务的评价,企业与客户之间的互动反馈更为及时与高效;对于社会个体,及时掌握自己在网络舆情中的状态,纠正负面舆论的影响,同时对于青少年等特殊群体的教育引导、净化网络环境等有重要意义。

网络舆情信息量巨大并且在互联网上广泛传播,这给及时、准确地分析舆情信息带来极大困难。国内外许多科研单位和企业相继开展了相关研究工作,取得了一定理论成果并开发出了网络舆情监测分析系统,如 TRS 网络舆情监测系统、天玑舆情监测系统等[3,4]。这些系统大都是基于关键词统计分析的文本挖掘方法[5-8],对网络舆情进行监测分析。但这类方法由于缺少对网络舆情文本必要的语义处理,往往导致分析结果不准确。近年来,专家学者一直在研究更加有效的方法,其中基于语义的内容识别方法[9]是当前研究的热点。现有的基于语义的成果都是以词为研究单位,并且人工参与操作过多,主观影响较大。网络舆情的精准分析需要对舆情

到稿日期:2013-11-14 返修日期:2014-03-01 本文受国家自然科学基金(91224006,61105040,61203284),科技部行业专项(201303107),北京市自然科学基金(4133085)资助。

侯圣峦(1989—),男,硕士生,主要研究方向为文本知识获取,E-mail:houshengluan1989@163.com;刘 磊(1979—),男,博士,副教授,主要研究方向为本体学习、数据挖掘;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。

信息进行智能处理，必须从语义和知识层面上增强对舆情信息的理解能力，句法分析[10]和语义分析[11]技术是解决该问题的重要手段。语义文法[12,13]能有效地实现语义分析，并能与句法分析紧密结合，产生对网络舆情文本准确唯一的理解，实现对网络舆情的精准分析。

本文提出了一种基于语义文法的网络舆情精准分析方法，首先是支持网络舆情精准分析的语义文法的设计，建立从形式文法符号到网络舆情语义的一个映射，然后以本体作为指导，利用语义文法将无结构的网络舆情文本转化成结构化的网络舆情语义表示。该方法包括两部分：可执行的网络舆情分析语言(Executable Internet public opinion accurate analysis language，Eipoaal)和网络舆情精准分析系统(Internet public opinion accurate analysis system，Ipoaas)。

其中，Eipoaal用以定义网络舆情分析所需的本体、模板常数、语义文法模式和语义动作，是为实现网络舆情精准分析而设计的通用编程语言，可满足对不同领域、不同结构的舆情文本的处理需求，具有一定的通用性。Eipoaal采用多主体[14]的思想，每个主体可针对不同特征的舆情文本采取不同的处理方式，每个主体功能专一、相互独立、协同合作，共同完成对网络舆情文本的处理。而Ipoaas以待处理的舆情文本和用Eipoaal编写的程序为输入，系统实现对Eipoaal程序的编译、调试和运行，输出结构化的语义表示。

本文第2节简述语义文法；第3节给出Eipoaal的具体定义及其详细设计方法；第4节介绍Ipoaas的实现；第5节通过实验对本文提出方法的优缺点进行全面讨论；最后总结全文。

2 语义文法简述

形式文法是句子结构分析的一个重要手段[15,16]，常被用于自然语言处理。根据形式文法可以将无结构的自然语言转化成结构化的语法分析树，进而生成自然语言的语义表示。从Chomsky形式语言分类的角度而言，自然语言一般是上下文相关的，但在实际应用中，上下文无关文法(Context-Free Grammar，CFG)常被用来表示自然语言。一方面是由于上下文无关文法形式简单，方便总结自然语言处理所需的文法。另一方面，上下文无关文法具有良好的代数性，便于计算，也便于在计算机上实现。

语义文法(Semantic Grammar，SG)是Burton[17]首次提出的概念，它是一种上下文无关文法。语义文法与普通句法文法的区别在于语义文法中的非终结符被赋予了领域语义，语义文法可以包含句法和语义层次上的非终结符，也可以只包含语义层次上的非终结符。语义文法的优点是可以直接从解析结果中获取句子的语义信息，而传统的基于句法文法的解析结果只能得到句子的句法信息。使用语义文法描述自然语言可以给出句子丰富的语义结构，但不足之处就是由于在文法规则上增加了语义信息，文法规则数量较大。

如何用语义文法来描述网络舆情文本是本文所研究的一个重要问题，文法符号本身是形式的，不具有任何语义信息，网络舆情语义文法须能确定从形式文法符号到网络舆情语义的映射。网络舆情语义文法描述了网络舆情文本的句子结构及语义信息，可直接从分析结果产生语义解释。因此，网络舆情语义文法是对网络舆情文本进行精准语义分析的依据。

3 Eipoaal的设计

由于网络舆情语义文法的多样性和复杂性，需要建立一种通用的机制，使网络舆情精准分析方法和不同语义文法相结合，完成网络舆情精准分析任务。依据这种想法，我们设计了一种面向网络舆情精准分析的通用编程语言Eipoaal，可根据舆情文本的结构和处理粒度需求设计Eipoaal程序，满足不同的处理需求。

3.1 Eipoaal的形式定义

Eipoaal包括本体的引用、常数定义、语义文法的定义以及主体定义，下面给出Eipoaal的形式定义：

定义1 可执行的网络舆情精准分析语言是一个四元组 $Eipoaal=\langle O,C,S,A\rangle$，其中：

(1)O是本体引用，用于指导主体中的操作；

(2)C是常数定义，将使用频率较高、意义相近、出现位置相似的词语或者符号定义成常数，供文法设计和主体定义调用；

(3)S是语义文法的定义，用于描述网络舆情文本内容和结构文法模式；

(4)A是主体定义，对网络舆情文本进行精准分析，每个主体包括激发条件、绑定的文法模式和语义动作。

Eipoaal的基本结构可以用BNF(Backus-Naur Form)表示，如图1所示。

```
〈Eipoaal〉::={〈本体引用〉}
              [〈常数定义〉]
              {〈语义文法定义〉}
              {〈主体定义〉}+
```

图1 Eipoaal的基本结构

Eipoaal以声明要加载的本体知识库开始，随后是常数定义，然后是语义文法和主体的定义。从BNF可以看出，Eipoaal可同时声明多个本体文件，常数定义只能出现一次，可以有多个语义文法的定义，但至少有一个主体定义。

Eipoaal包括以下关键字，这些关键字用于系统不同的内容模块，具有特殊含义，不允许被重新定义，分别是：

1)include：声明要加载的本体；

2)defconstant：定义常数；

3)defsyntax：定义用于描述待处理文本的文法模式；

4)defagent：定义进行舆情精准分析的主体；

5)上标号、下标号：用于界定主体可处理的文本；

6)模式：主体可处理的文本应满足的结构；

7)语义动作：主体所要执行的操作；

8)for，forall：用于说明语义动作执行的环境和条件，for的意义是在语法树中查找满足条件的结点并执行相应的操作，然后终止搜索；forall的意义是对于语法树中所有满足条件的结点都执行定义的操作。

Eipoaal还规定了一些具有特殊意义的符号，同样不允许被重新定义，分别是：

1)#：定义文本变量；

2)!：引用常数定义中的常数，形式为“〈!常数名〉”；

3)/：文本变量序列类型的分隔符；

4)$：定义文本变量中不允许出现的字符；

5)|:分隔定义在一起的左部相同的常数或产生式;

6)//、/*、*/:注释符号,分别是行注释和片段注释。

结合中文网络舆情文本的特点,为了简化程序设计,Eipoaal还提供了条件文本变量的定义方法。文本变量是一类特殊的终结符,用来表示一段长度和内容满足某种条件的字符串。条件文本变量包括:

1)非空字符串:表示文本变量不能是空串,定义方式为"〈#非空变量名〉"。

2)字符串序列:可表示由多个变量构成的序列,定义方式为"〈#变量名序列/〈!序列分隔符〉〉"。字符串序列需要指定序列各项的分隔符,当然,同一序列可以有多种分隔符,亦可指定不允许出现的字串。如"〈#字符串序列/〈!顿号〉/和$〈!阿拉伯数字〉〉"表示一个序列变量,序列的每个元素是一个不含有阿拉伯数字的字符串变量,元素之间用顿号或者"和"进行分隔。

3)特殊意义字符串:如表示正整数、无空格变量等,只需在相应的变量名前加"正整数"、"无空格",如"〈#正整数变量序列/〈! 顿号〉〉"表示以顿号隔开的数字串序列。

3.2 舆情本体的引用

一个领域的本体是指对该领域的一个清晰的、可共享的刻画[18]。针对网络舆情分析而言,舆情本体主要包括舆情所涉及的类(Class)、类间关系(Relationships)、类属性(Attributes)以及它们的各种逻辑约束(Axioms),舆情本体是一种特殊的领域本体。

网络舆情精准分析不仅仅是对舆情文本的简单处理,而且要在理解舆情文本的基础上实现内容分析。本文通过引入领域知识作为元知识,帮助理解舆情文本,并指导主体的正确执行行为。领域本体的研究和应用已经趋于成熟,本文不再赘述舆情本体具体设计方法。

为便于操作,我们用NKI本体语言(NKIL)来表示舆情本体,NKIL是一种框架描述语言[18]。舆情本体的引用在Eipoaal的开头进行声明,格式为:"# include〈本体文件路径〉"。

3.3 网络舆情语义文法的设计

3.3.1 网络舆情语义文法的定义

网络舆情语义文法(Semantic Grammar for Internet Public Opinion Analysis,SGIPOA)是Eipoaal程序的重要组成部分,是实现网络舆情精准分析的依据。结合网络舆情文本的特点,支持网络舆情精准分析的语义文法定义如下:

定义2 网络舆情语义文法 $SGIPOA=\langle V_N, V_T, T_K, T_C, T_V, \wp, \Im\rangle$,其中:

(1)V_N 是文法非终结符的集合,是一个非空有限集;

(2)V_T 是文法终结符的集合,是一个非空有限集,是用来最终判断输入文本是否与文法模式匹配的依据;

(3)T_K 是文法的关键字集合,直接以字符串的形式出现在文法定义的产生式中;

(4)T_C 是常数引用的集合,即对defconstant中常数的引用,例如"〈!逗号〉";

(5)T_V 是文本变量的集合,是一个非空有限集,例如"〈#非空字符串$〈!标点〉〉";

(6)$\wp$ 是由有限个文法产生式构成的集合,每个产生式以BNF表示,形式为 $N::=t_1|t_2|\cdots|t_n$,并且 $N\in V_N, t_i\in(T_K\cup T_C\cup T_V)$,定义方式可以是递归的,但只能以右递归形式定义;

(7)$\Im$ 是文法模式定义的集合,每个文法模式至少要在 $\wp$ 中某个产生式定义的左边出现一次,需要在defsyntax后明确指出,并且 $\Im\in V_N$。

并且满足:

(1)$V_N\cap T_K\cap T_C\cap T_V=\varnothing, T_c\cap T_v\neq\varnothing$;

(2)$T_K\cup T_C\cup T_V=V_T$。

另外,SGIPOA引入继承机制,文法的定义可以继承其他文法集合中的文法产生式定义,SGIPOA的定义也体现了面向对象的思想,即"一处定义,多处调用",降低了文法的冗余度。借助继承机制进行文法设计时,还允许被继承的文法继承其他文法产生式,即实现"多重继承"。借助BNF严谨的表示形式和灵活的继承机制,SGIPOA清晰灵活,满足网络舆情精准分析的需要。

下面给出一个SGIPOA示例,如图2所示。

```
defconstant 常数
{
  顿号:、
  标点:,|,|、|! |? |/|\|"|"|"|"
  强调标点:!!! |!! |!
  现在词:现在|如今|这年头
  程度副词:太|很|特|非常
  态度差形容词:差劲|恶劣|冷漠|差
}
defsyntax〈通用模式〉
{
  〈服务方〉::=〈#服务机构或人员序列/〈! 顿号〉$〈!标点〉〉
}
defsyntax〈服务态度语句〉:继承〈通用模式〉
{
  〈服务态度语句〉::=[〈!现在词〉]〈服务方〉[的][服务]态度[〈!程
  度副词〉]〈!态度差形容词〉[了][〈!强调标点〉]
}
```

图2 SGIPOA示例

其中"〈服务方〉"继承"〈通用模式〉"中的产生式定义,"〈#服务机构或人员序列/〈!顿号〉$〈!标点〉〉"表示以顿号为分隔符并且不含有标点的任意字符串,"[的]"、"[服务]"、"[了]"表示可选关键字,"[〈!现在词〉]"、"[〈!程度副词〉]"表示可选的常数引用,"〈!态度差形容词〉"表示必选的常数引用。利用该文法,可分析如"现在银行的服务态度太差了"、"莱州泰和驾校教练态度太差了"、"火车站工作人员服务态度真差劲"、"麦当劳、肯德基服务态度太差了!!"等语句。

3.3.2 SGIPOA的设计原则

SGIPOA是网络舆情分析的关键内容,是对网络舆情文本进行精准语义分析的基础。为了保证文法易读并且利于维护,在方便进行文法设计的基础上,我们认为,语义文法的设计应遵循以下原则:

(1)语料准备要全面:在准备进行文法设计的语料时,应针对领域的特点,语料涉及到该领域的内容要广泛,尽量涉及到不同类型不同结构的句子。只有语料全面、采样广,才能保证文法的泛化能力更强。

(2)命名要规范:非终结符和终结符的命名必须能有效地

说明所定义内容的含义，并且准确、无歧义，符合一般的语言表达习惯。例如用“发生于”表示事件发生的地点就不规范，因为“发生于”根据语境的不同可以理解为事件发生的地点也可理解为事件发生的时间，用“发生地点”表示事件发生的地点更合适。

(3)合理使用“无关内容”，确保无二义性：由于汉语表达方式的灵活性，一些句子中可能只有一部分是我们需要抽取的内容，可以将其他部分定义为“无关内容”。但如果过多不合理使用“无关内容”，程序会误将有用的知识当作无关内容处理，产生歧义，合理的做法是将“无关内容”进行语义约束，例如“〈＃无关内容＄〈!房产词〉〉”可用于关于“房产”的文法设计中。

(4)可扩展性：文法的可扩展性是指，可以在保证已有文法性能不受影响的前提下，在已有文法定义的基础上定义新的文法，而无须大量修改已有的文法定义。也就是说，文法应层次清晰、结构鲜明，能根据需要满足扩展和改进的需求。

3.3.3 SGIPOA 的设计方法

根据上述 SGIPOA 的设计原则，结合网络舆情文本的特点，我们总结归纳出 SGIPOA 的设计步骤如图 3 所示。

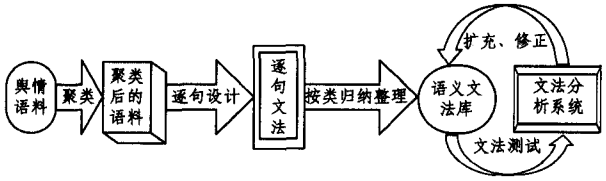

图 3 SGIPOA 设计步骤

步骤 1：对网络舆情文本 $D=\{d_1,d_2,\cdots,d_n\}$ 进行聚类，形成舆情文本的类别 $C=\{c_1,c_2,\cdots,c_m\}$，从整体上把握网络舆情的体系。

步骤 2：对步骤 1 聚类得到的每一类舆情文本 $c_i(i=1,2,\cdots,m)$ 以句子为单位进行初步舆情文法设计，人工总结形成文法规则。

在进行初步文法设计时，应注意以下问题：

①将使用频率较高、意义相近、出现位置相似的词语或者符号定义成常数，并利用同义词词林或者联想的方法进行扩充，例如“穿戴词”，还可扩充为“穿”、“拿”、“着”，因此“穿戴词”的常数定义为“穿戴词：戴|穿|挎|拿|着”；

②将常用短语定义成非终结符，并将定义类似或者定义中包含相同定义的非终结符合并，实现“一次定义，多次调用”；

③文法定义中的某个部分可能根据语境的不同出现或者不出现，因此将该部分设计成是可选的，在保证匹配效率的同时也减少了文法定义的数目，例如“〈省市县〉”定义为“〈省市县〉::=〈省〉[〈市〉][〈县〉]|〈市〉[〈县〉]|〈县〉”；

④为了提高文法的灵活性和匹配效率，合理使用递归给文法定义带来极大方便，例如“〈穿戴奢侈短语连用〉::=〈穿戴奢侈短语〉[〈!逗号〉][〈穿戴奢侈短语连用〉]”，但不能含有左递归，它会使分析陷入左循环。

步骤 3：将步骤 2 中逐句设计的文法进行归纳整合。利用文法模式的继承机制，将文法定义中公共部分整合，形成“通用模式”，供上层文法调用。

步骤 4：调用文法分析系统进行文法测试，并根据测试结果对文法匹配错误的、未匹配的语料再次分析。该步骤包括常数定义的扩充、文法的修改和扩充，对匹配错误的变量进行限制。

①扩充常数：对测试语料进行测试后，部分分析一句话，对未匹配的部分进行分析，基于语境，扩充常数。

②修改及扩充文法模式：如果文法定义不全，就必然导致查全率低，因此应尽可能全面地总结能描述所有语句的文法。具体方法是将未识别并且含有知识点的句子进行分析，总结该句文法，扩充到已有文法中。

③限制变量：一些情况下，某个变量中不允许一些字符出现，例如“〈＃官员姓名〉”可能匹配到的字符数很多或者包含标点符号，显然这是错误的，如果我们将该变量限制为不含有“〈!标点符号〉”，即“〈＃官员姓名＄〈!标点符号〉〉”，则可以避免该错误。

SGIPOA 的设计是一个不断补充和完善的迭代过程，最终形成领域内涵盖知识面广、准确率高的语义文法。通过以上步骤，可以得到支持网络舆情精准分析的语义文法。根据上述方法，我们以贪腐主题的网络舆情文本为例，设计了贪腐领域的语义文法，简称“贪腐文法”。贪腐文法的详细介绍及示例将在实验部分给出。

3.3.4 SGIPOA 的评价

SGIPOA 的设计好坏直接影响网络舆情精准分析系统的整体性能，SGIPOA 的评价主要通过具体的舆情分析系统分析得到的结果来衡量。这里我们从文法的泛化能力和文法匹配的准确率两方面进行评价，文法的泛化能力是指文法覆盖该领域范围内知识的广度以及是否能获取到语料中所包含的领域内知识，文法匹配的准确率即对句子识别的准确度。

3.4 主体的定义

在 Eipoaal 程序中主体是必不可少的部分，网络舆情文本的分析和处理工作都是在主体中进行的。Eipoaal 程序中可根据需要定义一个或多个主体，每个主体分析处理符合某一文法的网络舆情文本。每个主体都是自治的，多个主体协同合作，共同完成网络舆情精准分析任务。主体的定义格式如图 4 所示。

〈主体定义〉::=defagent〈主体名〉:主体本体
‘{’
上标号:〈标号序列〉
下标号:〈标号序列〉
模式:〈文法模式名〉
{语义动作:[〈语义动作序列〉]}+
‘}’
〈标号序列〉::=〈已定义常数**〉|〈文本字符串*〉
〈文法模式名〉::=〈语义文法名〉|〈已定义常数〉|〈文本变量〉
〈语义动作序列〉::=[〈条件表达式〉→]〈操作函数**〉{∧〈操作函数**〉}
〈条件表达式〉::=for(〈操作结点〉)(〈条件〉)| forall(〈操作结点〉)(〈条件〉)
〈条件〉::=〈布尔函数**〉{〈逻辑符号**〉〈布尔函数**〉}

图 4 主体定义格式

其中，标有 * 的表示是用户自己定义的或文本中可能出现的，所以不需要进一步展开定义。标有 * * 的表示系统提供的函数，也不需要进一步展开定义。系统定义函数如表 1 所列。

表 1　系统函数

函数类型	函数符号	函数功能
逻辑符号	∧	逻辑与
	∨	逻辑或
	¬	逻辑非
布尔函数	streq(〈参数 1,参数 2〉)	比较两个参数是否相等
	isstrlen(〈参数 1,Len〉)	判断〈参数 1〉的长度是否是 Len
	contain(〈参数 1,参数 2〉)	判断〈参数 1〉是否包含〈参数 2〉
	begin-with(〈参数 1,参数 2〉)	判断〈参数 1〉是否以〈参数 2〉开头
	end-with(〈参数 1,参数 2〉)	判断〈参数 1〉是否以〈参数 2〉结尾
字符串拼接操作函数	strcat(〈常数字符串〉,〈常数字符串〉{,〈常数字符串〉})	将多个常数字符串拼接起来,返回拼接后的新字符串
	prefix(〈字符串序列〉,〈前缀〉)	在序列类型的文本变量〈字符串序列〉的每一项前面加上〈前缀〉
	suffix(〈字符串序列〉,〈后缀〉)	在序列类型的文本变量〈字符串序列〉的每一项前面加上〈后缀〉
字符串存取操作函数	save(〈存储名〉,〈值〉)	将〈值〉以〈存储名〉为名保存在公共变量存贮区
	fetch(〈存储名〉)	返回公共变量存贮区中〈存储名〉对应的值
类及实例操作函数	close-category(〈本体类名〉)	关闭属于〈本体类名〉的所有实例框架
	close-all()	关闭所有本体类的所有框架,禁止访问和插入操作
	create-frame(〈实例框架名〉,〈本体类名〉,〈分隔符〉)	新创建一个〈本体类名〉类的实例,写成框架形式,命名为〈实例框架名〉
	close-frame(〈实例框架名〉)	关闭〈实例框架名〉对应的框架,禁止对其进行访问和插入操作
	insert-category(〈本体类名〉,〈槽名〉,〈槽值〉,〈分隔符〉)	在〈本体类名〉的框架中插入一个以〈槽名〉为名,以〈槽值〉为值的槽

图 5 所示为一个主体定义的示例,用以扼要说明主体基本结构的及执行机制。

```
defagent 生活腐败主体:主体本体
{
  上标号:"*"
  下标号:〈!标号〉
  模式:〈生活腐败〉
  语义动作:close-category("贪腐本体")
  语义动作:for(〈指名道姓主语〉)->create-frame(〈#官员名$〈!人物名干扰词〉〉,"贪腐本体")∧save(framename,〈#官员名$〈!人物名干扰词〉〉)∧insert-category("贪腐本体","官员名",〈#官员名$〈!人物名干扰词〉〉,",和 ")∧insert-category("贪腐本体","官职",strcat(〈省市县连用〉,〈官职名称连用〉),",和 ")
  语义动作:forall(〈生活作风差短语〉)->insert-category("贪腐本体","生活作风差",〈生活作风差短语〉,",和 ")
  语义动作:forall(〈穿戴奢华短语〉)->insert-category("贪腐本体","穿戴奢华",〈穿戴奢华短语〉,",和 ")
  语义动作:forall(〈违法配车辆和司机短语〉)->insert-category("贪腐本体","违规配车辆和司机",〈违法配车辆和司机短语〉,",和 ")
}
```

图 5　主体定义示例

图 5 所示的主体是处理贪污腐败主题的舆情文本中"生活腐败"文本的主体,当输入舆情文本满足主体的激发条件(上标号、下标号定义)时,用该主体绑定的文法模式"〈生活腐败〉"对文本进行语法分析。如果分析成功,则会根据分析生成的语法树继续执行主体定义的语义动作,进而得到预定义的语义表示结果。通常,语义动作都是根据舆情文本的结构特点和处理结果需求来定义的。因此,主体中语义动作的定义与绑定的文法模式有关,语义动作也是主体自治性的表现。主体的正确执行则完成了对满足激发条件的舆情文本的处理。

4　Ipoaas 的实现

本文研究实现了基于语义文法的网络舆情精准分析系统 Ipoaas,其为 Eipoaal 程序提供运行平台,也称为舆情语义文法分析系统。系统输入为 Eipoaal 程序和待处理文本,输出结构化的语义表示,同样以 NKIL 表示。系统功能是实现对网络舆情文本的语义分析,并按照主体定义的语义动作在本体的指导下将分析得到的语法树转化成结构化的语义表示。

相关研究中,王海涛[19]提出了一种基于本体的半结构化文本知识处理方法(Ontology-Mediated Knowledge Processing for Semi-structured Text,OMKP)。OMKP 方法把本体知识引入文本处理过程中,以半结构化文本为主要处理对象,利用多主体思想,结合模式匹配的技术,实现了文本知识的自动获取和理解。OMKP 把知识处理过程和具体应用分离开,但同时又以本体知识为桥梁把二者联系起来。从而使得 OMKP 方法具有较好的通用性,具有应用、领域、语言无关的特点。此外,OMKP 在自由文本处理方面也进行了尝试。OMKP 方法最大的问题在于处理内在规律性和结构性的半结构化文本时效率较高,但是对于无结构的自由文本,随着文法数量的增加知识处理效率较低。

本文在系统设计上吸收了 OMKP 的设计思想,并在此基础上加以改进和完善。在系统的文法匹配模块提出了一种启发式的匹配算法,该算法将依据启发式规则对文法非终结符的子树进行剪枝操作,以加快匹配速度,这使系统性能得到显著提升。系统结构图如图 6 所示。

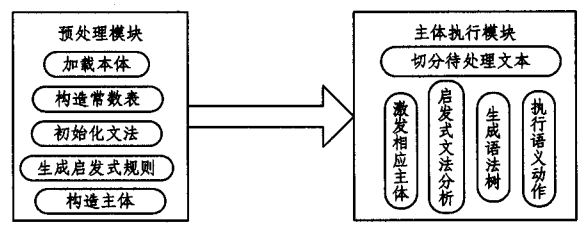

图 6　系统结构图

预处理模块是系统对 Eipoaal 程序的编译和初始化阶段,包括 5 个步骤,除生成启发式规则外,其余 4 个步骤都需要进行词法检查和语法检查,即保证 Eipoaal 格式正确。

(1)加载本体:本体知识用来指导主体对舆情文本的语义理解,加载本体知识库就是把以文本形式 NKIL 格式保存的知识库读入内存,构造本体表,供主体访问检索。

(2)构造常数表:在内存中对 Eipoaal 程序中定义的常数构造相应的常数表,供主体在需要时高效访问该表的内容。

(3)初始化文法:Eipoaal 程序中的文法规则是对上下文无关文法的扩展,并对文法产生式的书写进行了限制。因此,除 Eipoaal 的词法检查和语法检查外,初始化文法应检查文法定义是否符合规定格式。例如除关键字外,其他文法符号都必须以"〈"和"〉"括起来,文法产生式左端必须是非终结符,只允许左递归等,并处理文法继承。

(4)生成启发式规则:文法中的启发式规则是指找出每个非终结符的开始字符集和包含字符集,文法匹配时利用该规则可有效提升系统性能。

(5)构造主体:包括将主体的激发条件、绑定的文法模式存入内存的主体表中,将语义动作和语法节点关联起来,并解析语义动作的各个参数。

通过以上步骤,Eipoaal 程序就被成功编译完成。接下来就进入文法匹配阶段,读入待处理的舆情文本,执行 Eipoal 程序:根据主体定义可将待处理文本按句或者片段切分成小的单位,按照主体绑定的文法模式进行启发式的文法分析,若该主体分析失败则尝试下一主体,主体分析成功后根据生成的语法树在舆情本体的指导下执行语义动作,然后按照上述步骤继续处理剩余未处理的文本。Eipoaal 程序的执行过程如图 7 所示。

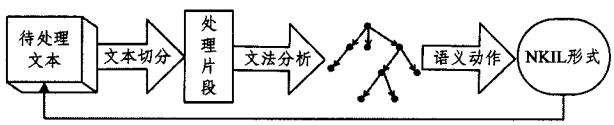

图 7 Eipoaal 程序的执行过程

Eipoaal 程序的执行主要包括以下 3 个步骤:

(1)文本切分:将待处理的网络舆情文本根据主体的上标号和下标号进行切分,则切分后的句子或者片段满足该主体的激发条件,主体被成功激活。

(2)文法分析:主体被激活后,系统按照主体绑定的文法模式进行启发式的文法匹配。若分析成功则生成语法树,若分析不成功则尝试下一个主体,若所有主体都失败,则该舆情文本片段分析失败。该步骤是主体执行的关键步骤,系统在文法分析算法中引入了启发式规则和回溯机制,在保证分析结果准确的前提下提高了文法分析的效率。详细匹配算法如下:

Step 1 将文法开始符(记为 N_0)压入待匹配的文法栈 L 中。

Step 2 将待匹配串 s 进行文法分析:如果 s 为空且 L 为空,文法分析成功,转到 Step 3;否则,如果 L 为空但 s 不为空,该主体文法分析失败,错误退出;其他情况下,从栈 L 中弹出一个节点,记为 v,并从 L 中删除 v。

CASE 1 若 $v\in V_N$:判断 v 的产生式定义是否满足启发式规则,若满足,则将其展开,并将展开的文法符号按照书写相反的顺序压入栈 L 中,重回 Step 2;若不满足,判断 v 是否是可选项,若是,v 的匹配串为空串,若否,回溯,重回 Step 2。

CASE 2 若 $v\in T_K$:判断 v 是否在 s 的开头出现,若是,从 s 中截取 v 对应的部分,并从 s 中删除该部分,重回 Step 2;若否,判断 v 是否是可选项,若是,v 的匹配串为空串,若否,回溯,重回 Step 2。

CASE 3 若 $v\in T_C$:调取 v 对应的常数表,顺序查找是否含有在 s 的开头出现的常数,若找到,从 s 中截取 v 对应的部分,并从 s 中删除该部分,重回 Step 2;若未找到,判断 v 是否是可选项,若是,v 的匹配串为空串,若否,回溯,重回 Step 2。

CASE 4 若 $v\in T_V$:弹出栈 L 的下一个元素(记为 v′),并从 L 中删除 v′:

①若 $v'\in V_N$,先按照 CASE 1 的方式处理 v′,后将 v 压入栈 L 中。

②若 $v'\in T_K$,处理方式与 CASE 2 类似,不同之处是此处判断 v′是否在 s 中出现,v 的可能匹配串为 s 中 v′之前(不包括 v′)的字符串。检查该部分是否满足变量 v 的语义限制,若满足则匹配成功,重回 Step 2;如不满足则回溯,重回 Step 2。

③若 $v'\in T_C$,处理方式与 CASE 3 类似,并按照②的方法验证是否满足变量的语义限制。

④若 $v'\in T_V$,由于我们规定文法定义中两个变量不能相邻,故该文法产生式定义是错误的,直接回溯,重回 Step 2。

注:以上回溯是指,假设匹配到节点 n 需要回溯,则恢复到即将处理 n 的情况,并将 n 已经匹配的串恢复到 s 中,考虑产生式定义的其他情况,若不存在其他情况,则向更上一层回溯。

Step 3 将参与文法匹配的文法符号及其对应结果按层次输出,生成语法树。

(3)执行语义动作:根据生成的语法树,在本体的指导下,执行主体定义中的语义动作,此处我们同样用 NKIL 语言来表示舆情精准分析后的处理结果。

目前,我们已经在 Windows 系统下用 Visual C++实现了该系统,系统界面可操作性强,并提供了丰富的编辑和调试功能,为网络舆情精准分析提供了良好的平台环境。

5 实验与讨论

为了验证上述方法的有效性,本文将上述方法应用到贪污腐败主题的网络舆情分析中,在上述系统环境下进行实验。我们从微博、博客、BBS 等热门网络应用中抽取了贪腐主题的网络舆情语料,构成该实验的语料库。首先按照本体建模的一般原则并结合本领域的特点,设计了贪腐本体,共 30 个属性,以 NKIL 表示,作为对贪腐主题的网络舆情分析的指导;然后根据上述 SGIPOA 的设计方法,设计了贪腐主题的网络舆情语义文法;最后定义了贪腐主题的网络舆情分析所需的主体,以上步骤的结果构成了贪腐主题的 Eipoaal 程序的各个部分,图 8 为部分 Eipoaal 程序示例:

```
#include〈贪腐本体.txt〉
defconstant 模板常数
{
  逗号:,|,
  顿号:、
  冒号::|:
  …
  存款词:银行存款|银行储蓄|银行存储|存款|储蓄|存储
  购买词:购买|购置|购下|购得|买下|买来|买了|购|买
  收入词:收入|得到|收得|收到|所得
  …
  时间量词:年|月份|月|日|周|天|时|点|分|秒
  阿拉伯数字:1|2|3|4|5|6|7|8|9|0
  …
  房产词:豪宅|别墅|房产|房子|房屋|私家住宅|楼房|房
  …
```

```
}
defsyntax〈通用模式〉
{
  〈指名道姓主语〉::=〈指名道姓主语 1〉|〈指名道姓主语 2〉|…|〈指
                名道姓主语 4〉
    〈指名道姓主语 1〉::=[姓]〈!姓〉[的]|〈!姓〉姓[的]|〈!姓〉
                〈行政单位官职〉
    〈指名道姓主语 2〉::=[〈省市县连用〉]〈官职名称连用〉〈#官
                员名 $〈!人物名干扰词〉〉|〈#非空官
                员名 $〈!人物名干扰词〉〉[〈行政单
                位〉]〈行政单位官职〉
                …
    …
  〈家属移民短语〉::=〈家属统称〉[〈!已经词〉][〈!全部词〉]〈!移
                民词〉[〈!到达词〉][〈国外目的地〉][了]
…
  〈生活腐败短语连用〉::=〈生活腐败短语〉[〈!逗号〉][〈生活腐败
                短语连用〉]
    〈生活腐败短语〉::=〈穿戴奢华短语〉|〈生活作风差短语〉|〈吃
                住奢华短语〉|…
      〈穿戴奢华短语〉::=[〈!身体部位〉]〈!穿戴词〉〈奢侈品〉
        〈奢侈品〉::=〈#奢侈品名 $〈!物品名干扰词〉〉〈!衣服首
                饰〉
  …
}
defsyntax〈生活腐败语句〉:继承〈通用模式〉
{
  〈生活腐败语句〉::=〈生活腐败语句 1〉|〈生活腐败语句 2〉|…
    〈生活腐败语句 1〉::=〈指名道姓主语〉〈利用职权短语〉[〈!逗
                号〉]〈生活腐败短语连用〉
    〈生活腐败语句 2〉::=〈指名道姓主语〉[〈!逗号〉][〈在任期间短
                语〉][〈!逗号〉]〈生活腐败短语连用〉[〈句
                尾带逗号无关〉]|…
  …
}
defagent 生活腐败主体:主体本体
{
  上标号:"*"
  下标号:〈!标号〉
  模式:〈生活腐败〉
  语义动作:close-category("贪腐本体")
  语义动作:for(〈指名道姓主语〉)->create-frame(〈#官员名 $〈!
          人物名干扰词〉〉,"贪腐本体")∧save(framename,〈#官
          员名 $〈!人物名干扰词〉〉)∧insert-category("贪腐本
          体","官员名",〈#官员名 $〈!人物名干扰词〉〉,",和 ")
          ∧insert-category("贪腐本体","官职",strcat(〈省市县
          连用〉,〈官职名称连用〉),",和 ")
  语义动作:forall(〈生活作风差短语〉)->insert-category("贪腐本
          体","生活作风差",〈生活作风差短语〉,",和 ")
  语义动作:forall(〈穿戴奢华短语〉)->insert-category("贪腐本
          体","穿戴奢华",〈穿戴奢华短语〉,",和 ")
  语义动作:forall(〈违法配车辆和司机短语〉)->insert-category
          ("贪腐本体","违规配车辆和司机",〈违法配车辆和司
          机短语〉,",和 ")
}
defsyntax〈贪污受贿语句〉:继承〈通用模式〉
{
  〈贪污受贿语句〉::=〈贪污受贿语句 1〉|〈贪污受贿语句 2〉|…
  …
}
defagent 生活腐败主体:主体本体
{
  …
}
…
```

图 8　贪腐主题的 Eipoaal 程序示例(部分)

系统根据我们定义的贪腐主题的 Eipoaal 程序将贪腐主题语料进行分析,得到 NKIL 形式的框架语义表示结构。上述示例中的 defframe 即为贪腐本体的 3 个类实例,系统将被分析语料中包含贪腐本体定义属性的知识抽取出来,并能结合上下文信息形成框架结构。由分析结果可知,该结构清晰明了,并利于进一步的处理。

对于网络文本,因为其口语化,会出现省略、倒装等情况。针对该问题,只要我们设计了分析这类语句的文法,并在主体定义中结合上下文信息,正确利用字符串操作函数,完全能得到精准的框架语义表示形式。

为了从整体上把握贪腐主题的网络舆情文本的特点,我们在进行语义文法设计时将测试语料分成生活腐败、操控工程等 18 个大类。由于一条测试语料中可能含有多个类别的知识点,故每类测试语料中允许含有其他类别的知识点。我们抽取含有生活腐败、贪污受贿、房产财产、操控工程 4 个类别知识点的语料作为实验用例,随机抽取 228 条语料,手工对语料中含有的 4 类知识点进行语义标注,标注后,将实验结果与标注结果对比,结果统计如表 2 所列。

表 2　实验结果统计表

分类	语料中包含数目	分析数目	分析正确数目	查全率	准确率
生活腐败	85	81	79	95.29%	92.94%
贪污受贿	62	56	56	90.32%	90.32%
房产财产	79	77	74	97.47%	93.67%
操控工程	43	41	40	95.35%	93.02%

由实验结果可知,利用该方法可实现网络舆情的精准分析,并且系统运行效率较高,说明了该方法的有效性。虽然分析的准确率和查全率都达到较高水平,分析错误的主要原因有两个:

(1)文法匹配错误:文法匹配错误是因为对文法符号的语义约束不强,产生错误匹配。解决方法是在不影响原有语义文法模式的泛化能力的前提下,对造成错误匹配的文法产生式加强语义约束或者增加新的文法产生式规则。

(2)文法未匹配到:文法未匹配到是因为不存在描述该语料的文法或者已有文法的语义约束太强。解决方法一般是在原有的文法产生式规则上增加新的文法规则。

结束语　为了克服传统的网络舆情分析方法对舆情文本缺少必要的语义处理的不足,本文提出了一种从语义和知识层面上实现网络舆情精准分析的新方法。首先定义了 Eipoaal 语言,然后设计实现了高效的网络舆情精准分析系统,

(下转第 237 页)

Routing Problem[J]. European Journal of Operational Research,2004,157(3):552-564
[7] 符卓.带装载能力约束的开放式车辆路径问题及其禁忌搜索算法研究[J].系统工程理论与实践,2004,24(3):123-128
[8] 钟石泉,杜纲.基于核心路径禁忌搜索算法的开放式车辆路径问题研究[J].计算机集成制造系统,2007,13(4):827-832
[9] 李湘勇,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
[10] 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
[11] Hansen P,Mladenovic N,Brimaerg J,et al. Variable Neighborhood Search[M]//Gendreau M,Potvin J. Handbook of Metaheuristics(Second Edition). New York:Spring Science+ Business Media,LLC,2010:61-86
[12] Mladenovic N,Hansen P. Variable Neighborhood Search[J]. Computers & Operations Research,1997,24(11):1097-1100
[13] Marti R, Moreno-Vega J M, Duarte A. Advanced Multi-Start Methods[M]//Gendreau M,Potvin J. Handbook of Metaheuristics(Second Edition). New York: Spring Science + Business Media,LLC,2010:265-281
[14] Marti R,Resende M G C,Ribeiro C C. Multi-Start Methods for Combinatorial Optimization[J]. European Journal of Operational Research,2013,226(1):1-8
[15] Salehipour A, Sorensen K, Goos P, et al. Efficient GRASP+VND and GRASP+VNS Metaheuristics for the Traveling Repairman Problem[J]. 4OR,2011,9(2):189-209
[16] 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
[17] Schittekat P, Kinable J, Sorensen 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
[18] Nguyen V P,Prins C,Prodhon C. Solving the Two-Echelon Location Routing Problem by a GRASP Reinforced by a Learning Process and Path Relinking[J]. European Journal of Operational Research,2012,216(1):113-126
[19] Beasley J E. Route First-Cluster Second Methods for Vehicle Routing[J]. Omega,1983,11(4):403-408
[20] Prins C. A Simple and Effective Evolutionary Algorithm for the Vehicle Routing Problem[J]. Computers & Operations Research,2004,31(12):1985-2002
[21] Archetti C,Speranza M G,Hertz A. A Tabu Search Algorithm for the Split Delivery Vehicle Routing Problem[J]. Transportation Science,2006,40(1):64-73

(上接第 231 页)

为 Eipoaal 提供编译和运行平台。通过在贪腐主题的网络舆情分析中的应用表明,该方法无论是处理长文本还是微博短文本都有效,生成的结果不仅可读性强,而且便于进一步处理。本研究将为新一代基于语义的舆情监测分析系统提供理论基础和技术支持,具有重要的理论意义和明显的实用价值。

基于语义文法的网络舆情精准分析方法可操作性强,系统执行效率高,满足对不同结构、不同处理粒度的网络舆情文本的处理需求,具有通用性。该方法可扩展到其他领域的 Web 自由文本分析处理中。今后的主要研究工作将包括:(1)语义文法的自动学习,手工总结网络舆情分析所需文法虽然准确率高但效率低,海量文本条件下,语义文法不易通过手工方式得到全面总结,文法的泛化能力难以保证;(2)在现有方法中增加正确性验证功能,由于网络舆情文本灵活多样,需要对分析结果的正确性进行验证,同时检测缺失、模糊、歧义、冗余等情况。

参 考 文 献

[1] 方薇,何留进,宋良图.因特网上舆情传播的预测建模和仿真研究[J].计算机科学,2012,39(2):203-205,235
[2] 许鑫,章成志,李雯静.国内网络舆情研究的回顾与展望[J].情报理论与实践,2009,32(3):115-120
[3] 李忠俊.基于话题检测与聚类的内部舆情监测系统[J].计算机科学,2012,39(12):237-240
[4] 丁杰,徐俊刚.IPSMS:一个网络舆情监控系统的设计与实现[J].计算机应用与软件,2010,27(4):188-190
[5] 戴媛,程学旗.面向网络舆情分析的实用关键技术概述[J].信息网络安全,2008(6):62-65
[6] 黄晓斌,赵超.文本挖掘在网络舆情信息分析中的应用[J].情报科学,2009(1):94-99
[7] 王灵芝.高校学生网络舆情分析及引导机制研究[D].长沙:中南大学,2010
[8] 万源.基于语义统计分析的网络舆情挖掘技术研究[D].武汉:武汉理工大学,2012
[9] 王兰成,刘晓亮.整合中文维基语义的网络论坛话题追踪方法研究[J].情报学报,2013,32(1):22-27
[10] 刘挺,马金山.汉语自动句法分析的理论与方法[J].当代语言学,2009(2):100-112,189
[11] Poon H,Domingos P. Unsupervised semantic parsing[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics,2009:1-10
[12] Roberts C W,Zuell C,Landmann J,et al. Modality analysis:a semantic grammar for imputations of intentionnality in texts[J]. Quality & Quantity,2010,44(2):239-257
[13] Rao G, Agarwal C, Chaudhry S, et al. Natural language query Processing using semantic Grammar[J]. International Journal on Computer Science and Engineering,2010,2(2):219-223
[14] 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
[15] Klein D. The unsupervised learning of natural language structure[D]. Stanford University,2005
[16] Hopcroft J E. Introduction to Automata Theory,Languages,and Computation,3/E[M]. Pearson Education India,2008
[17] Burton R R. Semantic grammar: an engineering technique for constructing natural language understanding systems[J]. ACM SIGART Bulletin,1977(61):26-26
[18] Cao C,Feng Q,Gao Y,et al. Progress in the development of national knowledge infrastructure[J]. Journal of Computer Science and Technology,2002,17(5):523-534
[19] 王海涛.文本知识处理方法及智能叙事生成应用研究[D].北京:中科院计算所,2008