一种本体学习中分类关系提取方法的研究

贾秀玲[1],文敦伟[1,2]
(1.中南大学 信息科学与工程学院,湖南 长沙 410083;
2.阿萨巴斯卡大学 计算机与信息系统学院,加拿大 阿萨巴斯卡 AB T9S3A3)

摘 要:本体学习技术是利用本体工程技术和机器学习技术等众多学科技术来实现本体的自动半自动构建,可解决本体手工构建的不足。根据本体学习目前的研究现状,提出了一种从文本中半自动获取本体中分类关系的实现,讨论了本体学习中概念抽取和概念间分类关系抽取等关键技术。实现了本体中分类关系提取,对于非分类关系的提取还有待研究。

关键词:本体;本体学习;概念;分类关系
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2007)10-0031-03

A Study on Taxonomic Relation Extraction from Ontology Learning

JIA Xiu-ling[1], WEN Dun-wei[1,2]
(1. School of Information Science and Engineering, Central South University, Changsha 410083, China;
2. School of Computer and Information Systems, Athabasca University, Athabasca AB T9S3A3, Canada)

Abstract: Ontology learning aims at constructing ontology (semi) automatically by integrating a multitude of disciplines such as ontology engineering and machine learning. This can lighten the burden of the manual construction of ontology. This paper introduces a framework of extracting the taxonomic relation semi-automatically for ontology learning from text. The key technologies of ontology learning such as domain concepts extraction and taxonomic relation extraction are discussed. The taxonomic relation of the ontology is realized, but the non-taxonomy relation need to be researched.

Key words: ontology; ontology learning; concept; taxonomic relation

0 引 言

目前对于本体(ontology)的研究在计算机科学领域变的越来越广泛,本体已经被广泛应用于语义 web、电子商务、信息提取、数字图书馆等领域。本体的概念最初起源于哲学领域,是客观存在的一个系统的解释或说明,关心的是客观现实的抽象本质。对本体的定义[1]引用最广泛的是由 Gruber 提出的"本体是概念模型的明确的规范说明"。

本体在众多领域的应用都是在构建本体的基础之上实现的,但本体的构建却是一项繁琐而辛苦的任务。虽然现在本体构建工具也日趋成熟,从最早的 Ontoligua[2], WebOnto[3]到 Protege2000[4], OntoEdit[5],这些工具提供了友好的图形化界面,借助这些工具,用户可以把精力集中在本体内容的组织上,而不必了解本体描述语言的细节,方便了本体的构建,但是这些工具支持的仍然是手工构建本体的方式。为了利用知识获取技术来降低本体构建的开销,采用了本体学习(ontology learning)技术。本体学习技术是综合本体工程技术、机器学习技术和统计等技术自动或半自动地构建本体。近年来,本体学习技术逐渐成为计算机科学领域的一个研究热点。概念之间的分类关系是本体中一个重要的组成部分,但目前主要是面向手工实现。

收稿日期:2006-12-14

作者简介:贾秀玲(1980-),女,内蒙古人,硕士研究生,研究方向为自然语言处理、语义网与本体工程;文敦伟,博士,教授,研究方向为分布式人工智能与知识工程、自然语言处理与机器学习、智能系统与计算智能。

1 本体学习研究现状

Alexander Maedche[6]对 ontology 结构定义为一个5元组 $O:=\{C,R,H^C,\mathrm{rel},A^O\}$。其中,$C$ 和 R 是两个不相交的集合,C 为概念集合;R 为关系集合;H^C 为概念层次,即概念间的分类关系(taxonomy relation),$H^C \subseteq C \times C$ 是一种有向关系,$H^C(C_1,C_2)$ 表示 C_1 是 C_2 的子概念;$\mathrm{rel}: R \rightarrow C \times C$ 是一个函数,表示概念之间的非分类关系(non-taxonomy relation),$\mathrm{rel}(R)=(C_1,C_2)$ 亦可表示为 $R(C_1,C_2)$;A^O 为使用某种逻辑语言表达的 ontology 的公理集(axiom)。从本体的结构

可以看出,本体学习的任务主要包括概念的获取、概念间关系即分类关系和非分类关系的获取以及公理的获取。这三个对象构成了本体学习从简单到复杂的层次。

目前国外对于本体学习的研究已经比较成熟,像Maedche[7]和Staab[8]提出了本体获取的框架,包括本体导入(ontology import)、本体抽取(ontology extraction)、本体裁剪(ontology pruning)、本体精炼(ontology refinement)和本体评估(ontology evaluation),并对如何从文本、字典和原有本体中获取新的本体进行了研究。Velardi和Missikoff[9]等人提出的基于文本的OntoLearn本体学习系统是采用一种基于语义解释的方法,即首先使用基于语言学和统计的方法从一组文本集中抽取出领域相关的术语,然后使用通用本体WordNet中的概念对这些术语进行语义解释,从而确定术语之间的分类关系以及其他语义关系。国内对于本体学习的研究则相对较少,2002年,李守丽[10]等人借鉴了国外经验,对利用奇异值和概念聚类进行汉语本体获取进行了初步的探讨。但是对于计算词频之前的准备工作和本体获取之后的评估却并没有作详细讨论。

本体学习技术[11]根据所面向的数据源的不同采用不同的学习技术,根据数据源的结构化程度不同数据源分为结构化数据、非结构化数据和半结构化数据。本研究中的本体学习技术主要是针对非结构化数据而言的,非结构化数据是指没有固定结构的数据,其中纯文本是Web中大量存在的一类非结构化数据,也是最重要的一类。笔者用它作为本体学习的数据源。

2 本体学习中分类关系的一种实现

文中研究的目的是实现从文本中半自动地抽取本体,其中包括概念抽取和关系抽取。在关系抽取中主要针对的是分类关系的抽取,分类关系被广泛地用于组织本体的知识,许多系统都把上下位关系(hyponymy relation)作为分类关系来处理。下位/上位关系也称为从属/上属关系,子集/超集关系,或ISA关系。像{枫树}是{树}的下位词,{树}是{植物}的下位词,如:"An x is a (kind of) y."为框架构造的句子,则同义词集合{X,X…}表示的概念称为同义词集合{Y,Y…}表示的概念的下位关系。概念之间的分类关系是本体中的一个重要的组成部分,目前的研究主要是面向手工实现,提出一些本体分类关系的建立模式和标准,利用这些规定建立的本体,其质量过多地依赖于个人水平及工作状态。一些半自动的方法主要是采用模式匹配、基于概念聚类的方法和基于词典的方法。本研究主要采用基于关联规则与模式匹配相结合的方法来获得概念间的分类关系。具体步骤如下(参见图1):

(1)收集领域文集(domain corpus)和一般对比文集(contrastive corpora);

(2)文档预处理;

(3)抽取候选术语集;

(4)过滤候选术语集生成概念集;

(5)通过关系提取算法抽取分类关系并建立分类层次体系。

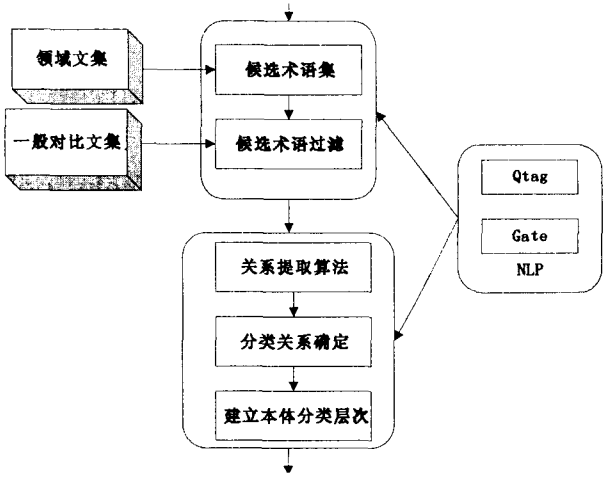


图1 系统基本框架

2.1 文档的预处理及候选术语集生成

在本研究的文档预处理中用到了应用相当广泛的英文标注器QTag[12],目前的标注策略基本上是采用基于概率统计和基于规则的两种技术。QTag是基于概率统计的方法。在研究中采用QTag对语料库进行词性标注。在文档的预处理中主要应用自然语言处理NLP技术。

候选术语的提取主要采用基于统计的方法,即提取术语通过计算术语的频率。根据专家经验知道,术语是一些相当频率的能代表领域特征的词或短语,因此一个词或短语在领域文集中出现一定频率是它作为术语的必要条件,对于出现频率很低的词或短语可以通过设置术语在领域中的出现频率阈值来过滤掉它们。对于一些出现频率特别高的常用词,它们在各个领域中基本都会频频出现但不能反映领域专有知识,可以通过停用词表把它们过滤掉。

2.2 概念集生成

通过选取,虽然从候选术语集中除去了常用词和出现频率较低的词,但集合中还包括一定数量的和领域无关的词,它们常常出现在多个领域文集中但又不在停用词表中,这时必须对候选术语集进行过滤。本研究采用的方法主要是基于两个量化公式,即领域相关度和领域一致度相组合的方法来对候选术语集进行

过滤,以生成真正的领域术语。这种方法需要一般对比文集做支撑。领域相关度的计算公式可由如下三个定义表示[13]:

定义 1(候选术语 t 在集合 D_K 中的领域相关度):领域集合 set = $\{D_1,\cdots,D_n\}$

$$DR_{t,k} = \frac{P(t \mid D_k)}{\max\limits_{1\le j\le n} P(t \mid D_j)}$$

其中条件概率 $P(t \mid D_k)$ 可以用下式来估算:

$$E(P(t \mid D_k)) = \frac{f_{t,k}}{\sum\limits_{t'\in D_k} f_{t',k}}$$, $f_{t,k}$ 是术语 t 在领域 D_k 中的频率。

领域相关度是通过和无关领域比较反映术语与特定领域的相关程度,仅仅使用领域相关度来衡量术语的重要程度还是不够的,因有可能一个术语仅仅在某领域的个别文档中大量出现,这就需要用领域一致度来反映术语在领域文集中的分布情况。

定义 2(候选术语 t 对于领域 D_k 的领域一致度):

$$DC_{t,k} = \sum_{d\in D_k} P_t(d)\log\frac{1}{P_t(d)}$$

其中:

$$E(P_t(d_j)) = \frac{f_{t,j}}{\sum\limits_{d_j\in D_k} f_{t,k}}$$, $f_{t,k}$ 是术语 t 在领域 D_k 中的频率。

所以可根据定义 1 与定义 2,每个候选术语 t 对领域 D_k 的量化公式为:

定义 3(候选术语对领域 D_k 量化公式):

$TW_{t,k} = \alpha DR_{t,k} + \beta DC_{t,k}$,其中 $\alpha,\beta \in (0,1)$

2.3 概念间分类关系的抽取

在本体学习中常采用模式匹配的方法抽取概念间的语义关系特别是分类关系,基于模式匹配的方法是指通过分析领域相关文本,总结出一些频繁出现的语言模式作为规则,然后判断文本中词的序列是否匹配某个模式,如果匹配则可以识别出相应的关系。例如 most European countries, especially France, England, and Spain. 根据模式 NP {,} especially {NP,} * {or and} NP 可判断分类关系(France, European country),(England, European country),(Spain, European country),这些模式可以是手工定义的,也可以是从某些样本句子中学习得到的,这类方法的主要缺点是准确度低,因为大量无用的概念对也往往匹配这些模式,而且模式的获取是否完备对于获取效果影响较大。

对于数据挖掘中的关联规则常用于抽取概念间的非分类关系,其基本思想[12]是如果两个概念经常出现在同一文档或同一段落或者是同一句子中,则这两个概念之间必定存在一定关系。但是大部分方法都停留在判断两个概念之间是否存在关系的层次上,无法进一步确定抽取出的概念之间具体是什么关系。因此可以看出两种方法都有一定的局限性。

本研究中所采用的是基于关联规则与模式匹配相结合的方法用于本体概念间关系的提取,其基本思想主要是首先利用关联规则中的 Apriori 算法在领域文集中发现频繁项目集,主要是发现频繁 2 - 项集,然后利用由频繁 2 - 项集产生的关联规则搜索领域文档集找出含此关联规则的句子,利用统计的方法发现其中的模式,并人工排除不是分类关系的模式,最后再用模式匹配的方法抽取出领域文档集中的分类关系并建立概念间的层次关系。

3 结　论

本体学习技术是当前研究的一个热点,是利用本体工程技术和机器学习技术等众多学科技术实现本体的自动或半自动构建。文中提出了一种从非结构化数据纯文本中半自动化构建本体的本体学习方法。该方法还处于探索阶段,还有许多深入的工作要做,如本体评价、本体实例的扩充、本体集成等。目前,国际上对概念间关系获取的研究很多,但是对于概念间非分类关系的获取,多数都只停留在判断两个概念之间是否存在关系的层次上,没有办法进一步确定概念间是什么关系。在今后的工作中还需要考虑概念间的非分类关系。

参考文献:

[1] Gruber T R. A translation approach to portable ontology specifications[J]. Knowledge Acquisition, 1993, 5(2): 199 - 220.

[2] Deng Z H, Tang S W, Zhang M, et al. Overview of ontology [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2002, 38(5): 730 - 738.

[3] Farquhar A, Fikes R, Rice J. The Ontolingua server: A tool for collaborative ontology construction[J]. Int'l Journal of Human - Computer Studies, 1997, 46(6): 707 - 727.

[4] Noy N F, Fergerson R W, Musen M A. The knowledge model of protégé - 2000: Combining interoperability and flexibility [C]//In: Dieng R, Corby O. Proc of the EKAW2000. Heidelberg: Springer - Verlag, 2000: 17 - 32.

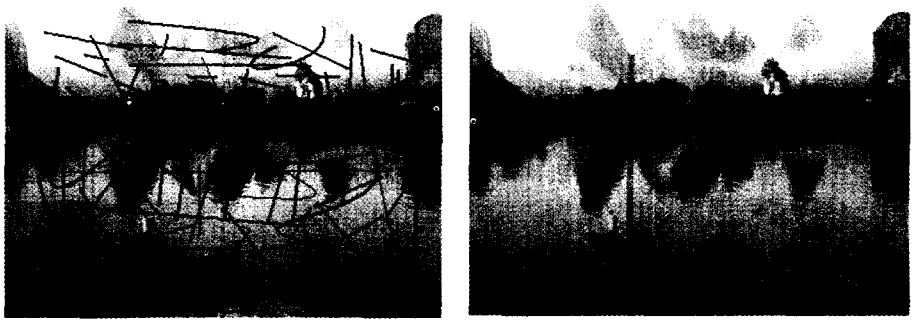
[5] Arpirez J C, Corcho O, Fernandez - Loperz M, et al. WebODE: A scalable Ontological engineering workbench[C]//Gil Y, Musen M, Shavlik J. Proc of the KCAP 2001. New York: ACM Press, 2001: 6 - 13.

[6] Maedche A, Staab S. Ontology learning for the semantic web [J]. IEEE Intelligent Systems, 2001, 16(2): 72 - 79.

(下转第 36 页)

(449×307) 带文字的图像　Oliveira算法修复的结果　文中的修复结果（用时8.125秒）

图 4　新奥尔良



(361×256) 覆盖面占原图像的 11.94%　修复后的图像（用时 4.141 秒）

图 5　风景

(365×272)覆盖面占 17.89%　修复后的图像（用时 7.992 秒）

图 6　草原

ACM Press,2000:411－424.

[2] Chan T, Shen J. Mathematical models for local deterministic inpaintings[EB/OL].2000. http://www. math. ucla. edu.

[3] Chan T, Shen J. Non－Texture Inpainting by Curvature－Driven Diffusions(CCD)[J]. SIAM journal on applied mathematics,2002,62:1019－1043.

[4] Oliveria M M, Bowen B, Mckenna R, et al. Fast Digital Image Inpainting [C/OL] // Proceedings of the International Conference on Visualization, Imaging and Image Processing (VIIP 2001), 2001:261－266. http://www. inf. ufrgs. br/～oliveira/pubs_ files/inpainting. pdf.

[5] 丁　雯. 一类非线性扩散问题及其在图像修复中的应用[J]. 上海交通大学学报, 2004, 38(1): 153－156.

[6] 邵肖伟,刘政凯,宋　璧. 一种基于 TV 模型的自适应图像修复方法[J]. 电路与系统学报,2004,9(2):113－117.

[7] 周廷方,汤　锋,王　进,等. 基于径向基函数的图像修复技术[J]. 中国图像图形学报,2004,9(10):1190－1196.

[8] Efros A A, Leung T K. Texture synthesis by non－parametric sampling[C] // Proceedings of International Conference on Computer Vision. Greece:[s. n.],1999:1033－1038.

[9] 王　晨,杜建洪. 基于图像修复技术的压缩方法的研究[J]. 电子与信息学报,2006,28(5):848－851.

[10] Buades A, Coll B, Morel J M. On image denoising methods [EB/OL]. 2004. http://www. cmla. ens－cachan. fr/Cmla/.

(上接第 33 页)

[7] Maedche A, Volz R. The Text－To－Onto ontology extraction and maintenance environment[C]//Proceedings of the ICDM Workshop on Integrating Data Mining and Knowledge Management. California:[s. n.],2001.

[8] Maedche A, Staab S. Semi－automatic engineering of ontologies from text[C]//Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering. Chicago:[s. n.],2000.

[9] Missikoff M, Navigli R, Velardi P. Integrated approach to web ontology learning and engineering[J]. IEEE Computer, 2002, 35(11):60－63.

[10] Liao L J, Cao Y D, Li S L. A semantic web architecture and its implementation[J]. Computer Engineering and Application, 2003,15:157－161.

[11] Du X Y, Li M, Wang S. A survey on ontology learning research[J]. Journal of Software, 2006, 17(9):1837－1847.

[12] QTag Software[EB/OL]. 2005. http://bham. ac. uk/o. mason/software/tagger/index. html.

[13] Navigli R, Velardi P. Learning domain ontologies from document ware－houses and dedicated web sites[J]. Computational Linguistics, 2004, 30(2):151－179.