# 一种基于文本的领域本体进化需求自动生成模型①

欧阳柳波　兰小飞　伍振兴

(湖南大学 软件学院 湖南 长沙 410082)

**摘　要：** 本体进化研究已经成为领域本体研究的一个重要内容。本体进化需求是本体进化的前提，提出一种基于文本的本体进化需求自动生成框架，首先从自然语言描述的纯文本中提取相关概念，然后利用加权词频算法判断是否为领域关键概念，最后结合本体搜索算法和进化需求生成规则，自动生成本体进化需求。

**关键词：** 本体进化；　进化需求；　自动生成模型；加权词频

## Model for Auto-Generated Requirements of Domain Ontology Evolution Based on Text

OU YANG Liu-Bo, LAN Xiao-Fei, WU Zhen-Xing

(College of Software, Hunan University, Changsha 410082, China)

**Abstract:** Ontology evolution has become more and more important in domain ontology research. The evolution requirements are the basis of Ontology evolution. This paper gives the model for auto-generated requirements of domain ontology evolution based on text. First, it extracts relevant concepts from natural text, and then uses the weighted-frequency-algorithm to determine whether the relevant concepts are the key concepts. Last, it combines the ontology-search-algorithm with the rules of evolution requirements, and automatically generates the requirements of ontology evolution.

**Keywords:** ontology evolution; evolution requirement; auto-generate model; weighted frequency

自 W3C 主席 Tim Berners-Lee 首先提出语义 Web[1,2]的概念后，它正在成为计算机信息处理领域当前研究的热点之一。而本体是语义 Web 实现的关键技术，通常在系统建立之初，根据系统的应用需求和应用环境构造出相应的本体，然后，以该本体为中心，对系统做全面的应用设计。然而，现实世界是不断变化的，这样，固定的本体与变化的知识源之间的数据一致性就可能遭到破坏，本体已经不能正确地反映知识源的新状态。

## 1　引言

### 1.1 文章安排

本文第 2 节介绍本体进化需求自动生成模型的框架和实现该模型的关键技术。第 3 节给出实验数据以及分析结果。第 4 节给出结论以及未来工作。

#### 1.1.1 基本介绍

如何让本体适应动态变化的外部世界，并根据外部知识源的变化做出及时的调整，即本体进化，已成本体研究中的一个重要内容。但是当前领域本体的研究还多停留在本体概念集的构建、描述以及本体的开发工具上，本体进化的研究还在起步阶段。

从本质上来说[3]，本体进化要做的工作就是根据进化需求对系统内的所有相关部分进行修改，以保证系统各部分的一致性，因此进化需求是本体进化的前提和依据。如何从初始文档中提取关键概念、属性、实例，如何围绕着这些概念、属性、实例自动生成进

① 基金项目:国家自然科学基金(60970098;60803024)

收稿时间:2009-10-13;收到修改稿时间:2009-11-18

化需求就是本文要做的工作。

## 2 需求自动生成框架

整个框架流程如图 1：当外界环境发生改变时，系统获取进化初始领域文本，该文本是自然语言描述的纯文本；第二步是对获取的领域文本进行预处理，依据非用词表和 POS tags 词性标注工具对领域概念进行过滤，得到备选概念集合；第三步是利用加权词频公式，对备选概念集合进行筛选，得到与领域相关度高的关键概念集合；最后通过搜索算法，结合领域本体进化需求生成规则表，自动生成领域本体进化需求。具体处理过程如下：

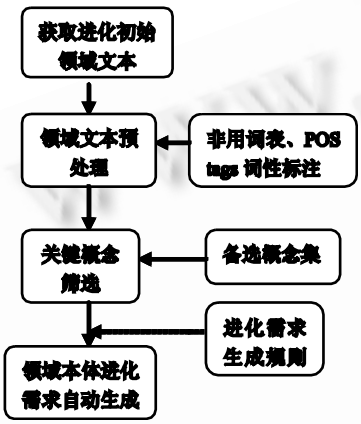


图 1 进化需求自动生成框架

### 2.1 获取自然语言描述的进化文本

进化需求产生的第一阶段是感知并捕获进化需求的初始文档，也就是捕获进化数据源。现实世界中的数据种类很多大致可分为 3 大类[4]：①结构化数据；②非结构化数据；③半结构化数据。本文主要研究如何通过非结构化数据源(自然语言描述的纯文本)产生进化需求。

### 2.2 领域文本预处理

系统获取的非结构化领域文本中包括大量的与领域本体无关的概念，在进行关键概念筛选之前必须对领域文本进行预处理，把与领域本体无关概念过滤掉，形成备选概念集合。领域文本预处理过程中用到的工具包括：非用词表和POS tags 词性标注器。本文定义了一个非用词表，非用词包括：with、for、and、or、a、the、this、of 等，非用词表中包含的词在各个领域的文本中出现频率都非常高，但不能反映领域的专业知识，不能成为领域本体的相关概念，在计算概念词频前必须将文档中出现的非用词找出并且过滤掉。

概念通常是名词或名词短语，非名词性的概念不能成为关键概念，因此必须对文本进行词性标注，过滤非名词性概念。本文所选的词性标注工具是 POS tags，POS tags 是应用相当广泛的英文词性标注器，它是 Stanford NLP Group 开发的。利用 POS tags 对句子 HNU is a famous university.标注结果如下：HNU/NNP is/VBZ a/DT famous/JJ university/NN。其中 NNP 表示专有名词单数，NNPS 表示专有名词复数，NN 表示名词单数，NNS 表示名词复数。

通过非用词表和 POS tags 词性标注工具对领域文本进行预处理的结果是得到备选概念集合。备选概念集合中不含非用词，只包括名词概念。

### 2.3 关键概念筛选

关键概念筛选的结果是得到与领域本体相关度高的关键概念。要抽取文本中的关键概念集并对其反映文本主要内容的重要性进行排序，需要解决两个关键问题：①关键概念的识别和抽取；②关键概念重要性的衡量和筛选方法。关键概念的识别和抽取在领域文本预处理过程中已完成，并且得到备选概念集合，在备选关键概念集合中衡量概念的重要性以及筛选关键概念是关键概念筛选过程要做的工作。关键概念挑选流程如图 2 所示：

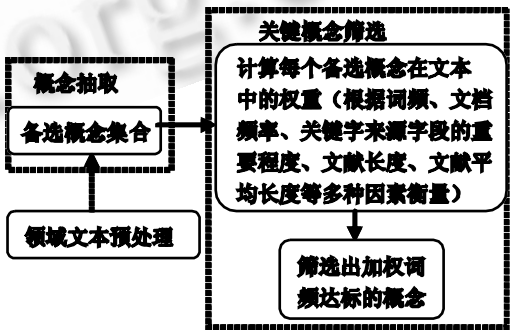


图 2 加权词频挑选关键概念

本文采用基于统计的加权词频算法[5]从备选概念集合中筛选领域关键概念。基于统计的加权词频算法是 Robertson 提出的，它是典型的概率检索模型，比较适合对关键概念进行加权的词频统计，它保证了词频的作用不会成为衡量关键概念是否领域相关的唯一标准。该算法综合考虑了词频、文档频率、关键字来

源字段的重要程度、文献长度、文献平均长度等多种因素衡量。加权词频算法的基本思路是计算出概念集合中每个候选概念 j 的$w_j(\bar{d},C)$值(加权值)，再利用加权值计算出每个候选概念 j 的$F_j$值(加权词频值)，最后规定一个阀值 W，对于每一个概念 j 的$F_j$值，如果$F_j \geq W$，则概念 j 成为关键概念，如果$F_j < W$，则概念 j 不能成为关键概念。本文采用 Robertson 提出的领域词频加权的权重计算公式 BMF25 如下：

$$w_j(\bar{d},C)=\frac{(k_1'+1)tf_j'}{k_1'((1-b)+b\frac{dl'}{avdl'})}\log\frac{N-df_j+0.5}{df_j+0.5} \quad (1)$$

其中，$tf_j'$指加权后的第 j 个概念的词频，$dl'$是加权后的文档长度，$avdl'$是加权后的平均文档长度，$k_1'$[6] 是加权后的自由参数，取值 2，$df_j$是指文档频率，b 是自由参数，取值 0.75，N 为文档集合 C 中的文档数。文档集合 C 是某特定领域的文档集合。对于不同域的词频系数$w_f$具体设为：$w_f${题名、关键词、摘要、正文}={1，0.8，0.5，0.1}。

计算概念集合中的每个备选概念 j 在文档集合 C 中的每篇文档 d 的权重$w_j(\bar{d},C)$以及该概念 j 在该文档中的词频$tf_j$。

$$F_j=\sum_{\bar{d}\in C} w_j(\bar{d},C)\times tf_j \quad (2)$$

根据公式(1)(2)计算的结果，按照每个概念的$F_j$值由大到小顺序排列，把加权词频值大于阀值 W 的概念挑选出来，成为领域关键概念。阀值 W 的大小与文档集合 C 中的文档数量有关。文档数越多，则关键概念加权词频值就越大，阀值 W 的取值也越大。但是通过非用词表过滤后，与领域相关性小的名词概念，在领域文档集 C 中出现的概率小，则它们的加权词频值与关键概念的加权词频值有明显的差别。

然而此步得到的关键概念一般是以某种形式出现的(如复数形式等)，此时的概念不符合本体中概念的定义，必须将筛选出的关键概念进行概念还原，本文采用 Martin Porter 提出 Porter Stemming algorithm。经该算法处理的结果如：stems→stem。

### 2.4 进化需求生成

在该过程中用到本体搜索算法，该算法的思路是：利用 WordNet，将领域关键概念与领域本体中的概念一一对比分析，得出领域关键概念与领域本体的关系。关键概念与领域本体的关系分为关键概念包含在领域本体中和关键概念与领域本体相关两大类，领域关键概念可能是领域本体中的类、属性、实例或是领域本体中概念的父类、子类和相关类，结果如表 1 所示：

表 1　关键概念与领域本体关系

| 关键概念包含在领域本体中 | |
|---|---|
| InOntologyAsClass | 类 |
| InOntologyAsObjectProperty | 对象属性 |
| InOntologyAsDataTypeProperty | 数据属性 |
| InOntologyAsInstance | 实例 |
| 关键概念与领域本体相关 | |
| SubClassInOntology | 子类 |
| SuperClassInOntology | 父类 |
| RelateToOntology | 相关类 |
| NotInOntology | 无关类 |

通过在 WordNet 中的搜索，得到关键概念的同义词关系(Synonyms)、反义词关系(Anto-nyms)、上位关系(hypernyms)，下位关系(holonyms)，附属关系(Meronyms)。将关键概念各种关系下的语义分别与领域本体作比较，即得出关键概念与领域本体间的关系。如：通过查询 WordNet，可以得到 Digital_Camera 与 Camera 是下位关系，则 Digital_Camera 是 Camera 的子类。尽管 point 不包含在领域本体中，但通过 WordNet 搜索，得到 point 的同义词关系 interval 包含在领域本体中，则关键概念 point 与领域本体的关系为 InOntologyAsClass。

得到领域关键概念与领域本体的关系后，必须依照本体进化需求生成规则才能最终生成进化需求。进化需求生成规则与本体搜索算法得到的结果息息相关，如关键概念与领域本体的关系为 SubClassInOn- tology，则新增该关键概念，并且需要新增该关键概念与领域本体中概念的关联属性。具体内容如表 2 所示：

表 2　本体进化需求生成规则表

| 关键概念包含在领域本体中 | |
|---|---|
| InOntologyAsClass | 合并类 |
| InOntologyAsObjectProperty | 合并对象属性 |
| InOntologyAsDataTypeProperty | 合并数据属性 |
| InOntologyAsInstance | 合并实例 |
| 关键概念与领域本体相关 | |
| SubClassInOntology | 新增类、新增关联属性 |
| SuperClassInOntology | 新增类、新增关联属性 |
| RelateToOntology | 新增类、新增关联属性 |
| NotInOntology | 无 |

定义：领域本体或关键概念集中与关键概念相关联的概念称为源概念。

本体进化需求的表示形式：关键概念名称(与领域

本体关系，概念类型，源概念，操作类型，新增关联属性)。

说明：关键概念与领域本体的关系如表 1 所示；操作类型包括：合并和新增；若新增概念，则新增关键概念与源概念之间的关联属性，否则省略。

例如：关键 Hen(母鸡)与动物领域本体中源概念 Herbivore(草食动物)的关系是 SubClassInOntology，则进化需求可表示为：

Hen(SubClassInOntology，Herbivore，新增概念，新增 kind-of 关联属性)。

## 3 实验

实验平台： OWL 本体编辑器为 Protégé 3.4；英文词性标注器： POS tags；语义相关性判断：WordNet；开发工具：Jena2.6.0 和 MyEclipse6.0；测试本体:简单的动物本;以 2008 下半年 Zoological Record 数据库中所有动物领域文献集合(C)为数据来源。进化需求初始文档如下：

First came the three dogs, Bluebell, Jessie, and Pincher, and then the pigs, Pincher was the oldest animal on the farm, who settled down in the straw immediately in front of the platform. The hens perched themselves on the windowsills.

从该篇文献中提取的概念集合以及通过加权词频公式计算得到的结果如下：

表 3 加权词频公式计算结果

| 概念 | 词频 | 来源文档数 | 加权词频 |
|---|---|---|---|
| Dogs | 2080 | 932 | 254.6 |
| Pigs | 3085 | 1245 | 658.9 |
| Hens | 2750 | 1087 | 487.2 |
| Animal | 4052 | 2510 | 869.4 |
| Platform | 28 | 19 | 22.4 |
| …… | | | |

假定阀值 W 为 200，则加权词频大于 200 的概念成为关键概念，得到的关键概念还原后有：Dog、Pig、Hen、Animal。

实验测试用的领域本体类图如图 3 所示：

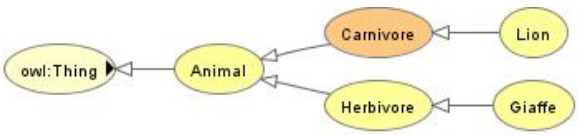


图 3 实验领域本体类图

利用本体搜索算法，得出 Dog、Pig、Hen、等关键概念与动物领域本体的关系，然后结合进化需求生成规则(表 2 所示)，生成进化需求如表 4 所示：

表 4 进化需求表

| 关键概念 | 进化需求 | |
|---|---|---|
| Animal | InOntologyAsClass | 合并 Animal 类 |
| Pig | SubClassInOntology | 新增 Pig 类<br>新增与 Herbivore 的 kind-of 关系 |
| Hen | SubClassInOntology | 新增 Hen 类<br>新增与 Herbivore 的 kind-of 关系 |
| Dog | SubClassInOntology | 新增 Dog 类<br>新增与 Carnivore 的 kind-of 关系 |

## 4 结论

实验表明，利用本文提出的本体进化需求自动生成模型，能自动从文档中过滤、提取出关键概念，并且生成进化需求，但是在提取关键概念阶段并没有考虑到概念的二义性，这就导致进化后的本体概念存在歧义，使得以该本体为基础的应用系统不能正确推理从而导致局限性。如何去除关键概念的二义性也是本人下一步的要考虑的工作。

## 参考文献

1 张子振.本体进化关键技术研究.计算机与网络学报, 2008,10(4):202－204.

2 Berners-Lee T, Hendler J, Lassila O. The Semantic Web. Scientific American, 2001,284 (5):34－43.

3 周明建,高济,李飞.面向 OML 的本体进化框架. 计算机辅助设计与图形学学报, 2005,17(3):102－106.

4 杜小勇,李曼,王珊.本体学习研究综述.软件学报, 2006,17(9):54－58.

5 Robertson SE, Walker S. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval.

6 Robertson SE, Walker S, Jones K S, et al. Okapi at TREC-3. Proc. of 3rd Text Retrieval Conference (TREC-3), 1995. 109－126.