本体进化需求自动化生成模型的构建与实现

尹绍宏 李 敏
(天津工业大学计算机科学与软件学院 天津 300387)

摘 要 通过对本体进化的需求分析,提出了一个进化需求的自动生成模型。在此模型中主要通过对领域文本进行分词获得候选概念,先进行概念还原,再通过筛选、简约、转换获得最终的关键概念。采用ATF*PDF算法实现了关键概念的筛选,并引入叙词表把概念转换成规范化格式,把复合的进化需求分解成相应的原子变化,并结合相应的进化策略实现了一个本体自动进化系统。对实验的评测和分析说明了该模型能获得得良好的实验效果。

关键词 本体进化,进化需求,自动化生成,ATF*PDF算法
中图法分类号 TP391 **文献标识码** A

Design and Realization of Ontology Evolution Requirement Auto-generated Model

YIN Shao-hong LI Min
(School of Computer Science and Software Engineering, Tianjin Polytechnic University, Tianjin 300387, China)

Abstract By analysing the requirement of ontology evolution, it proposed a model of ontology evolution requirement which can be generated automatically. By segmenting the domain texts, we got the candidate concepts, and got the final key concepts by restoration, extraction, conciseness and shift. This paper adopt ATF * PDF algorithm to get the key candidate concepts, and bring in the thesaurus, which is used to convert the natural concept to normalized format. Then, we decomposed the compound demands of evolution to atomic changes, and implemented an evolution system by using corresponding evolution strategy. The experiment result indicates the model can get a good result by evolution and analysis of the experiment.

Keywords Ontology evolution, Evolution requirement, Auto-generated, ATF * PDF algorithm

1 引言

目前,鉴于本体在知识共享和重用过程中的关键作用,大部分知识系统都使用本体作为系统的骨架[1]。建立系统时依照系统的应用环境和需求建立相应的本体,并以此为中心,设计出满足需求的应用系统。但现实世界无时无刻不在改变,本体中的知识的含义、层次关系、存在方式也应向前发展,另外,用户需求也是不断改变的[2]。因此,最初构建的本体不可避免地变化着,原有的本体已经不能正确地反映知识源的新状态。如何让本体与时俱进,并根据外部知识源的变化作出及时的调整,即如何实现本体的动态进化,这已经成为本体研究中的一个重要内容。

本体进化的目的是为了使本体更好地满足用户需求,而对本体做出适当的变化。为了考察变化的起因,我们依据Studer的本体定义来分析可能导致本体变化的因素:概念的变化;概念模型的变化;规范的变化;领域的变化。本体进化的原因正是由于本体所在的信息系统环境发生了改变。本体信息系统需要适应环境而对本体做出相应的调整,这样本体才能正确地反映知识源的新状态,适应变化后的环境需求[3]。

尽管领域本体在知识组织体系中的地位日益重要,但目前对领域本体的研究主要集中于本体的构建、表示和开发工具上,已有的本体大多数规模较小,应用范围窄,而本体管理及评价还没有形成相应的标准[4]。本体进化的相应研究和实践则表现得更加不完善,到目前为止还没有形成一套共同认可的方法论、指导方针和进化框架。本体进化的研究之路还需要进行很多相关工作,因此对其的研究是有价值和现实意义的。

2 本体进化需求自动生成模型设计

进化需求是本体进化的首要阶段,本体进化根据需求来执行领域更新。全面、正确的进化需求是衡量本体进化系统好坏的关键[5]。

本体进化需求的数据可以是结构化的数据(如数据库)、半结构化数据(如HTML、XML文本)和非结构化数据(如自然语言形式的文档)。自然语言文档,主要是相关领域中的期刊、文献、著作等,这些文档里包含了相关领域中的主要概念和关系,是进化需求的主要数据源。本文的需求来源是非结构化的数据,在此基础上实现需求的自动化生成。经过词性标注、剔除应删除词等操作,利用领域关系查找算法,并利用相应的进化需求生成规则,实现领域进化需求的自动化生成。

到稿日期:2012-04-11 返修日期:2012-08-09
尹绍宏(1966—),女,副教授,硕士生导师,主要研究方向为数据挖掘、图形图像处理;李 敏(1987—),女,硕士生,主要研究方向为数据库应用技术。

进化需求的自动生成模型设计如图1所示。

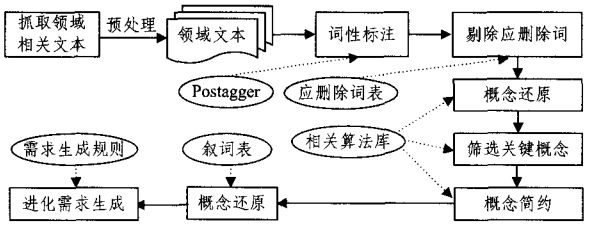

图1 进化需求自动生成模型设计

2.1 领域文本的获取及预处理

首先通过Web信息抓取工具WebZip来获取系统所需要的领域文本，由于文本中可能包含一些无关信息和噪声页面，先对这些数据进行处理，使之满足领域文本的要求，本文研究的是英文本体，因此采用词性标注工具开源工具Postagger。

针对Postagger的使用给出一个简单的例子。给定的句子为：Lisa is a lovely girl，使用Posttagger对该句子标注的结果显示如下：Lisa/NNP is/VBZ a /DT lovely/JJ girl/NN，其中单词和词性用'/'分隔。

词性标注后的语料中包含了很多无关词，考虑到系统运行的高效性和结果准确性，我们利用已有的应删除词表，剔除语料中的虚词、连词、介词等。例如：a，the，an等。由专家经验可知，通常名词带有大量的数据信息，从而成为概念的备选。因此，本文只获取语料中的名词作为备选的关键概念集合。

2.2 对概念进行还原

获取到的备选关键概念集合一般是以某种形式出现的，比如大小写不一致、复数形式。通常极少数概念在领域文本中以其原形形式出现，而这显然不是我们想要的。因此，我们必须对关键概念进行还原，本文采用Martin Porter提出的Porter Stemming算法。该算法是目前最有效的概念还原算法。Porter Stemming算法规则及实例如表1所列。

表1 Porter Stemming算法规则及实例

还原的形式规则	还原的具体实例
将单词尾有元音的es、e、ed、y替换掉	searched→search
把单词尾为tional、fulness、iveness等替换为tion、ful、ive等	conditional→condition
把单词尾为icate、alize、iveness等替换为ic、al、ive等	specialize→special
删除剩余的标准单词尾，例如al、ance、er、ic	critical→critic
去除单词尾没有元音的e	because→becaus

通过对概念的还原，把还原后相同的概念合并，并把它们出现的频次相加，这样在关键概念抽取时，增加了该概念出现的权重，也增大了此概念成为关键词的可能。

2.3 关键概念的筛选

筛选关键概念是为了获得与领域本体紧密相关的概念集合。由于每个概念对文本的贡献程度是不同的，比如题目中的概念比正文中的概念更能反映文章的主题内容，因此不能简单地仅用词频统计次数来衡量概念的重要度，在关键词权重的计算中最广为应用的方法是"词频与逆文档频度"(Term Frequency * Inverse Document Frequency，TF * IDF)方法，其基本思想是：一个词在特定文档中出现的频率越高，说明它在区分该文档内容属性方面的能力越强(TF)；一个词在文档中出现的范围越广，说明它区分文档内容的属性越低(IDF)。TF * IDF的经典计算公式如式(1)所示。

$$w_{ij}=tf_{ij}\times idf_j=tf_{ij}\times\log(N/n_j) \tag{1}$$

式中，tf_{ij}是特征项t_j在文档d_i中出现的次数；idf_i指出现特征项t_j的文档的倒数，N是指总的文档数，n_j指出现特征项t_j的文档数。

国内外很多学者对TF * IDF的应用进行了深入研究，很多学者分析了TF * IDF的缺陷，对其进行改进，并通过实验验证了改进的有效性，因此本文并不直接使用TF * IDF，而是采用改进的关键词抽取算法ATF * PDF，以保证其更适合我们的研究领域。

ATF * PDF(Average Term Frequency * Proportional Document Frequency)方法的词语权重公式如式(2)、式(3)所示。

$$W_i=\frac{\sum_{i=1}^{N}|tf_{ji}|}{N}e^{\frac{n_i}{N}} \tag{2}$$

$$|tf_{ji}|=\frac{tf_{ji}}{\sqrt{\sum_{i=1}^{n_j}tf_{ji}^2}} \tag{3}$$

式中，W_i是词语的权重，N是领域文本包含的文档数，n_i为领域文本中包含词的文档数，n_j为第j个文档的词表大小，tf_{ji}为词语i在文档j中出现的频度。由式(2)、式(3)可知，ATF * PDF公式由两部分组成，ATF表示词语在整个文档集合中的平均词频，PDF是词语的比例文档频率。抓取而来的领域文本中每个文档的大小不同，文档越大词语在文档中出现的次数就可能越多，为了降低文档大小对词语频度的影响，该方法对词语在文档中的频度进行正规化，然后取词语在单个文档中词频的平均值作为词语在文档集中的词频。同时词语存在的文档数不同，对文档集主题的反映度也不同，词语的文档频率越大，就越可能反映文档集的主题，PDF给较多文档中出现的词语赋予更大的权重。因此该算法具有良好的关键词筛选效果。本文中对词语的权重进行了排序，并设置了一个阈值T，如果满足$w\geqslant T$，则w成为关键词；否则，舍弃。

2.4 概念的简约

通过上述步骤之后获得的概念集合并不是最终的关键概念集合，其中可能存在不同概念表示同一个语义的情况，从而导致概念冗余，因此需要对概念集合进行概念简约，去除或合并概念中语义相似的概念。因此本文采用基于WordNet的语义相似度算法。这里使用Hirst、St-Onge提出的Hirst-St-Onge算法，该算法简单，易于实现。其主要思路是：假如两个单词在语义上足够近，那么需要符合以下要求，即在WordNet中连接这两个单词的同义词组的路径不能过长，且在这条路径上改变方向的次数不能过多。其式如(4)所示：

$$rel_{HS}(c_1,c_2)=C-pathlength-k\times d \tag{4}$$

式中，rel是关系relation的缩写，即概念的相似度；HS是算法名称的缩写；c_1，c_2是指某两个同义词；C和k均是常量；$pathlength$是指c_1，c_2这两个同义词间的路径长度；d表示在此路径上方向的变化次数。如果没有路径，则c_1，c_2无关系，否则计算相似度。相似度大于给定阈值S，则被认为是同一含义的概念具有不同的表达形式。

2.5 概念的转换

简约之后的关键词是以自然语言形式来表达的，但对本体而言，其自身已经发展为一个成熟体系，有自己的描述语言标准、推理机制、查询语言和建模工具，而自然语言形式的表

达并不适合本体。叙词表的完善成熟体系在本体的构建和进化时能提供良好的参照作用，而且叙词表的规范化表达十分利于本体进化时概念合并的选择性。因此本文引入叙词表对自然语言形式的关键词进行转换。

3 构建需求生成规则

获取到关键概念后，下面则需要利用这些概念来生成本体的进化需求。这需要依靠领域关系查找算法来完成。利用WordNet等语义词典，将关键概念和领域本体中的概念一一对比分析，以获得关键概念和领域概念的关系映射。关键概念和领域本体的对应关系如表2所列。

表2 关键概念与领域本体对应关系表

关键概念与领域本体的关系			
关键概念包含在领域本体中		关键概念与领域本体相关	
InOntologyAsClass	类	SuperClassInOntology	父类
InOntologyAsDataTypeProperty	数据属性	SubClassInOntology	子类
InOntologyAsObjectProperty	对象属性	RelateToOntology	相关类
InOntologyAsinstance	实例	NotinOntology	无关类

其中，关键概念可能是本体中的类、属性或是某个领域概念的父类、子类、相关类。

通过在WordNet中的搜索，可以得到关键概念的同义词关系、反义词关系、上位关系、下位关系等，把关键概念各种关系下的语义分别与领域本体作比较，就能得到关键概念和本体概念的关系映射。例如通过查询WordNet，得到Digital_Camera和Camera存在下位关系，则Digital_Camera是Camera的子类，其关系是SubClassInOntology。相应的生成规则，如表3所列。

表3 本体进化需求生成规则

关键概念包含在领域本体中		关键概念与领域本体相关	
InOntologyAsClass	合并类	SuperClassInOntology	新增关，新增关联属性
InOntologyAsData TypeProperty	合并数据属性	SubClassInOntology	新增关，新增类联属性
InOntologyAsObjectProperty	合并对象属性	RelateToOntology	新增类，新增类联属性
InOntologyAsInstance	合并实例	NotInOntology	无操作

我们获取到的本体进化需求，并不是单一的原子变化，往往是由简单变化叠加而成的复合变化。考虑到本体进化过程本身的复杂性，如果一次处理多个变化操作所引起的作用将是相互交织的，将极大地增加本体进化处理的复杂度。因此需要对本体的复合需求进行分解，使其变成原子变化的组合。

4 系统实验及分析

4.1 系统开发工具及环境

采用Java编写，Java运行环境为JDK1.6.0_25，编程平台采用Eclipse3.5.2，以及Jena的开发工具包。系统的测试环境运行于XP之上，RDF解析器：Jena2.6.4，本体编辑器：Protégé3.4.6，英文词性标注器：Postagger；语义相关性判断：WordNet；推理工具：Pellet；测试本体：简单的动物本体。

4.2 实验结果及过程分析

实验测试预先使用Protégé工具构建了一个初始动物本体，层级关系如图2所示。

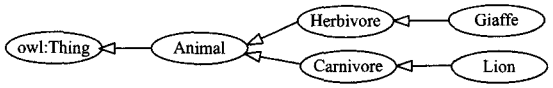

图2 初始动物本体Animal类的层次关系

用ATF*PDF算法来完成关键概念的筛选，在此通过对1000篇领域文本的操作获取到候选关键词及其出现词频、权重，其结果如表4所列。

表4 ATF*PDF算法下的词频计算结果

概念	词频	来源文档统计	ATF*PDF权重
Creature	2663	783	567.5
Animal	3201	894	746.3
Dog	3000	859	625.7
Tiger	1600	630	522.5
Tail	55	36	38.4
Cow	1903	527	480.2
Sheep	1327	410	183.7
…	…	…	…

得到词频的统计结果后，再经由概念简约、转换后，根据概念权重的大小，选择大于阈值T的概念作为系统待进化的概念，并系统呈现所有待进化的关键概念。用户可以根据自己系统的应用方向或目的有选择地选择需要进化的概念，并且可以自动设置阈值的大小，适当地调节进化概念的数目。得到关键词之后根据进化需求规则，其进化需求生成如表5所列。

表5 本体进化需求生成表

待进化的关键概念	进化需求	
合并Animal类	InOntologyAsClass	
Creature	SuperClassInOntology	新增Creature类，新增与Animal的is-a关系
Plant	SubClassInOntology	新增Plant类，新增与Animal的is-a关系
Tiger	SubClassInOntology	新增Tiger类，新增与Carnivore的is-a关系
Dog	SubClassInOntology	新增Dog类，新增与Carnivore的is-a关系
Sheep	SubClassInOntology	新增Tiger类，新增与Herbivore is-a关系
Cow	SubClassInOntology	新增Cow类，新增与Herbivore的is-a关系
Hen	SubClassInOntology	新增Hen类，新增与Herbivore的kind-of关系
Branch	SubClassInOntology	新增Branch类，新增与Plant的is-part-of关系
…	…	…

由表5可知，获取到的概念进化并不是简单的原子变化，而是对照第3节中的需求分解表对复杂变化的分解，然后概念才依据进化需求生出规则进行进化，对于本体系统中已存在的概念，为避免无效操作，我们不进行任何进化动作。表5中增加子类时有的选择的是is-a关系，其实也可以定义其他新增关系，这可以由用户来指定进化的关系类型和名称。例如表5中，Hen与Herbivore类是kind-of关系，关系的名称并不局限于本体中已有的类型。本文中需要注意的是Plant类的增加。如果已经向系统增加了Creature类，那么Plant将按照表5中的进化需求操作进行。但如果系统中没有Creature类，在进化时是不成功的。这是因为Plant类是被当

（下转第272页）

[5] Cootes T, Taylor C, Cooper D, et al.. Active shape model-their training and application[J]. Computer Vision and Image Understanding, 1955, 61: 38-59

[6] Shen Tian, Li Hong-sheng, Qian Zhen, et al. Active volume models for 3D Medical Image Segmentation[M]. Computer Vision and Pattern Recognition-CVPR, 2009: 707-714

[7] 林正春，王知衍，张艳青. 最优进化图像阈值分割算法[J]. 计算机辅助设计与图形学学报，2010，22(7)：1201-1206

[8] 陈果，左洪福. 图像分割的二维最大熵遗传算法[J]. 计算机辅助设计与图形学学报，2002，14(6)：530-534

[9] Otsu N. A threshold selection method from gray-level histogram [J]. IEEE Transactions on Systems, Man And Cybernetics, 1979, 9: 62-66

[10] 杨静宇，曹雨龙. 计算机图像处理及常用算法手册[M]. 南京：南京大学出版社，1997

[11] Soda P, Pechenizkiy M, Tortorella F, et al. Guest editorial Knowledge Discovery and Computer-Based Decision Support in Biomedicine[J]. Artificial Intelligence in Medicine, 2010, 50(1): 1-2

[12] Campadelli P, Casiraghi E, Pratissoli S. A segmentation framework for abdominal organs from CT scans[J]. Artificial Intelligence in Medicine, 2010, 50(1): 3-10

(上接第 243 页)

作无关类看待的，因此无法加入到本体中。

4.3 结果评测及分析

本体进化的实质是对现实环境变化的自动更新。目前计算机领域专家们并未对详细的进化过程给出严格的步骤方法，因此，只要能实现本体对现实变化的自动化更新，那么就可以称之为可行的进化手段。

为了更好地检验本文的 ATF * PDF 关键词抽取算法的可行性和效果，本文进行了测验。当前，对信息系统的评估方法主要采用查全率、准确度以及 F1 值来进行评价。

本文考虑到领域文本的大小对进化需求的影响，特地分析了不同规模下的文本数对实验结果的影响，不同文档篇数下查全率和准确度分别如表 6 所列。

表 6 1000 篇和 2000 篇文档情况下的查全率及准确度

领域文本数	系统获取到的概念	正确概念	实际概念	查全率	准确度	F1 值
1000	46	38	49	77.55%	82.61%	79.79%
2000	71	56	68	82.35%	78.87%	80.57%

由表 6 可以看出，不同文档篇数下实际的概念数是不同的，获取到的正确数目，以及由系统抽取到的概念数目也是不相同的。这也造成了系统的查全率、准确度及 F1 值也不相同。为了更好地查看查全率、准确度及 F1 值的变化，特地给出更多不同文档篇数下准确度、查全率及 F1 值的比较，如图 3 所示。

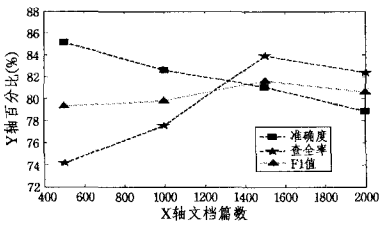

图 3 不同文档数下的准确度、查全率及 F1 值比较

由图 3 可以看出，查找的准确度基本随着文档的增多呈下降趋势，在 500 篇时达到 85%左右，在 2000 篇时准确度也能保持在 80%左右，可见 ATF * PDF 关键词抽取算法的准确度是相当高的，能够很好地满足大规模下的概念准确定位。查全率在 500 篇时是 74%左右到 2000 篇时是 82%左右，在 1500 篇达到 84%，可以看出查全率基本呈现上升趋势，其平均查全率在 78%左右。尽管该算法的稳定性有待改进，但这样的查全率效果还是令人满意的。F1 值基本保持在 75%～82%之间，呈较稳定状态。由此可以认为该算法可以良好地实现关键词的抽取，并具有不错的效果。

本文在概念简约时，借鉴了 WordNet 提供的词语相似度算法，凡是相似度大于给定阈值的多个候选关键词可以认为表达的是同一含义，将会分为同一组。在执行进化时会以一个概念来进行进化，其他概念可以作为被进化概念的属性，以避免概念冗余。为了更好地确定阈值 S 的选择，本文在 1000 篇文档下对不同阈值下的准确度、查全率、F1 值进行了实验，其示意图如图 4 所示。

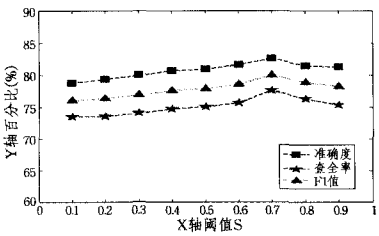

图 4 1000 篇文档数下不同阈值 S 取值下的准确度、查全率及 $F1$ 值

由图 4 可以看出，查找的准确度基本保持在 80%左右，在阈值 $S=0.7$ 时达到最大值 82.61%。查全率基本保持在 73%左右，在阈值 $S=0.7$ 时达到最大值 77.55%。F1 值基本保持在 77%左右，呈现较稳定状态。因此阈值 $S=0.7$ 时，准确度和查全率都能达到最大值，这直接使得 F1 值在 $S=0.7$ 时达到最大值，因此，本文的语义相似度算法在阈值 $S=0.7$ 时能获得最佳性能，故本文的概念简约的阈值设定为 0.7。

结束语 设计了一个需求自动生成的模型，并根据模型建立了相应的一个本体进化系统，实验结果表明该系统具有较好的实验效果，但系统采用本体进化模型所建立的系统还相对简单，亟待完善。对系统进行改进是下一步的研究工作。在未来的研究过程中，可以引入数据挖掘技术来发现数据变化源，使得经过挖掘的文本更具有领域相关性和实时性。

参 考 文 献

[1] Hussain S, Roo J D, Daniyal A, et al. Detecting and Resolving Inconsistencies in Ontologies Using Contradiction Derivations [C]// COMPSAC. 2011: 556-561

[2] Sathya D, Uthayan K R. Proposal for semantic metric to assess the quality of ontologies[C]// ICSCCN. 2011: 754-756

[3] 兰小飞. 基于文本的领域本体进化需求自动生成模型研究[D]. 长沙：湖南大学，2010

[4] Kasisopha N, Wongthongtham P. Semantic Wiki-based Ontology Evolution[C]// 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies. 2009: 493-495

[5] 孙中铁. 服务环境驱动的本体进化需求自动产生[D]. 上海：上海交通大学，2007