非成熟领域的本体构建方法

杨圣洪[1,2],贾 焰[1]
YANG Sheng-hong[1,2],JIA Yan[1]

1.国防科技大学 计算机学院,长沙 410073
2.湖南大学 计算机与通信学院,长沙 410082
1.School of Computer,National University of Defense Technology,Changsha 410073,China
2.College of Computer and Communication,Hunan University,Changsha 410082,China
E-mail:yangshenghong8@yahoo.com.cn

YANG Sheng-hong,JIA Yan.Building method on domain ontology in not growthful knowledge system.Computer Engineering and Applications,2008,44(24):153-155.

Abstract: To obtain the application domain ontology whose concept system is not growthy,a kernel concept word set is proposed and are refined according to the sample documents,to find the nouns and verbs in context of the kernel concept word,and to select the core relation word and to form the semantic surroundings.Finally,to obtain the synonymy word,approximate word and antonym by matching the semantic surroundings,the selected relation word are used for the kernel concept word of next iterativeness,and the algorithm is developed.

Key words: ontology;ontobuilder;concept system

摘 要: 为解决文本数据挖掘等知识体系欠发育领域之本体的构建问题,先构建领域本体的最基本概念词集,利用样本库检测与优化基本概念集,利用样本库构造基本概念词的上下文名词与动词,从而筛选出基本概念词的相关名词,构造其语义环境,通过比对语义环境确定同义词、近义词、反义词。将相关名词作为下一代概念词,迭代计算直至构造整个领域本体,并设计了具体算法,证实了该方法的可行性。

关键词: 本体;本体构建;概念体系
DOI: 10.3778/j.issn.1002-8331.2008.24.046 **文章编号:** 1002-8331(2008)24-0153-03 **文献标识码:** A **中图分类号:** TP311

1 引言

凡需要借助语义的研究领域都在使用“本体”工具,如自然语言理解、语义 Web、文本挖掘、信息检索、知识管理、机器学习,很多本体已进入了工程实施阶段,如 NKI 计划。

目前本体建设多数是面向知识体系已发育成熟的领域,如数学、物理、化学、医学、生物,建立本体时可直接使用该领域的概念体系,知识间的层次关系可作为本体的概念之间的相互关系,并采用本体学习或本体演变的方法[1-2,4-7,12-17]。

对于知识一般比较零散、直观的领域,概念体系往往不成熟或难以建立,主要是采用人工方法或统计方法[8-9,11]来获取其概念体系,效果并不太好。

统计方法是先将文本转化为词,基于 TFIDF 建立词-文档矩阵, 对该矩阵进行奇异值分解以获取词与概念之间的关系,基于相似度聚类获取上下位关系[9]。当样本文档较多时,词-文档矩阵的奇异值分解数据计算量相当大,易出现病态,收敛速度较慢,所获概念集针对性不强。

为此本文采用人工与统计相结合的方法,由领域专家提出初始的核心概念词集, 在样本文档中搜索相关的名词与动词,并按相关性选择第一代概念词汇,辅以适当的人工决策,滚动寻找其后代词汇, 同时建立概念词汇间的层次关系与横向联系,收敛速度较快,针对性较强,提高了非成熟领域本体的构建效率。

2 基本知识

2.1 建立本体的首要工作是建立概念集

概念是对某类事物或行为的总结、归纳,并以规范的文字予以描述。该描述称为概念的内涵。对于概念体系发育成熟的领域,一个概念常用一个准确的词语来表示,如“人、太阳、爱”,因此建立本体概念集就是建立“基本词汇集”。

2.2 本体词汇集主要包括名词

按照亚里士多德的观点,“本体名词是语义中心,一个句子至少包含一个本体名词,它直接出现,或根据所在语境可明确推知”[10]。“句子的重心在动词上,此外凡动作之所由起,所于止,以及所涉的各方面,都是补充这个动词把句子的意义说明白”(吕淑湘)。

所以“动词”与“本体名词”是语义的重点,这两者是同一问题的两个方面,“动词”是指句法关系上,指一个动词与一个或多个本体名词发生句法关系(主语或宾语),“本体名词”是指语

作者简介: 杨圣洪(1965-),男,博士在读,副教授,主要研究领域为分布计算,数据库;贾焰,女,博士,教授,博导,主要研究领域为分布计算,数据库。
收稿日期: 2007-10-33 **修回日期:** 2008-01-21

义,是动词的说明对象。

因此构造本体概念集时,主要是从真实的样本语料库中提取名词与动词,其他的词汇予以忽略,其中名词用于构造概念集,动词用于构造名词的语义环境即上下文环境。

2.3 基本概念集

本文所述基本概念集,是由专家或知识工程师,通过对该领域的关键词、主题词、分类词进行分析而获取,并给出这些基本词汇的语义、相关词、同义词、近义词及经常出现在其上下文环境中的其他名词或动词, 这个基本概念集可理解核心词汇集,形象地称为"生长基",选择不同的核心词汇集,会得到不同的概念体系,这相当于从不同的角度去构建本体,所得到的本体有所不同。

3 本体概念的描述

本体中的每个概念就是一个类,与面向对象的"类"非常类似,如继承、实例、类与类之间的关系等可以直接引入到本体的概念类中,但两者又稍有区别。本体中的概念是为语义服务,它关心的是概念的内涵、外延、上下文环境,而面向对象中的"类"是封装服务,关注的是该类所具有的属性与方法。该算法类如下例所示:

```
defcategory 经济适用房
{
 name:经济适用房
 mean:政府提供政策优惠,限定建设标准、供应对象和销售价格,具有保障性质的政策性商品住房。
 before_noun:        //上文名词
 before_verb:        //上文动词
 next_noun:          //下文名词
 next_verb:          //下文动词
 synonymy:经济房、廉租房          //同义或近义词
 relationWord:政府,保障,政策性,商品,住房        //相关名词
}
```

一个本体概念所需要的属性,应根据所在领域来确定。为了掌握概念词的语义,设计以上属性,这是因为:

(1)"名词"与"动词"是构成语义的重要词汇,在构造词汇的上下文环境即语义环境时, 主要选择在名词周围的其他名词、动词,因此设置 before_noun(上文名词)、before_verb(上文动词)、next_noun(下文名词)、next_verb(下文动词)属性。

(2)为计算语义距离,设置了 mean(含义)、synonymy(同义词)、relationword(相关词汇)属性。

4 算法

对于"生长基"中的概念,尽管可从公共本体(如 Hownet)中获取确定它们的 mean、synonymy、relationWord, 但其针对性不强,效率不太高,因此主要从样本库中获取。

通过样文库,获取每个词汇的"上下文词汇",从上下文词汇中按一定的算法,挑选出与基本概念词相关性较强的词作为"相关词汇",并将此相关词汇作为下一代词汇集,迭代计算,直至计算整个本体的概念词集。

编码实现时,将概念词汇保存到关系数据库中,一个词对应一条记录,词的属性作为字段名,并增加"代码 wordcode"作为词语标识符"name(名字)"的编码,其值为 name 中各汉字的 unicode 码之拼接,再建立一个"层 generation"字段,"生长基"中的词汇为 0 代,并按"generation+wordcode"建立索引。

4.1 基本思路

以初始概念集即生长基为概念树的根 S_0,在此基础上长出第一代、第二代……第 n 代词汇集,以构造出本体概念树,同时得到概念之间的层次关系。

在计算过程中,若当前词汇集为 S_k,由它出发计算与 S_k 距离小于某个阈值的所有词汇得到 S'_{k+1},再剔除其中已在其辈结点中已出现的词汇,即 $S_{k+1}=S'_{k+1}-S_k-S_{k-1}\cdots-S_1-S_0$,要求算法在在挑选词汇时,避免选择父辈词汇,或在已选词汇做某个标志。

形成概念树后,通过"相关词汇、同义词汇、上下文词汇"得到概念之间的横向联系。

4.2 算法

(1)分词

获取每个文档 $doc(i)$ 中的名词与动词, 并分别保存到 $Noun(i)$与 $Verb(i)$,同时记录每个词汇在文档中的位置。

采用比较成熟中科院的的汉语词法分析工具 ICTCLAS,根据领域不同可能要适时调整其词汇表,如对于短文本与网络语言可能要适当增加网络用语与缩略语。

在分词的基础上,利用董振阳的知网(HowNet),确定名词与动词,并确定每个名词的基本语义、相关词汇。动词的语义暂时不考虑。

(2)判断基本词汇集 S_0 与样本文档的关联性

如果基本词汇集与真实的样本语料库的相关性不高,那么基于该语料库而构建的本体质量肯定有问题,因此必须进行检测。

①将基本词汇集 S_0 读入内存,如保存在一个数组,或保存一个内存关系(如游标或 ADO.Net 中的 dataSet 对象中)。

②读入文档 $doc(i)$的名词表 $Noun(i)$的第 j 个词 wj,总的名词计数器增 1。

③若 wj 在基本词汇集 S_0 中出现,则 S_0 中相应词汇的计数器加 1。

④不断重复,当将文档 $doc(i)$中所有名词读完后,再读下一篇样本文档。

⑤计算 S_0 中各词汇 $S_0(k)$的频率 $FS_0(k)$:

$$FS_0(k)=\frac{\text{词汇 } S_0(k)\text{在样本库出现的次数}}{\text{样本中名词总数}}$$

⑥S_0 中满足条件"$FS_0(k)<\alpha$"(α 成为基本词汇集的阈值)的名词,是与样本文档的关联性不强,将其从 S_0 中删除,若 S_0 中剩余的词汇太少, 要求知识工程师或领域专家重新挑选,并重复以上过程。

(3)产生第 k+1 代概念集

①为第 k 代中每个概念词语 $word(i)$ 建立四个链表 $before\_noun(i)$、$before\_verb(i)$、$next\_noun(i)$、$next\_verb(i)$。

②将第 k 代的每个词汇读入内存,只读入其"名字 name",一般领域本体(非顶级本体)的概念词汇一般应在 100 万即以下,故应该能读入内存。

③遍历每个样本文档 $doc(j)$的名词链表 $Noun(j)$,若词汇 $word(j,m)$第于 k 代词汇 $word(i)$,则将其标记为第 k 代。

④将文档 $doc(j)$中名词链表 $Noun(j)$中第 k 代词 $word(i)$之前所有名词 $word(j,n)$,转抄到 $word(i)$的上文名词链表 $before_noun(i)$中,之前的动词转抄上文动词链表 $before_verb(i)$中。

若 $word(j,n)$已在 $before_noun(i)$中出现,则相应的计数器加上,若是新出现则计数器的值置为 1。对于转抄到 $before_verb(i)$中的动词也类似处理。

类似转抄到下文名词链表 $next_noun(i)$与下文动词链表 $netx\_verb(i)$。

⑤当遍历完所有样本文档后，对第 k 代词的每个词语 $word(i)$中的链表 $before\_noun(i)$、$before\_verb(i)$、$next\_noun(i)$、$next_verb(i)$中词汇进行筛选，将“词频<样本总名词数×β”(β 成为“上下文词汇”的阈值)删除，将“词频<样本总动词数×β”的动词删除。

⑥在已选定的上文名词 $before\_noun(i)$、下文名词 $next_noun(i)$的基础上，将“词频≥样本总名词数×γ”(γ 为“相关词汇”的阈值)挑选出来，并交由领域专家或知识工程师审查后，并填入到相关词汇 $relationword(i)$中。

⑦将第 k 代概念词汇的所有相关词汇 $relationword(i)$汇聚起来，保存至 S'_{k+1}，剔除其中的父辈词汇得到第 k+1 代词汇，即 $S_{k+1}=S'_{k+1}-S_{k-1}-S_{k-2}\cdots-S_1-S_0$。

(4)计算同义词、近义词、反义词

①在以上算法得到的本体概念集 $S_0\cup S_1\cup S_2\cdots\cup S_n$ 中，让 i=1~$total\_of\_word$，j=i+1~$total\_of\_word$。

②如果 $word(i)$与 $word(j)$的属性“相关词汇 $relationword$”中相同词汇数超过某个阈值 δ，则将 $word(j)$写入 $word(i)$的“近义词”集中，同时将 $word(i)$也写入到 $word(j)$的“近义词”集中。

如果 $word(i)$与 $word(j)$的“上下文名词”与“上下文动词”的相同词汇数也超过某个阈值 γ，则可认为这两个词是“同义词”，可将这两个词条合并。

③步骤②中产生的同义词、近义词也可能是反义词，所以在确定这三种词语时，需要人工启发。

5　分析

(1)后代词汇是通过父辈词汇的“$relationword$(相关词汇)”产生，从而基于此属性建立了概念词汇之间的层次关系。

在建立后代“上下文名词、动词、相关词汇”时，并没有剔除祖辈词汇，因此算法自动生成了词语之间的横向联系，所以本算法一次生成词语之间的层次关系与横向联系。

(2)后代词汇 S_k 与祖辈词汇之间 $S_j(j<k)$的没有重复，这样可提高算法的效率，避免重复计算而引起的收敛速度慢的缺点。

寻找后代词汇的过程不断进行下去，会将本领域的所有词汇关联起来，但正如网络路由算法中，正常情况下当“跳数hop”达到一定数值后，会到达指定的服务器，从而对 hop 值做出限制一样，寻找后代词汇也必须做出限制，一般有两种停机策略：

①后代数 k 达到某个指定值；

②$|S_0\cup S_1\cup S_2\cdots\cup S_k|$达到指定值，即本体中的概念数达到一定数据。因为当概念数多到一定的程度后，可对本体概念进行细分，既可以提高运算速度，更能提高准确度。

(3)由于后代词汇 S_k 是父辈词 S_{k-1} 的相关词汇，并且有人工启发，故本体词之间的相关性高、针对性及汇聚性强，避免了基于“词频-文档”矩阵的奇异值分解等统计手段所建立本体词汇集，针对性不强的缺点。

(4)根据上文与下文的动词、名词来构建其上下文环境，与自然语义的理论一致，使算法不仅具有统计性，还具有可理解性。

(5)当利用本体来标记待分析文档，可以先将本体概念集读入内存，然后：

①每读入文档的某个词语，便从本体概念集的“name(名字)”属性中寻找，若找到则为标记为此本体词，吻合度可标为100%；

②若没找到则从“同义词、近义词、反义词”中去查找，若找到则如上标记，吻合度视情况来定；

③若没找到则从“相关词汇”中去查找，若找到若找到则如上标记，吻合度更低；

④若没找到则只从“上下文名词、动词”中去查找，若找到了则如上标记吻合度最低。

6　总结

由领域专家或知识工程师给出的基本词汇，利用真实样本文档的检测与优化基本词汇集，然后寻找基本词汇的上文、下文环境中的名词、动词，然后确定其相关词汇，剔除其与祖辈相同的词汇得到新一代词汇集，最终考虑整个本体概念集生成各个概念的近义词、同义词，遵循自然语言的理论，收敛速度与针对性都较强。这是一种基于核心本体进化得到完整本体的方法。

本体词语的构造是一项复杂的工程，需要机器学习、数据挖掘、自然语言的理解、内容分析法、统计方法与专家系统的方法等工具综合运用，将在分析现有文本聚类、分类的方法基础上，多作对比分析，提高算法的效率。

参考文献：

[1] 王海涛，曹存根.基于本领域本体的半结构化文本知识自动获取方法的设计和实现[J].计算机学报，2005(12).

[2] 罗贝，曹存根.从文本中获取植物知识方法的研究[J].计算机科学，2005(10).

[3] 贾焰.基于本体论的文本挖掘技术概述[J].计算机应用，2006(9).

[4] 曾庆田，曹存根.基于本体的数学知识获取与知识继承机制研究[J].微电子学与计算机，2003(9).

[5] 周肖彬，曹存根.基于本体的医学知识获取[J].计算机科学，2003(10).

[6] 王丽丽，曹存根.基于本体的民族知识获取与分析[J].计算机科学，2003(5).

[7] 雷玉霞.基于知识本体的属性分析以及概念联通[J].计算机科学，2004(3).

[8] 杨明华.特定领域本体的构造方法[J].计算机工程，2006(6).

[9] 董慧.中文本体的自动获取与评估算法分析[J].情报理论与实践，2005(4).

[10] 姚振武.论本体名词[J].语文研究，2005(4).

[11] 陈治平，林亚平.基于最小类差异的无关信息预处理算法[J].电子学报，2003(11).

[12] 杜小勇.本体学习研究综述[J].软件学报，2006(9).

[13] Avigdor G，Giovanni M，Hasan J.OntoBuilder：fully automatic extraction and consolidation of ontologies from web sources[C]//Proc of the ICDE 2004.Boston：IEEE Computer Society，2004：853-853.

[14] Volz R，Oberle D，Staab S，et al.OntoLiFT prototype.IST Project 2001-33052 Wonder Web Deliverable 11，2003.

[15] Hotho A，Staab S，Stumme G.Wordnet improves text document clustering[C]//Proceedings of the Semantic Web Workshop at SIGIR-2003：26th Annual International ACM SIGIR Conference，2003.

[16] Hotho A，Maedche A，Staab S.Ontology-based text document clustering[C]//Proc of the Conf on Intelligent Information Systms.Zakopane：Springer-Verlag，2003.

[17] Jing L P.Zhou L X，Ng M K，et al.Ontology-based distance measure for text clustering[C]//2006 SIAM Conference on Data Mining，2006.