Agent 驱动的中文本体智能构建研究

向 阳,张 波,韩 婕
XIANG Yang,ZHANG Bo,HAN Jie

同济大学 电子信息与工程学院,上海 201804
College of Electronics and Information Engineering,Tongji University,Shanghai 201804,China
E-mail:shxiangyang@vip.sina.com

XIANG Yang,ZHANG Bo,HAN Jie.Agent driven intelligent construction of Chinese ontology.Computer Engineering and Applications,2009,45(10):133-137.

Abstract: Intelligent construction of Chinese ontology is the hot topic in ontology research field now.Based on the features of Chinese ontology,technology of Agent,which has the characteristics of agility and autonomy,is introduced to construct an agent based intelligent construction of Chinese ontology model.This model is composed of three layers:data extraction layer,ontology construction layer and ontology treating layer.In each layer,there are different professional agents to do the targets of ontology construction.Then,in order to enhance the accuracy of ontology construction,improving strategies of term extraction and relationship extraction are proposed.

Key words: Agent;Chinese ontology;intelligent construction

摘 要:中文本体的智能构建是当前的研究热点。在中文本体特点的基础上,引入 Agent 技术,利用其灵活性、自治性等特点,提出了一个基于 Agent 的中文本体智能构建模型。该模型分为数据抽取层、本体创建层和本体处理层。每一层中由不同的职能 Agent 完成本体构建的任务。进而提出了术语抽取和关系抽取改进策略,提高本体构建的准确性。

关键词:Agent 技术;中文本体;智能构建
DOI:10.3778/j.issn.1002-8331.2009.10.040 **文章编号**:1002-8331(2009)10-0133-05 **文献标识码**:A **中图分类号**:TP391

近年来,语义网技术的发展以及本体应用领域的广泛使构建本体的需求激增。而传统的本体构建大多为人工方式,耗时耗力,无法满足人们对大量使用的需求。为了符合这种需求,利用计算机技术来自动构建本体的构想应运而生。然而,本体自动构建的研究尚在起步阶段,自动构建的本体质量不高;本体构建自动化的相关研究也多停留在对本体构建中某一步骤的自动化或是基于某个应用领域的本体构建[1]。同时,对于国内的用户而言,中文本体大量缺乏,直接影响到用户的使用。现存的研究大都是基于外文数据源来探讨本体自动构建的问题,由于不同语种语言特点和思维方式的不同,导致基于外文数据源的本体自动构建方法对中文数据源的适用性较差。

智能化的中文本体构建需要在符合本体构建的原则基础上[2],创造一个能够自动进行语法分词、本体术语抽取、关系抽取、本体编码等工作的计算环境。目前已经出现许多本体构建的技术[3-5],然而这些研究并非专门针对中文本体提出的,而且自动构建过程中存在智能化不足,准确度不够等问题。因此,引入 Agent 技术,利用 Agent"在特定环境中,能够满足特定设计要求,具有高度的灵活性、自治性"[6]的特点,来进行智能化的中文本体构建。

本文将本体构建分为三个主要层次:数据抽取层、本体创建层和本体处理层,其中,数据抽取层是整个本体自动构建体系的核心。设计了分工明确的不同类型 Agent,包括分词 Agent、术语抽取 Agent、关系抽取 Agent、本体创建 Agent、本体编码 Agent 等,由这些 Agent 来进行自动的实施构建工作。进而,为了提高本体构建中术语抽取、关系构建等方面的准确性,本文提供了一系列的改进策略,使中文本体的构建能够达到最佳效果。

1 基于 Agent 的本体智能构建系统

根据本体自动构建本身的特点,以及通用的本体构建步骤,为基于 Agent 中文本体智能构建体系设计了三个主要层次:数据抽取层、本体创建层和本体处理层。用户首先导入中文文本数据到数据抽取层,数据抽取层对数据进行处理,从中抽取出术语和关系;最后,本体创建层根据数据抽取层抽取到的数据来创建本体;本体创建完成后,本体处理层可对其进行各种操作,例如本体编码、本体文档化等等。其基本结构以及Agent

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.70771077);国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2008AA04Z106);上海市科委基金项目(No.08DZ1122301)。
作者简介:向阳(1962-),教授,博士生导师,研究方向为语义网、人工智能;张波(1978-),博士研究生,研究方向为语义网、本体论;韩婕(1978-),硕士,主要研究方向为本体论、Agent 技术。
收稿日期:2008-09-11 **修回日期**:2008-12-01

的组成情况如图 1 所示。下面分别介绍实施本体构建的各类 Agent 及其功能。

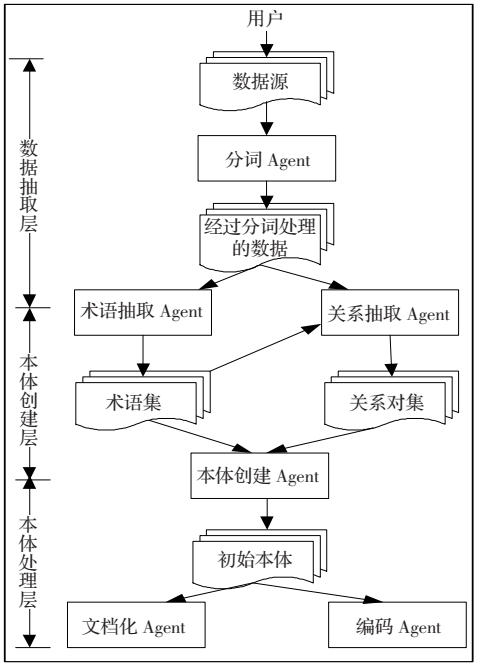

图 1 基于 Agent 的中文本体智能构建体系结构

1.1 分词 Agent

汉语分词是数据抽取层的第一步。词是最小的能够独立活动的有意义的语言成分[7],分词 Agent 的任务就是对经过导入的文本数据进行词法分析,分割出词片段,并为每个词片段标注词性。该 Agent 的输入、输出和处理分别如下:

(1)输入:经过预处理的文本 Txt,分词词典库 WordBase;

(2)输出:经过分词和词性标注处理的文本 SegD;

(3)处理:SegD=POS&Seg-tag(Txt,WordBase)。其中 POS(Part Of Speech)表示词性标注,Seg-tag 表示分词。

本文采用汉语词语分析框架的中科院分词系统(Institute of Computing Technology,Chinese Lexical Analysis System,ICTCLAS)[8]及其提供的分词词典库来实现实现分词 Agent。本文采用文献[9]提出的实现 ICTCLAS 系统 Java 调用的方法,将 ICTCLAS 的分词和词性标注功能融入到 JADE 的 Agent 框架中。ICTCLAS 支持当前广泛承认的分词和词类标准。同时本文采用北大标准[10]的一级标注来对自然语言文本进行分词和词性标注。

1.2 术语抽取 Agent

术语抽取 Agent 的任务是从经过分词和词性标注处理的文本中抽取出该文本领域相关的术语。术语抽取 Agent 的输入、输出和处理分别如下:

(1)输入:经过分词和词性标注处理的文本 SegD;

(2)输出:术语集 TermSet;

(3)处理:TermSet=ExtractTerm(SegD)。

其中 ExtractTerm()是抽取术语算法,具体在第 2 章中介绍。

1.3 关系抽取 Agent

关系抽取 Agent 的任务是从经过分词和词性标注处理的文本中抽取出该文本领域相关的术语间的关系。该模块的输入、输出和处理分别如下:

(1)输入:经过分词和词性标注处理的文本 SegD,术语抽取模块抽取出的术语集 TermSet,模式集 pattenSet;

(2)输出:关系集 RelationSet;

(3)处理:RelationSet=ExtractRelation(SegD,TermSet,patternSet)。

其中,关系集的每一个元素是一个三元组,包括一个术语对和术语对中两个术语间的关系名称。即有术语关系三元组(*pre*,*post*,*relationName*),(*pre*,*post*)是一个术语对,*pre* 和 *post* 分别指前项术语和后项术语,它们都是术语集中的术语;*relationName* 表示由前项术语指向后项术语的关系的名称。ExtractRelation()是抽取关系算法,在第 3 章中介绍。

1.4 本体创建 Agent

本体创建 Agent 的任务是根据术语抽取 Agent 和关系抽取 Agent 传送过来的术语集和关系集来创建本体。术语集的每一个元素是一个术语;关系集的每一个元素是一个三元组(*pre*,*post*,*relationName*)。

本文提出的本体创建的算法如下:

(1)定义术语集中每一个元素为一个类,类名为术语名;

(2)对关系集中关系名称为 is-a 的元素,定义它的 pre 所对应的类为其 post 所对应的类的子类;

(3)对关系集中关系名称不为 is-a 的元素,定义它为属性,它的 *relationName* 定义为属性名,它的 *pre* 定义为该属性的定义域(domain),它的 *post* 定义为该属性的值域(range),并定义它为它的 *pre* 所对应的类的属性;

(4)返回类集 classSet 和属性集 propertySet。

1.5 本体处理 Agent

本体处理 Agent 的任务是,根据本体应用的不同需求,对本体创建 Agent 创建的初始本体进行一些处理工作。本体编码主要是将初始本体转换成通用的格式,例如 OWL,以方便本体应用中的机器自动化操作;而本体文档化主要是将初始本体转换成一定格式自然语言文档,便于人们阅读、查询等。其中,OWL 本体编码对于本体创建 Agent 创建的初始本体中的数据进行的转换主要包括以下几个方面。

(1)对 classSet 中的每一个类:

```
<owl:Class rdf:ID="类名"/>
```

(2)对 classSet 中每一个有子类的类的子类:

```
<owl:Class rdf:ID="子类名">
<rdfs:subClassOf rdf:resource="# 类名"/>
</owl:Class>
```

(3)对 propertySet 中的每一个属性:

```
<owl:ObjectProperty rdf:ID="isComposedOf">
<rdfs:domain rdf:resource="#domain 类名"/>
<rdfs:range rdf:resource="#range 类名"/>
</owl:ObjectProperty>
```

2 术语抽取策略改进

在抽取术语时,将术语分成“常用词”与“合成词”两种类型,并根据它们各自的特点采用不同的策略进行抽取。现存的常用词抽取基本策略是将从分词标注后的文本中抽取出的词片段采用基于词频的统计学公式来计算每个词片段在领域文本中的权重,根据词片段的权重大小来筛选出领域术语。合成词抽取策略则是将抽取出的k-gram 词根据互信息和上下文依

赖等评价标准来筛选出该领域的合成词。这些术语抽取方法的时间复杂度和空间复杂度较大,无法客观地反映词片段的领域相关性,准确率较低,同时统计学计算,算法效率较低。

2.1 改进后的常用词抽取策略

针对现存的常用词抽取方法的缺点,提出了改进后的常用词抽取算法如下所示。

算法 1 改进的常用词抽取算法

输入:经过分词和词性标注处理的文本 SegD;

输出:领域常用术语集 keyComWdSet;

处理:

(1)seqList=storeArray(SegD); //存储词序列(词片段+词性标注)

(2)wdList=firstFilter(seqList,POSSet); //词性过滤

(3)weightdList=calculateWeight(wdList); //计算词片段权重

(4)keyComWdSet=domainFilter(weightdList,topN); //领域聚焦

(5)return keyComWdSet; //返回领域常用术语集

其中,(2)和(3)是本文对现有算法的改进。下面分别介绍:

词性过滤:在计算词片段的权重之前,先根据中文术语的语言特点对词片段进行第一轮筛选。对于中文术语而言,一般术语都不会是如下词性:连词(c)、副词(d)、叹词(e)、前接成份(h)、量词(p)、代词(r)、助词(u)和语气词(y),所以默认将这些词性的词过滤掉。考虑到各个领域的术语词性的不同特点,所以词性过滤集 POSSet 可根据具体情况由用户自行定义。

计算词片断权重:词片段权重计算需要计算的是某个词在文集中的重要性,一般主要根据词出现的次数、文件数量等方面计算得到[11]。词片段的权重定义 $TF\times IDF\times W$,其中 TF 是词片段在领域文本集中出现的次数;$IDF=\log(N/n)$,N 是文集中的文本总数,即领域文集中的文本数和参考文集中的文本数之和,n 是领域文集和参考文集中出现该词片段的文本总数,IDF 用于筛掉 TF 值大却非该领域的术语;W 为词片段的地位因子,某词片段地位因子 w 定义为该词片段在该文本中的位置所反映的其在特定领域中的重要性。

对于地位因子 w 值的设定,首先建立等级特征规则库。等级特征规则库用三元组的形式表示:(等级名称,w 值, 特征)。等级名称可以是该位置的名称;w 值即赋予该位置的权值;特征是该位置在文本中表现出来的特点,可以理解成一种查找处于该位置的词片段的搜索算法。处理领域文本的过程中,词片段的地位因子 W 初始值为 1, 然后考察该词片段在文本中的位置, 如果它出现在等级规则库中任一等级所对应的位置时,则该词片段的地位因子 w 等于原值乘上所对应等级的 w 值。

2.2 改进后的合成词抽取策略

针对现存的合成词抽取方法的缺点,提出了改进后的合成词抽取算法如下所示。

算法 2 改进后的合成词抽取算法

输入:经过分词和词性标注处理的文本 SegD,领域常用术语集 keyComWdSet;

输出:领域合成术语集 keyCmpdWdSet。

处理:

(1)cwdList=constructCmpd(SegD,maxGram,keyComWdSet);

//构造 k-gram 词

(2)MIList=calculateMI(cwdList,SegD);

//计算互信息

(3)LCDList=calculateLCD(cwdList,SegD);

//计算上下文依赖度

RCDList=calculateRCD(cwdList,SegD);

(4)weightList=calculateWeight(cwdList);

//计算权重

(5)keyCmpdWdSet=cmpFilter(MIList,LCDList,RCDList,cwdList,weightList);

//筛选

(6)return keyCmpdWdSet;

//返回领域合成术语集

其中,(1)和(4)是本文对现有算法的改进。

构造 k-gram 词:一反以往将常用词和合成词的抽取割裂开来的方式,而是利用已经抽取出的常用词来辅助合成词的抽取,选择权重较大的词片段作为 k-gram 词的一部分以构成 k-gram 词。k-gram 词的基本构成方法就是将从经过分词标注的文本中抽取出的前后相连的 k 个词片段进行组合。k-gram 词的构造元素由原来的所有词片段改为之前抽取出的领域常用词, 构造方式转变成以每一个已抽取出的常用词作为右边界,将之与文本中左邻近的 k-1 个词片段组成 k-gram 词(k=2,3,…,maxGram)。

计算权重: 如 2.1 节中所讨论,k-gram 词的权重定义为 $TF\times IDF\times W$,与常用词抽取的词片断权重公式相同。

3 关系抽取策略改进

在抽取概念间的关系时,一般将关系分成两大类:分类关系和非分类关系。现存的抽取关系方法主要以语言学方法和统计学方法为主。语言学方法受到已知有限知识的限制,无法获得准确的对应关系;而统计学方法的时间复杂度较大,且只能确定术语间的关系,无法确定关系的方向。因此准确率均不高。针对以上缺陷, 本文将语言学和统计学方法有效地结合在一起,充分利用了它们各自的优点互补各自的缺点,提出了关系抽取的混合策略:一方面利用模式抽取和统计学语义计算的方法进行关系抽取,一方面利用模式扩充算法[12]不断丰富模式库以扩大模式抽取方法的覆盖面。在关系抽取时,用经过用户筛选后的领域术语来辅助关系抽取,并结合现存方法和本文提出的新方法来对关系进行多次抽取,最后对抽取到的关系进行修剪;在首次抽取关系时,采用 Hearst 模式[13]和 LSA 语义相似度计算[14]的相结合、非分类模式和关联规则[15]的相结合的方法;在扩充抽取关系时,结合对等模式[16]和签名语义距离计算[12]的方法。改进后的关系抽取算法描述如下:

算法 3 改进后的关系抽取算法

(1)利用合成词特点进行抽取;

(2)结合 Hearst 分类模式和 LSA 进行抽取;

(3)利用对等模式和签名进行扩充抽取;

(4)修剪非直接的分类关系;

(5)利用非分类模式和关联规则进行抽取。

下面分别具体介绍对算法 3 的改进。

3.1 利用合成词特点进行抽取

利用合成词构成的特点,添加具有分类关系的术语对。主要过程如下:

(1)对合成术语集 keyCmpdWdSet 中的每一个 n-gram 合成术语进行解构得到构成它们的词片段,以右边界的词片段为右边界构成要素,提取出 n-1 个 k-gram 词,其中 k=1,2,…,

$n-1$。定义有 n 个常用词 $xn,x(n-1),\cdots,x2,x1$ 构成合成术语 $xn\ x(n-1)\cdots x2\ x1$,则以右边界的词片段为右边界构成要素提取出的 $n-2$ 个 k-gram 词为 $x2\ x1,x3\ x2\ x1,\cdots,x(n-1)\ x(n-2)\cdots x2\ x1,x1$。

(2)将这些 k-gram 词与常用术语集 keyComWdset 以及合成术语集 keyCmpdWdSet 中的每一个术语进行对照,筛选出构成领域合成术语的领域常用术语。

(3)将筛选出的领域常用术语和它们构成的领域合成术语构成具有分类关系的术语对,并这些术语对加入到集合 isaSet 中。

3.2 结合 Hearst 分类模式和 LSA 进行抽取

主要过程如下:

(1)取出 Hearst 模式库中的表示分类关系的模式集 isaPatternSet,并准备好术语抽取模块输出的合成术语集 keyCmpdWdSet 和常用术语集 keyComWdset;

(2)将术语集中的术语放入模式中构成含有正则表达式的字符串,并用字符串到领域文本中去匹配;

(3)对匹配到的句段进行处理,抽取出其中的词片段,对照术语及进行筛选,留下其中已是术语集中术语的词片段,并根据抽取术语对时所用的模式,确定可能具有分类关系的候选术语对,将之加入到集合 isaCSet 中;

(4)采用 LSA 法[14]计算每个术语对的相似度,并根据计算结果筛选出具有分类关系的术语对,并将之加入到集合 isaSet 中。

3.3 结合对等模式和签名进行扩充抽取

利用文献[16]中提出的表示对等关系的 Hearst 模式,抽取出具有对等关系的术语对,以此抽取出更多具有分类关系的术语对。本文通过抽取具有对等关系的术语对来间接抽取分类关系术语对,具体步骤如下:

(1)取出 Hearst 模式库中的表示对等关系的模式集 coordPatternSet,并准备好存有具有分类关系的术语对的集合 isaSet;

(2)将具有分类关系的术语对中的子类术语放入对等模式中构成含有正则表达式的字符串,并用字符串到领域文本中去匹配,再对匹配到的句段进行处理,提取出可能具有对等关系的候选术语对,并将之加入到集合 coordCSet 中;

(3)提取集合 coordCSet 中所有术语对中每个术语的签名以构造术语向量,计算每个术语对的语义距离,并根据计算结果筛选出具有对等关系的术语对,并将之加入到集合 coordSet 中;

(4)对于集合 coordSet 中的每一个术语对 (A,B),将术语对中所包含的术语 A 和 B 与所有具有分类关系的术语对 (C,D) 中的子类术语进行比对,如果找到相同的(例如,$A=C$),则将与该子类术语 $A(=C)$ 组成具有对等关系的术语对的另一个术语 B 与父类术语 D 组成术语对 (B,D),并将之加入到集合 isaCSetTwo 中;

(5)采用 LSA 法计算集合 isaCSetTwo 中每个术语对的相似度,并根据计算结果筛选出具有分类关系的术语对,并将之加入到集合 isaSet 中。

3.4 修剪不存在直接的分类关系的术语对

这一步主要将集合 isaSet 中的不存在直接分类关系的术语对修剪掉。修剪不存在直接分类关系的术语对的步骤如下:

(1)查找具有相同子类术语的术语对 (A,B) 和 (A,C);

(2)提取出这些术语对的父类术语 B 和 C,将它们两两组合成临时术语对 (B,C);

(3)在集合 isaSet 中搜索是否存在这样的临时术语对 (B,C) 和 (C,B),如存在,则将具有相同子类术语的两个术语对中拥有临时术语对的父类术语的术语对删除,即如果集合中 isaSet 存在临时术语对 (B,C),则删除 (A,C);如果 isaSet 中存在临时术语对 (C,B),则删除 (A,B)。

3.5 非分类模式和关联规则

本文用 Hearst 模式来抽取典型的非分类关系术语对的方式与用 Hearst 模式抽取具有分类关系术语对的方法雷同,此处不再赘述。关联规则抽取那些有着一定关系却无法定义其间关系名称的术语对,具体步骤如下:

(1)从术语抽取模块输出的合成术语集 keyCmpdWdSet 和常用术语集 keyComWdset 中抽取任意两个术语,进行两两配对组成若干组术语对,作为候选术语对加入到集合 noIsaCSet中。

(2)对照存有具有分类关系的术语对的集合 isaSet,对集合 noIsaCSet 中的术语对进行修剪,去掉其中同时存在于集合 isaSet 和 noIsaCSet 中的术语对。

(3)计算集合 noIsaCSet 中每个术语对的支持度和可信度,计算方法如式(1)和式(2)所示。

术语对支持度为:

$$support=f(xy)/f(all) \tag{1}$$

术语对可信度为:

$$confidence=f(xy)/f(x) \tag{2}$$

其中 $f(xy)$ 为同时出现术语 x 和 y 的句段的数目,$f(x)$ 为出现术语 x 的句段的数目,$f(all)$ 为领域文本中的句段数。

(4)当且仅当两种情况:①术语对的支持度 $support$ 大于等于最小支持度阈值(min_sup);②术语对的可信度 $confidence$ 大于等于最小可信度阈值(min_conf)时,才可判定该术语对中的两个术语之间的关联规则成立,即该术语对中的两个术语间存在非典型非分类关系。

4 模式扩充算法

模式扩充的主要思想是用已有的正确的关系术语对去学习模式以实现模式库模式的自动扩充。文献[12]中给出了一种扩充算法,但是这种算法对于已经存在关系的术语对而言,无法为其发现新的关系。本文借鉴文献[12],采用一种改进的模式扩充算法,具体如下:

(1)对于术语对的集合,若存在没有确定关系的术语对时,转步骤(2);而对于存在关系的术语对,转步骤(3);

(2)从不具备关系的术语集合中选取术语对 (A,B) 转步骤(4);

(3)对于存在关系的术语对,每隔固定的单位时间,触发新关系发现,以及已存在关系检测,转步骤(8);

(4)从来源于各种资源的文本集中提取包含术语 A 和 B 的句段,这些文本集来源可以是网络资源或用户提供的文本集;

(5)对步骤(4)提取的句段进行分析,抽取出 X 关系模式;

(6)用户对步骤(5)抽取出的模式进行检查和修正;

(7)若所抽取的关系模式不在模式库中,则将该模式加入模式库,退出;否则转步骤(9);

(8)从来源于各种资源的文本集中提取包含已存在关系术语对的文集,重新分析并抽取关系模式。若抽取到的关系模式未变,则不改动关系模式库,算法退出;若关系改变,则转(9);

(9)将新关系和已存在的关系模式进行冲突检测,若无冲突,则将新关系加入关系模式库,算法退出;若有冲突,则消除冲突后加入模式库,算法退出。

5　实验分析

利用本文提出的中文本体构建算法,进行计算机领域知识本体的构建。抽取对象来源为一组由 300 篇计算机领域学术论文组成的文本集。共构建了 6 个计算机领域本体,表 1 中给出了抽取结果。共抽取术语 794 个,其中可在 2 个以上的本体中出现的重复术语 175 个;抽取关系 1 074 个,可在 2 个以上本体中出现的重复关系 297 个。表 1 中的错误率是指术语与关系抽取中出现的不准确术语和关系的总错误率。

表 1　本体术语和关系抽取情况

	计算机工程本体	计算机应用本体	计算机系统结构本体	计算机软件本体	计算机基础本体	人工智能本体
术语	181	223	163	127	206	231
关系	293	297	242	257	284	309
错误率/(%)	23.87	17.16	14.83	19.67	20.30	22.83

同时,利用未改进的算法和改进的算法进行了对比实验。实验结果如图 2 中所表示。图 2 中分别给出了 6 个本体利用两种抽取算法所获得术语和关系的总准确度。对比中可以看出,除了第 4 和第 6 个本体准确率相近以外,其余 4 个本体改进算法明显准确率高于未改进算法。总体而言,准确率保持在 18% 左右,这是现今自动抽取术语和关系方法中比较能够接受的。

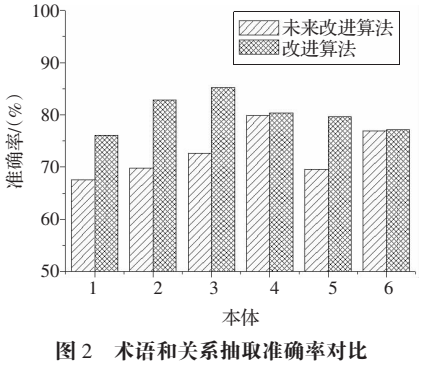

图 2　术语和关系抽取准确率对比

6　结束语

提出了一个基于 Agent 的中文本体自动构建系统,包括三层结构(数据抽取层、本体创建层和本体处理层)和实现各层的功能的多个 Agent 的详细设计,重点阐述了术语抽取 Agent 和关系抽取 Agent 所使用的策略,并对现有算法进行创新改进,同时使用 Agent 技术将改进后的方法综合运用起来,以实现一个基于中文数据源的中文本体自动构建系统。今后的研究将集中在术语和关系准确率的提高,以及智能构建本体的过程上。

参考文献:

[1] Maedche A,Staab S.Ontology learning for the semantic Web[J].IEEE Intelligent Systems,2001,16(2):72-79.
[2] 邓志鸿,唐世渭,张铭.Ontology 研究综述[J].北京大学学报,2002,38(5).
[3] 陈刚,陆汝钤,金芝.基于领域知识重用的虚拟领域本体构造[J].软件学报,2003,14(3):350-355.
[4] 王红滨,刘大昕,王念滨,等.基于非结构化数据的本体学习研究[J].计算机工程与应用,2008,44(26):30-33.
[5] 杨圣洪,贾焰.非成熟领域的本体构建方法[J].计算机工程与应用,2008,44(24):153-155.
[6] Wooldridge M,Jennings N R.Intelligent agents:Theory and practice[J].The Knowledge Engineering Review,1995,10(2):115-152.
[7] 朱德熙.语法讲义[M].北京:商务印书馆,1982.
[8] 刘群,张华平,俞鸿魁,等.基于层叠隐马模型的汉语词法分析[J].计算机研究与发展,2004,41(8):1421-1429.
[9] 夏天,樊孝忠,刘林.利用 JNI 实现 ICTCLAS 系统的 Java 调用[J].计算机应用,2004,24:177-182.
[10] 孙斌,朱学锋,段慧明,等.北京大学现代汉语语料库基本加工规范[J].中文信息学报,2002,16(5):49-64.
[11] Salton G,Buckley C.Term-weighting approaches in automatic text retrieval[J].Information Processing & Management,1988,24(5):513-523.
[12] 方卫东,袁华,刘卫红.基于 Web 挖掘的领域本体自动学习[J].清华大学学报:自然科学版,2005,45(1):1729-1733.
[13] Hearst M A.WordNet:An electronic lexical database[M].Cambridge MA:MIT Press,1998.
[14] Cederberg S,Widdows D.Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction[C]//Conference on Natural Language Learning(CoNNL),2003.
[15] Nakaya N,Kurematsu M,Yamaguchi T.A domain ontology development environment using a MRD and text corpus[C]//Fifth Joint Conference on Knowledge-based Software Engineering Frontiers in Artificial Intelligence and Applications.[S.l.]:IOS Press,2002,80:242-251.
[16] Widdows D,Dorow B.A graph model for unsupervised lexical acquisition[C]//19th International Conference on Computational Linguistics,Taipei,Taiwan,August 2002:1093-1099.

(上接 101 页)

[5] Karp B,Kung H T.Greedy perimeter stateless routing for wireless networks[J].Mobicom,2000:243.
[6] Seet B C,Liu G,Lee B S,et al.A-STA:a mobile ad hoc routing strategy for metropolis vehicular communications[J].Networking,2004.
[7] Su W,Gerla M.IPv6 flow handoff in ad-hoc wireless networks using mobility prediction[C]//Proceedings of IEEE Global Communications Conference,Riode Janeiro,Brazil,1999:271.
[8] ETSI,Universal Mobile Telecommunication System.Selection procedures for the choice of radio transmission technologies of the UMTS[S/OL].(1998-04).http://www.3gpp.org/ftp/Specs/html.
[9] Perkins C E,Royer E.RFC3561 Ad hoc on-demand distance vector(AODV) routing[S].2003-07.