领域本体术语抽取研究*

汤　青[1]　吕学强[1,2]　李　卓[1]　施水才[1,2]
[1]（北京信息科技大学网络文化与数字传播北京市重点实验室　北京 100101）
[2]（北京拓尔思信息技术股份有限公司　北京 100101）

摘要：【目的】尽可能多地抽取多字词本体术语，以保证本体构建的质量。**【方法】**提出基于部件扩展的本体术语抽取方法。利用部件的领域聚合性和词性特征，采用领域词频比较的方法抽取部件；考虑术语长度、术语词性构成以及术语内部结合度等因素，设计合理的扩展规则对部件扩展以形成候选术语；利用上下文关联信息、语境信息从候选术语集中筛选出本体术语。**【结果】**利用该方法在 IT 领域实验数据集上进行测试，实验结果准确率为 83.5%，召回率为 87%，准确率相比 Baseline 方法要高出 2.5 个百分点。**【局限】**部件抽取方法需要借助于平衡语料库，部件的质量直接影响术语抽取效果。**【结论】**实验结果表明该方法是有效的，对本体学习、本体构建具有积极意义。

关键词：本体术语　术语抽取　术语部件　部件扩展
分类号：TP391.1

1　引　言

随着大数据时代的到来，如何将网络上的海量数据形成一个互相关联的网络以实现信息的高速运作的问题，推动着基于本体的知识检索、知识工程等领域的快速发展。但本体在构建和维护上的困难制约着这些依赖于本体的相关领域的发展。本体是概念模型的明确规范说明[1]，是概念间的关系模型，因而概念是本体中最重要的组成部分之一。而术语在国家标准规范 GB/T 19101-2003《建立术语语料库的一般原则与方法》中被定义为“特定专业领域中一般概念的词语指称”[2]，它作为概念的一种描述，可以用于表示概念的实例。故本体术语抽取成为本体构建的首要工作，对本体学习以及基于本体的应用技术的发展具有重要意义。

本文中“本体”指的是领域本体，所以，本体术语不是一般意义上的术语，而是领域内的核心术语。本体术语具有很强的领域性，主要以极具领域内涵的多词型术语为主。本文从术语在语料上的分布特征、术语形成的方式等方面进行分析，提出了基于部件扩展的本体术语抽取方法。

2　相关工作

本体术语抽取是本体构建的基础，也是知识抽取等信息技术中的关键步骤。目前，本体术语抽取研究采用的方法有基于规则的方法[3]、基于统计的方法[4]、基于规则与统计相结合的混合方法[5]。其中，混合方法是当前本体术语抽取的主流方法。

2010 年，Yang 等[6]提出一种不依赖领域特征的术语抽取方法，根据边界分隔符抽取候选术语，借助领域相关句与领域术语之间的相互强化关系抽取领域术语，但边界分隔符的准确获取本身就是一个难点。2011 年，章成志[7]提出多层术语度的一体化术语抽取方法，并提出了句子术语度的概念，将术语所在句子的所有词语均作为训练特征，用 CRF 识别术语，但该方法依赖于大量训练数据。2012 年，Lee 等[8]提出了一种不

收稿日期：2013-09-27
收修改稿日期：2013-11-22
*本文系国家自然科学基金项目“基于本体的专利自动标引研究”（项目编号：61271304）和北京市教委科技发展计划重点项目暨北京市自然科学基金 B 类重点项目“面向领域的互联网多模态信息精准搜索方法研究”（项目编号：KZ201311232037）的研究成果之一。

依赖词典、以规则作为特征的 SVM 分类抽取术语的方法，但召回率偏低。2012 年，王卫民等[9]提出了一种半监督的基于种子迭代扩充的专业术语识别方法，该方法仅利用少量训练样本通过方法自身的迭代来增加训练样本，同时生成新的模型，将迭代生成的最终模型作为专业术语识别模型，但需部分人工参与。

上述研究较好地利用了语法规则、统计方法两者的优点，大大提高了术语抽取的准确率，但是由于这些方法缺乏对本体术语构成方式的分析，使得部分极具领域内涵的较长本体术语没有被召回，不太适合本体术语的抽取。

针对现有方法抽取本体术语的不足，将规则与统计相结合，提出了基于部件扩展的本体术语抽取方法，综合采用了领域词频比较的部件抽取方法、基于规则扩展的候选术语抽取方法、基于上下文信息的本体术语筛选方法，抽取本体术语。与前人研究相比，该方法以平衡语料库作为支撑，利用部件的领域聚合性特征获得部件；对部件进行单、双向扩展，获得几乎所有较长术语，提高了本体术语抽取的召回率。

3 基于部件扩展的本体术语抽取方法

本体术语一般分为单词型术语和多词型术语，吴云芳等[10]对信息科学领域的术语进行了研究，发现26%的为单词型术语，74%的为多词型术语。单词型术语由单个词组成，如“相机”、“硬盘”等；而多词型术语一般由多个词语通过复合、派生、拼凑形成，如“云计算存储”、“固态硬盘”。多词型术语构成成分中的有些词生成术语的能力较强，GB/T 19102-2003《术语部件库的信息描述规范》中将这类生成术语的能力较强的词或者词缀称作“术语部件”[11]，其定义如下：

特定领域中结合紧密、生成能力强、使用稳定的语言片段称为术语部件，简称部件，即用来生成多词术语的词或者词缀。

某一特定领域，设 T 是该领域中一个多词型术语，对任意词或者词缀 c，其构成的术语集合为 $seT=\{T|T=c_1\cdots c_{m-1}cc_{m+1}\cdots c_n\}$，集合 setT 的元素个数为 Num；若 $Num\geq\alpha>0$，则 c 称为部件，此处，阈值α表示部件生成术语的个数的下限值。如：“新浪微博、腾讯微博、搜狐微博”、“网易微博”、“天涯微博”，“微博”即为部件，此处，部件生成术语的个数是 5。

因此，可将术语看作是由部件按一定逻辑关系与其他词缀组合成的一个整体。本文提出基于部件的本体术语抽取方法，如图 1 所示，主要包括部件抽取、候选术语抽取、本体术语抽取三个部分。其中在部件抽取部分利用部件的出现频率特征和词性特征，候选术语抽取部分通过分析术语的形成方式和结构特征，提出了基于部件扩展的候选术语抽取方法；本体术语抽取阶段，综合考虑词频、语义和上下文关联信息。

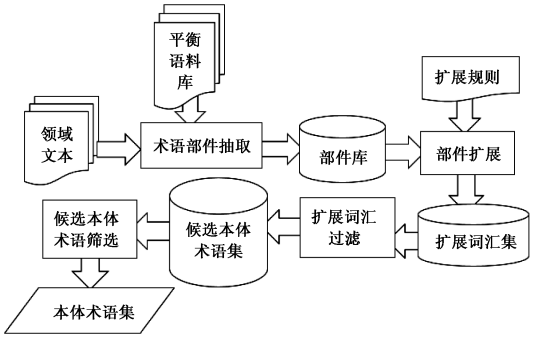

图 1 本体术语抽取方法

3.1 基于统计的部件抽取

根据部件的定义可知，部件具有一定的领域聚合性，也即领域部件在该领域语料中出现的频率要高于在其他领域语料中的出现频率。吴云芳等[10]曾提出部件在领域语料中的两个特征，即部件的出现频率特征和部件词性特征。词性特征指部件一般为名词、动词、形容词等实词词性的词，因为部件能生成术语决定着部件通常由实词充当。利用部件的以上两个特征抽取部件：

（1）部件和术语具有领域聚合性特征：在某一领域中词频较高或只出现在某个领域中，而在不相关领域中词频相对较低。因而，采用语料库中词频统计量相比较的方法抽取部件。即：将词语在平衡语料库和领域语料库中的词频相比较，把在领域词表中词频排序较高而在平衡语料库中其他领域词表中词频排序较低的词语，初步选为部件。此处，将本领域词频排序前 5 000 的高频词作为入选部件的候选集，针对词频比较抽取部件的方法，设计公式如下：

$$C_i(w)=\log(F(w))\times P_i(w) \quad (1)$$

$$\begin{cases} |\overline{A_i}|>\lambda \\ R=\{A_1\cap A_2\ldots\cap A_i\cap\ldots\cap A_n\} \end{cases} \quad (2)$$

其中，i 表示平衡语料库中第 i 领域语料，$i \in [1,n]$；w 为本领域词语，F(w)表示 w 在本领域语料中的词频，$P_i(w)$ 表示 w 在平衡语料库第 i 领域语料上的正向词频位序，$C_i(w)$ 为领域词语和平衡语料库中第 i 个领域比较的结果，λ为阈值；A_i 为领域词频按照比较结果 $C_i(w)$升序前λ的词语构成的集合，R 为术语部件集合，n 为平衡语料库所含领域语料的总数。

（2）上述所选部件中，存在少量满足部件聚合性特征的无效词串，比如各种虚词类词性的词。利用部件在词性上的明显特征，将不能成为部件的词语统计出来，并观察发现诸如时间词、量词、方位词等虚词类词性的词，不能成为领域部件，如表 1 所示。故根据词性特征筛除无效词串，提取部件。

表 1　垃圾词串的词性

时间词（t）	数词（m）	副词（d）	动词性惯用语（vl）	方位词（f）	代词（r）
上月	百万	刚刚	也就是说	之外	每个
周二	众多	飞速	成千上万	后面	本文
周三	批量	远远	众所周知	背后	何人
今年	万部	共同	不为人知	上面	这家
中午	数百万	快速	尘埃落定	下面	此文

3.2　基于部件扩展的候选术语抽取

冯志伟[12]在 2010 年描述术语形成的经济律，即由少量的单词构成大量的术语。李萍等[13]发现 IT 领域术语主要由部件通过复合法、派生法、转化、拼缀等几种方式构成。即多词型术语可以看成是由部件和其他词语结合形成。本文利用以上思想，对部件库中的部件进行扩展，形成候选术语。

要通过部件扩展而获得候选术语，首先分析术语在结构上的特征，如表 2 所示：

表 2　术语的结构特点

术语内部结构特点		举例
词长	2-8 个中文字符	磁盘、移动云计算
词性	至少含有一个名词性成分的词语（n，vn，an，ng） 首词的词性一般不能是量词（q）、介词（p）和区别词（b） 术语中没有连词（c，cc）、代词（ry）、语气词（rz）	编程/vn 语言/n [vn+n] 片式/n 服务器/n [n+n] 阿里巴巴/nz [nz] 开/v 源/ng 软件/n [v+ng+n]
边界	术语末尾不是停用词	多媒体、操作系统

部件扩展是以部件所在句子为介质，以部件为中心，以词语为单位向左或向右单向扩展、向左向右双向扩展。扩展规则如下：

规则 1：部件向左扩展的第一个词语不能为介词（p）、量词（q）或区别词（b）；部件向右扩展的最后一个词必须为动词性或名词性成分的词，且不能为停用词。

规则 2：部件扩展形成的词语至少有一个动词、名词或名词性成分的词。

规则 3：部件扩展形成的词语长度取值区间为[2,8]。

采用以上规则对部件扩展后，获得部件扩展集。在部件扩展集中，只有内部字串结合较为紧密的多词短语才可能是术语。互信息可以用来表示一个字串的内部结合强度，而本体术语是领域内出现次数相对较多的核心术语，本文对部件扩展集中出现次数大于 20 次的扩展词汇计算互信息值，获得候选术语。互信息值计算公式[14]如下：

$$MI(c) = \log\left(\frac{f(c)}{f(a)\times f(b)}\right) \qquad (3)$$

其中，c 表示扩展形成的多词短语，a 表示部件，b 表示被部件结合的串，f(a)、f(b)、f(c)指 a、b、c 分别独立出现的词频。如果 c 中字串结合紧密，f(c)与 f(a)或者与 f(b)的值相差不大，此时，计算的 c 的互信息较大；反之，f(a)或 f(b)的值远大于 f(c)，这样计算的 c 的互信息就较小。

把由同一个部件生成的所有扩展词汇看成一个集合，对集合中的扩展词汇 c 按照互信息值大小进行排序，排序靠前的词语保留，保留后的词语构成候选本体术语集。

3.3　多策略候选本体术语筛选

为了召回更多的术语，允许部件扩展规则的设定相对宽松，因而扩展过程中不可避免地会产生无效词串，仅依靠互信息值也不能完全将无效词串过滤掉。例如：部件“存储”在句子“大量/存储/寻址/需要校对”中的扩展结果为“大量存储”、“存储寻址”和“大量存储寻址”，其中下划线短语不是术语。为了提高术语抽取的准确率，需要将无效词串筛除。

根据术语扩展形成过程分析，发现存在两种无效字符串：第一种是由正确的部件错误扩展而来；第二

种是由错误的部件扩展而来。通过分析总结发现：第一种无效字符串通常是一些与某个术语具有公共字串的词串，即是某个术语的子串或母串，比如“如雅虎”、“计算机科学与”，本文将这类词串称为缺陷术语；第二种无效字符串通常是一些与术语具有相同的结构特点，但不是本领域术语的词，包括通用词汇和其他领域词汇，比如“信息业”、“基础理论”，本文将这类词串称为弱领域性术语。

（1）基于上下文关联信息的缺陷术语过滤

在领域文档集中，候选术语之间通过共同的上下文建立关联关系，而缺陷术语被关联的可能性相对较小。在领域语料中的某个候选术语，被其他候选术语关联越多，说明它越重要，是术语的可能性就越大。因而，利用基于候选术语之间的关联关系，采用 PageRank 算法[15]计算候选术语的重要度 PR(A)，重要度指的是对术语在领域语料中重要位置的度量。根据重要度值的大小，筛选缺陷术语。

将候选术语所在句子作为其上下文，定义共现在同一个上下文的两个候选术语具有关联关系。据此建立候选术语间关联信息的关系图。借鉴 PageRank 思想，术语重要度计算方法如下：

$$PR(A)_{doc} = (1-d) + d \times \sum_{i=1}^{i=n} \frac{PR(B_i)_{doc}}{C(B_i)_{doc}} \quad (4)$$

$$PR(A) = \sum_{i}^{N} PR(A)_{doci} \quad (5)$$

其中，A，B_i 均为候选术语，且 A，B_i 共现在同一个上下文。$PR(A)_{doc}$ 表示候选术语 A 在任意一篇文档中的重要性，$PR(B_i)_{doc}$ 初始值都为 1，$C(B_i)_{doc}$ 表示任意一篇文档中 B_i 指向的候选术语的总个数，d 为调节因子，便于程序正常结束，取值范围为[0,1]，本文将 PR(A)值大于设定阈值 α 的候选术语保留，将其余候选术语从候选术语集中删除。

（2）基于上下文语境的弱领域性术语过滤

弱领域性术语主要是由于错误的部件扩展形成，可能是通用词或者和领域关系较小的词。这里以相似度来量化术语间的关系，当一个候选术语和大部分候选术语都不相似，说明它是弱领域性术语。Resnik[16]认为术语的相似度取决于它们语境的相似程度，术语相似性可以通过它们之间共有的信息量来衡量，共有信息量越高，则相似性越高。本文基于该思想，以候选术语共有信息量表征语境，计算术语的语境相似度，筛选弱领域性术语。

本文选择句子为语境单元，取出候选术语所在句子中的实词，构建候选术语的语境向量，此处实词为名词或名词性短语、动词或动名词短语。

设 t 表示候选术语，则其语境向量可以表示为：$s(t)=<(w_1,f_1),(w_2,f_2)\cdots(w_i,f_i)\cdots(w_n,f_n)>$，$w_i$ 表示实词词汇，f_i 为当前语境下实词 w_i 和候选术语 t 的共现词频之和，$i \in [1,n]$。例如：{编程语言：安全编程语言开放平台//面向对象的编程语言开放平台}，则候选术语“编程语言”的上下文语义向量可表示为：

s(编程语言)=<(安全,1),(开放平台,2),(面向对象,1)>

本文用余弦相似度方法计算任意两个候选术语的语义相似度，计算公式[17]如下：

$$sim(s(t_i),s(t_j)) = \frac{\sum_{k=1}^{n} f_{ik} \times f_{jk}}{\sqrt{\sum_{k=1}^{n} f_{ik}^{\ 2}} \sqrt{\sum_{k=1}^{n} f_{jk}^{\ 2}}} \quad (6)$$

其中，t_i 和 t_j 均表示候选术语，$s(t_i),s(t_j)$分别表示 t_i、t_j 所对应的语义向量，f_{ik} 和 f_{jk} 分别表示 t_i，t_j 对应语义向量中第 k 个特征值，$sim(s(t_i),s(t_j))$表示两个候选术语之间的语义相似度值。对任意候选术语，计算该候选术语与其他所有候选术语间的相似度值，对这些相似度值排序，设定阈值 β，如果 TopN 的相似度值均大于 β，则将该候选术语保留，否则将被删除，最后保留的集合即为抽取的本体术语结果。

4 实验结果及分析

4.1 语料介绍

（1）平衡语料库

以搜狗实验室官方网站提供的 2012 年分类语料①为来源，选取其中交通（Travel）、文化（Cul）、学习（Learning）、女性（Women）和汽车（Auto）5 个领域的分类数据构建而成，每个领域的数据均为两万篇以上。

（2）测试数据

以开发者技术社区 CSDN 上的网页数据为来源，选取业界、移动开发、软件研发和云计算共 4 个频道在 2012 年 10 月–12 月期间发表的 61 075 篇文档数据

① http://www.sogou.com/labs/dl/cs.html

作为测试数据。

4.2　评价指标

对部件抽取和本体术语抽取两个部分的结果进行评价，部件抽取采用抽取的正确结果数量和准确率进行评价，本体术语抽取采用准确率和召回率评价。设抽取的正确部件数记为 A，抽取的非正确部件数记为 B，抽取的正确本体术语数记为 C，抽取的非正确本体术语数记为 D，部件扩展结果中正确的本体术语数记为 E。

部件抽取的准确率为：

$$P(\text{部件})=\frac{A}{A+B} \qquad (7)$$

本体术语抽取的准确率为：

$$P(\text{术语})=\frac{C}{C+D} \qquad (8)$$

本体术语抽取的召回率：

$$R(\text{术语})=\frac{C}{E} \qquad (9)$$

4.3　结果与分析

（1）部件抽取

采用在 3.1 节方法抽取部件，抽取结果与方法中设定阈值λ有关，阈值λ表示领域词汇在平衡语料库各领域的比较结果 A_i 中，参与求交集的词汇个数。阈值对部件抽取结果的影响如表 3 所示：

表 3　实验结果-部件抽取评价

阈值λ	所抽部件数	正确部件数	P（部件）
1 000	444	333	0.75
1 200	530	391	0.738
1 800	797	590	0.74
2 500	1 139	831	0.73
3 500	1 598	1 107	0.653

部件抽取的理论依据是：在本领域词频较高，而在其他领域词频较低（排序位序较大）的词作为选为部件的基本条件。词语在集合 A_i 中排序越靠前，表明该词语在领域语料中的词频较高而在平衡语料库中第 i 个领域语料中的词频较低，这样的词语越有可能是部件。经过和平衡语料库中 5 个领域比较后的 5 个 A_i 中求交集即为部件抽取结果，随着λ取值越大，准确率整体呈下降趋势，这是因为 A_i 排序相对靠后的词语，可能在其他领域词频较高，不符合部件的基本条件。

本体术语是经过部件扩展获得，要求部件抽取保证正确结果数以及准确率，部件过少或者抽取准确率较低均会对本体术语抽取结果有不良影响。故部件抽取过程中，阈值的选择很重要。由表 3 看出，阈值λ取 2 500 的对应结果准确率为 0.73，抽取的正确部件为 831 个，满足下一步用于扩展的部件在数量和质量上的要求，故选择该阈值对应结果用于下一步部件扩展。部件抽取结果如下所示：

模式 开发 运算 安装 并行 计算 设备 木马 系统 百科 编译器 量子 存储器 内核 芯片 游戏 协议 网软件 索引 贴吧 腾讯 粒度 磁盘 接口

其中的结果基本符合部件的定义：即结果可以是术语，如“接口”和“磁盘”；结果均能产生多个术语，如：“开发—>[软件开发、开发程序、系统开发]”；“分布式—>[分布式系统、分布式数据库、分布式运行]”。这为用部件扩展抽取术语提供了事实依据。

（2）部件扩展

采用 3.2 节方法中对抽取的部件进行扩展，从术语长度、术语词性构成、术语内部结合紧密程度三个方面考虑，设定合理的扩展规则，向左、向右以及同时向左向右以一个词语为窗口单位进行扩展。扩展结果示例如下所示：

光影魔术手百度 云计算 信息业 基础理论
光影魔术手百度贴吧 百度贴吧 网格计算互联
无线鼠标 北桥芯片 固态硬盘 磁盘 移动设备

可知，通过部件扩展生成多字词短语，比如，以“百度”作为部件进行扩展的结果有“光影魔术手百度贴吧”、“百度贴吧”、“光影魔术手百度”，前两个为多字词本体术语，说明以部件扩展提取本体术语是有效的，而“光影魔术手百度”则为无效词串。

扩展结果中无效词串来源有：

①部件抽取结果中存在错误，任何一个错误的部件进行扩展均可能产生三类错误的结果；

②部件扩展窗口是以词语为单位，分词误差产生的碎片引起部件扩展结果错误，如“新浪微博”中“微”为分词碎片，“新浪”扩展为“新浪微”；

③扩展规则中关于术语长度、词性等要求的设定使得满足条件的扩展结果都被保留下来，包括那些仅仅在形式上符合术语标准的无效词串。

总之，经过部件扩展引起的无效词串不可避免。

（3）本体术语抽取

采用 3.3 节中方法从扩展结果中筛选本体术语。

对扩展结果中的无效词串分析，根据结构特点将其分为缺陷术语和弱领域性术语。第一个阶段过滤缺陷术语，设定过滤阈值α为6.5。第二个阶段过滤弱领域性术语，设定过滤阈值β为0.0015。经过两次过滤，剩余共2 750个词即为本体术语抽取结果。为了验证该方法的有效性，将何琳[18]提出的基于术语分布度、术语活跃度、术语主题度的多策略领域本体术语抽取方法作为Baseline方法，在本文的实验数据集上进行实验测试，采用准确率、召回率评价，评价结果如表4所示：

表4　实验结果–本体术语抽取评价

评价结果	本文方法			对比方法
	部件扩展	+关联信息筛选	+语境信息筛选	多策略融合方法
P（术语）	0.76	0.80	0.835	0.81
R（术语）	1.0	0.92	0.87	0.83

从表4可以看出，部件扩展结果获得0.76的准确率，证明了部件抽取的合理性和扩展规则设计的合理性。引入两个本体术语筛选方法，均使得召回率下降，准确率提高，其中，召回率计算是以部件扩展结果作为基础。通过过滤无效词串筛选出本体术语，两次过滤均使得召回率下降，因为上下文数据稀疏问题使得关联信息和语境信息无法将部分本体术语与无效词串区别开，导致部分本体术语被过滤。两次过滤均使得准确率提高，说明用上下文关联信息和上下文语境信息筛选出本体术语是合理的。与Baseline方法相比，准确率高出2.5个百分点，主要是由于Baseline方法中N-Gram模型获得候选术语产生的垃圾较多，增加了本体术语抽取的难度，而本文方法是立足于部件扩展获得候选术语，降低了本体术语抽取的难度。本体术语抽取结果示例如下所示：

超级计算机 图形图像处理　硬盘 数据库服务器 静态存储器 加速软件 软件工程 海量数据挖掘 分布式处理系统 平板电脑 串行数据接口

可知，单词型本体术语和多词型本体术语都能被准确抽取，主要由于部件作为候选术语保留使得单词型术语被召回，部件扩展结果作为候选术语使得多字词术语被召回；候选术语进一步筛选，获得本体术语。

5　结　语

本文针对术语生成方式和结构特点，提出了一种基于部件扩展的本体术语抽取方法。采用领域词频比较的方法抽取部件，对部件扩展形成的候选术语进行分析，提出了基于关联信息和基于语境信息的本体术语筛选方法，实验取得较好的结果。本体术语抽取为本体关系挖掘提供基础支持，抽取的术语为本体构建提供来源参考。下一步工作是利用抽取的本体术语，辅助抽取本体术语间的关系。

参考文献：

[1] Gruber T R. A Translation Approach to Portable Ontology Specifications [J]. Knowledge Acquisition，1993，5（2）：199-220.

[2] 中国国家标准化管理委员会.GB/T 19101-2003，建立术语语料库的一般原则与方法[S]. 北京：中国标准出版社，2003：1-4.（Standardization Administration of the People's Republic of China. GB/T 19101-2003，General Principles and Methods of Establishing Terminology Corpus[S]. Beijing：China Zhijian Publishing House，2003：1-4.）

[3] Chambers N，Jurafsky D. Template-based Information Extraction without the Templates [C]. In：Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics（HLT'11）. Stroudsburg：Association for Computational Linguistics，2011：976-986.

[4] 韦小丽，孙涌，张书奎，等.基于最大熵模型的本体概念获取方法研究 [J]. 计算机工程，2009，35（24）：114-116.（Wei Xiaoli，Sun Yong，Zhang Shukui，et al. Ontological Concept Extraction Method Based on Maximum Entropy Model [J]. Computer Engineering，2009，35(24)：114-116.）

[5] 游宏梁，张巍，沈钧毅，等.一种基于加权投票的术语自动识别方法[J]. 中文信息学报，2011，25（3）：9-16.（You Hongliang，Zhang Wei，Shen Junyi，et al. A Weighted Voting Based Automatic Term Recognition Method[J]. Journal of Chinese Information Processing，2011，25（3）：9-16.）

[6] Yang Y，Lu Q，Zhao T.A Delimiter-based General Approach for Chinese Term Extraction [J]. Journal of the American Society for Information Science and Technology，2010，61（1）：111-125.

[7] 章成志.基于多层术语度的一体化术语抽取研究[J]. 情报学报，2011，30（3）：275-285.（Zhang Chengzhi.Using Integration Strategy and Multi-level Termhood to Extract Terminology [J]. Journal of the China Society for Scientific

and Technical Information，2011，30（3）：275-285.）

[8] Lee C，Huang C，Tang K，et al. Iterative Machine-Learning Chinese Term Extraction [C]. In：Proceedings of the 14th International Conference on Asia-Pacific Digital Libraries. 2012：309-312.

[9] 王卫民，贺冬春，符建辉.基于种子扩充的专业术语识别方法研究[J]. 计算机应用研究，2012，29（11）：4105-4107.（Wang Weimin，He Dongchun，Fu Jianhui. Research of Professional Term Identification Method Based on Seed Expansion[J]. Application Research of Computers，2012，29（11）：4105-4107.）

[10] 吴云芳，穗志方，邱利坤，等.信息科学与技术领域术语部件描述[J]. 语言文字应用，2003（4）：34-39.（Wu Yunfang，Sui Zhifang，Qiu Likun，et al. The Approaches and Strategies to Describe the Term Component in Information Science and Technology [J]. Applied Linguistics，2003（4）：34-39.）

[11] 中国国家标准化管理委员会.GB/T 19102-2003，术语部件库的信息描述规范[S]. 北京：中国标准出版社，2003：1-4.（Standardization Administration of the People's Republic of China GB/T 19101-2003，Specification of Description of Term Component Database [S]. Beijing：China Zhijian Publishing House，2003：1-4.）

[12] 冯志伟. 术语形成的经济律——FEL 公式[J]. 中国科技术语，2010，12（2）：9-15.（Feng Zhiwei.Economic Law of Term Formation——FEL Formula [J]. China Terminology，2010，12（2）：9-15.）

[13] 李萍，黄崇岭. IT 领域的专业术语构词特点及功能意义[J]. 桂林电子工业学院学报，2004，24（2）：48-51.（Li Ping，Huang Chongling. The Morphological Formation and Functional Significance of Technical Term in IT Field [J]. Journal of Guilin University of Electronic Technology，2004，24（2）：48-51.）

[14] 陈士超，郁滨. 面向术语抽取的双阈值互信息过滤方法[J]. 计算机应用，2011，31（4）：1070-1073.（Chen Shichao，Yu Bin. Method of Mutual Information Filtration with Dual-threshold for Term Extraction[J]. Journal of Computer Applications，2011，31（4）：1070-1073.）

[15] Page L，Brin S，Motwani R，et al. The PageRank Citation Ranking：Bringing Order to the Web[R]. Stanford InfoLab，1999.

[16] Resnik P. Using Information Content to Evaluate Semantic Similarity [C]. In：Proceedings of the 14th International Joint Conference on Artificial Intelligence（IJCAI'95）. San Francisco：Morgan Kaufmann Publishers Inc.，1995：448-453.

[17] Tan P，Steinbach M，Kumar V. Introduction to Data Mining [M]. Addison-Wesley，2005.

[18] 何琳.基于多策略的领域本体术语抽取研究[J]. 情报学报，2012，31（8）：798-804.（He Lin. Domain Ontology Terminology Extraction Based on Integrated Strategy Method [J]. Journal of the China Society for Scientific and Technical Information，2012，31（8）：798-804.）

作者贡献声明：

汤青：提出研究思路，设计研究方案和完成实验，论文的起草、撰写；

吕学强，李卓：负责设计论文框架和论文的修改；

施水才：提出研究课题，负责论文的修订工作。

（通讯作者：汤青 E-mail：tangqing20062008@126.com）

Research on Domain Ontology Term Extraction

Tang Qing[1]　Lv Xueqiang[1,2]　Li Zhuo[1]　Shi Shuicai[1,2]

[1]（Beijing Key Laboratory of Internet Culture and Digital Dissemination Research，Beijing Information Science and Technology University，Beijing 100101，China）

[2]（Beijing TRS Information Technology Co.Ltd.，Beijing 100101，China）

Abstract：[**Objective**] Ontology terms are extracted as more as possible for the quality of Ontology construction. [**Methods**] This paper proposes an Ontology term extraction method based on term component extension. It uses the polymerization characteristics and POS features of the terms，extracts term components by word frequency comparison approach. Considering the factors of term length，term POS and term internal associative strength of character strings，reasonable extended rules are designed for components extension to get the candidate terms. Then，Ontology terms are

filtered from candidate terms by using the relational information and the contextual information. [**Results**] Experimental result shows that accuracy rate is 83.5%，the recall rate is 87%，the accuracy rate is 2.5 percentages over the baseline. [**Limitations**] It needs a balanced corpus to extract term component，and term extracting effect is effected by the quality of the term. [**Conclusions**] The method is effective and has a positive significance for Ontology learning and Ontology construction etc.

Keywords：Ontology term　Term extraction　Term component　Component extension

《现代图书情报技术》特邀专栏组稿

《现代图书情报技术》是中国科学院主管、中国科学院国家科学图书馆主办的计算机信息管理技术方面的学术性刊物。刊物拥有清晰的定位，即以跟踪技术的研究、应用、交流为主体，服务于广大信息技术人员。

本刊从 2004 年起开设不定期栏目——《特邀专栏》，每一期专栏集中发表关于某个特定方面的技术研发与应用的研究型文章，汇集科研成果、聚焦研究前沿。

1　《特邀专栏》操作办法及流程

（1）本栏目特邀国内外知名专家、学者、教授担任专栏主编，专栏的设立一般由期刊的策划编辑和特邀专栏主编沟通，根据国内外图书情报技术学科的发展需要提出选题。

（2）选题一旦确定后，由特邀专栏主编承担稿件的组织，审核并撰写前言。一期特邀专栏一般为 4-6 篇文章为宜。稿件组织过程中，策划编辑将与特邀专栏主编进行定期的沟通，及时掌握稿件的撰写情况，并对稿件的撰写提出适当的建议和意见。

（3）稿件经特邀专栏主编审核通过，提交给编辑部。后期由策划编辑负责与作者的联系沟通及安排出版等事宜。

（4）专栏的选题一旦确定后，将确定基本时间表。一般的操作周期为 3-5 个月。以正式确定特邀专栏题目为起始点，在 1 个月内确定约请论文的作者和题目，3 个月内确定初稿，5 个月内确定采用稿。

2　《特邀专栏》稿件内容要求

（1）深入反映本专栏选题方向的前沿研究成果或重大应用成果，侧重理论研究、技术分析、系统论证或设计等，注意理论与实践相结合。

（2）特邀专栏稿件应该主要是原始性和原创性研究论文，也可以有一篇综述性论文，但综述性论文必须可靠地覆盖该方向的原始核心文献。

（3）文章按照严谨的学术文章体例写作，即明确扼要地界定研究问题，简要说明研究方法，系统精炼地描述国际国内发展状况，进而详细地描述作者自身研究工作的技术线路及研究结果。

（4）特邀专栏的一系列文章应注意覆盖专栏选题所涉及的各个研究方向和多个研究单位，充分覆盖可能存在的多种观点和技术线路。

（5）充分承认前人/别人的工作，充分引证所参考引用的文献（尤其是本研究工作中的原始核心文献和国内最先出现的研究文献），严格遵守著录规范。

3　《特邀专栏》稿件格式要求

（1）论文版式请参照本刊网站“下载专区”中“论文模板”。

（2）多个作者时，请注明通信作者，并注明各个作者的单位。

（3）每篇稿件以 6-8 千字为宜（按篇幅字数计算，包括图、表）。