面向文本的本体学习方法综述

A Survey of Ontology Learning Methods from Text

傅　魁　聂规划

（武汉理工大学经济学院 武汉 430070）

摘　要：文本是本体学习最主要的数据源，面向文本的本体学习因此成为目前本体学习研究的重点。本文对国内外面向文本的本体学习方法进行了综述，主要包括领域概念学习方法、概念间继承关系学习方法、属性关系学习方法、整体部分关系学习方法以及公理获取方法等各种本体学习方法的分析与评述。

关键词：本体，本体学习，概念，关系

中图分类号：TP181, TP391

本体能够支持人机之间、机器之间的信息交换、知识共享与重用，而得到越来越广泛的重视、研究和应用。然而，领域本体的匮乏却是困扰本体理论研究与现实应用的最主要瓶颈之一，本体学习应运而生，它能够以自动或半自动化的机器学习方式从多种不同的数据源中获取本体，其中文本是最广泛的数据源。从文本中获取本体也因此成为目前本体学习研究的重点。相比国外较多本体学习研究而言，中文环境下本体学习刚刚拉开序幕。本文对面向文本的本体学习方法的研究现状进行了综述，分析与评述了从文本中自动/半自动化获取领域概念、继承关系、其它领域关系和公理的主要方法。

1 领域概念获取方法

1.1 概念获取

大多数的本体学习方法和本体学习系统直接将术语识别为概念。术语的获取方法大体上可以分类三类：基于语言学的方法、基于统计的方法和混合方法。

基于语言学的方法主要使用浅层解析技术（Shallow Parsing Technique）或模板方法获取术语。浅层解析技术是在已进行词性标记的文本中，探测句子中词语边界、发现词语间语法关系（如主谓、动宾关系）的技术，文献[1]均采用了浅层解析技术获取术语。Bourigault[2]认为术语单元有一个固定的词法形式，如名词短语，在他描述的 LEXTER 系统中，在“表面语法分析”基础上第一次抽取出了最大长度的名词短语。Justeson 等[3]进一步扩充了术语的词法形式(即术语的构成模式)。另外，为消除术语的歧义性，Smeaton[4]开始了对语义知识的利用。模板方法[5,6]是根据领域术语的特殊词法结构或模板，寻找和抽取结构符合这些特定模板的字符串。基于语言学知识的术语抽取方法在术语消歧、准确率上有非常明显的优点，但在大多数情况下，基于语言学的方法是与具体语言相关的，因此这类方法具有较高的语言依赖性。

基于统计的方法[1,7]主要根据领域术语与普通词汇在语料中拥有不同的统计特征来鉴别出领域术语，常用的统计方法有互信息(MI)、词频逆文献频率(TFIDF) 、术语相关频率(RTF) 、信息熵(Entropy) 和 C 值/NC 值方法等。基于统计的方法适合于大规模文本处理，但缺乏必要的语义逻辑基础。目前，统计方法是国内外相关研究的主流。Salton 等[8]简单地加权两个相邻的字来抽取术语。Damerau[9]使用互信息来测量两个字之间的联合强度。Cohen[10]利用对数似然参数来避免一些低频词的遗漏，从而较有效地弥补了互信息的不足。Frantzi[11]提出的 C/NC-Value 的方法，联合了语言学和统计学方法。Ellen 提出了一种基于语料库的方法提取给定类的词汇。Patrick[12]将互信息和对数似然两个参数相结合进行术语提取。基于统计的术语自动抽取方法，不考虑句法、语义上的信息，所以实现起来较简单，并且这种方法不局限于某一专门领域，也不依赖任何外部资源。

事实上，大多数本体学习系统往往采用语言学和统计方法相结合的混合方法来获取领域术语。其中，语言规则主要用于抽取候选术语，而统计方法则主要用于抽取前或抽取后的过滤，以更有效、准确地得到领域术语。如在本体学习系统

基金项目：国家科技支撑计划“国家科技支撑计划-电子商务与现代物流共性集成技术研究开发(2006BAH02A08)”，国家自然科学“基于知识网格的电子商务智能推荐系统（70572079）”

作者简介：傅魁,男,1977 年生,讲师,博士生,研究方向为商务智能,知识工程；聂规划,男,1957 年生,教授,博士生导师,研究方向为信息资源管理,商务智能

TextToOnto[7]和 OntoLearn[1]中均采用了浅层解析技术从文本中获取候选词语，然后再采用统计方法对术语进行过滤。

中文领域概念的自动抽取的研究工作起步较晚。2003 年，东北大学的陈文亮等人[13]提出利用 Bootstrapping 的机器学习技术，从大规模无标注真实语料中自动获取领域词汇。梁健等[14]研究基于种子概念的术语获取方法。华南理工大学的方卫东等人[15]都采用统计分析与自然语言规则相结合的方法实现了术语抽取。

1.2 概念的领域性判断

抽取术语可能是领域内概念，也可能不是领域内概念，因此需要对术语的领域性作判断，这一工作被称为领域术语的过滤。一般而言，领域术语的过滤可以通过分析术语在领域相关文档集中和普通文档集中的统计特征来实现。Paola 等[16]提除了术语的领域相关度和领域一致度的概念用于判断术语的领域性。领域相关度是术语与特定领域的相关程度的度量，它可以通过计算术语在特定领域文本集中出现的概率值与其在不同领域文本集中出现的概率值总和的比值来计算。领域一致度是术语在其特定领域的所有文档中分布的一致性的度量，它可以通过术语在特定领域文档中的分布使用的熵值来计算。在领域文档集和对比文档集质量较高的情况下能够较为准确判断术语的领域性，实现领域术语的过滤。

1.3 同义词消歧

术语并不等同于概念，概念是语义层面的处理单位，而术语只是语法层面的处理单位，因此在大多数本体学习系统中直接将抽取的术语作为概念并不完全恰当。多个不同的术语可以表达同一个概念语义，这些术语之间构成了同义词关系，例如术语“电脑”和“计算机”实际上表达是同一个概念，它们互为同义词。同义词的实现可以通过基于术语相似度的方法、基于语义解析的方法、基于语境的方法或基于统计的方法。基于术语相似度的方法首先计算术语之间的相似度，相似度越高，则术语之间构成同义词的可能性越大。术语相似度的计算方法有字面相似度计算、词素相似度计算[17]或基于词典的语义相似度计算[18]，如 Wordnet、HowNet、同义词词林。OntoLearn 系统借助 WordNet 对获取的术语进行语义解释，构造概念森林，较好地解决术语和概念间关系，OntoLearn 是少数能够区分术语与概念的本体学习系统之一。除了可以使用基于术语相似度的方法、基于语义解析和基于语境的方法识别同义词外，还可通过潜在语义分析（Latent Semantic Analysis, LSA）方法、逐点互信息的信息检索法和术语相关熵等统计方法进行识别术语之间的同义词关系。

2 概念间继承关系获取方法

概念间继承关系，又称分类关系或上下位关系，它是领域概念之间的一种最基本的重要关系，它和领域概念一起构成了领域本体的骨干。继承关系也是本体学习中研究的最为广泛的一种概念间关系，常见的继承关系获取方法可分为：基于语境的方法、基于语言学的方法、基于统计的方法、基于词典的方法和混合方法。

2.1 基于语境的方法

基于语境的方法的基本思想是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。Hearst[19]利用手工构造了“ such *NP* as {*NP*,}*{(or | and)} *NP* ”，“ *NP* {,*NP*}*{,} or other *NP* ”等 6 个上下位关系的词汇句法语境从百科全书的英文语料中提取了 152 个概念间上下位关系。基于语境的方法的缺点是准确度较低，因为大量无用的概念对往往也会匹配这些模式，另外语境的完备性对于获取效果影响较大。另外，很多研究人员从不同的角度，如附加语境、采用提问语料、基于 Web 的语境匹配、精度及覆盖度的改进、语境学习等，对 Hearst 的方法做了扩展以概念间获取继承关系。国内的研究人员方卫东等人[15]作了类似研究，通过语境“ <某些>*N0*<如>*N1*[*N2*,...,<及|或>*Ni*<等>] ”来提取概念间的显式的 is-a 关系。方卫东等人还使用了分布语义假设来发现概念间潜在的继承关系，其基本假定是：两个在语义上相近的概念，与它们共同出现的词的规律(主题签名) 和它们所处的上下文(上下文签名) 也必定相似，分布语义模型可被看作是一种基于统计的方法。

2.2 基于语言学的方法

基于语言学的方法通过语形分析、句法分析、依存结构分析以及语义分析等来获取概念间继承关系，其特点是抽取概念间继承关系准确率高，但不够强壮和效率低。本体学习系统 OntoLT[20]、OntoLearn 及 Text2Onto 中均用到了基于语言学的方法。OntoLT 系统的研究人员 Buitelaar 等人对语料进行语言学分析与标注，然后通过定义的映射规

则将标注得到的语言学实体映射成概念和关系[20]，其中规则“ HeadNounToClass_ModToSubClass ”能够将标注实体中的主要名词映射为类（对应上位概念），将主要名词及其修饰词的组合映射为子类（对应下位概念），该规则的基本思想是修饰词限定了被修饰名词的意义。例如“ 国际信用卡 ”是一个被标注实体，标注主要名词“ 信用卡 ”，修饰词位“ 国际 ”，由此可得到 is-a（“ 国际信用卡 ”，“ 信用卡 ”）。OntoLearn 系统中应用语言学启发式方法来获取概念间的继承关系，通过解析概念术语定义的句法词性规则抽取 is-a 关系。该系统中语言学方法是由 Missikoff 和 Navigli 等人[1]提出的，他们提出利用机器学习技术基于已有的通用本体对抽取出来的术语进行语义解释，即为这些术语关联上明确的概念标识符；然后，基于这些语义解释来确定概念之间的继承和相似关系，生成一个领域概念森林。与其它方法相比，该方法的主要特点是对术语进行语义解释，然后使用这些语义解释来获取除继承关系以外的其它概念间的关系，而其它方法都是将术语等同于领域概念。

2.3 基于统计的方法

基于统计的方法的共同的主要思想是词语的语义特性由它在不同上下文的分布来反映，因此词语的含意可以通过共现词语及共现频率来描述。目前研究的较多的概念聚类方法和关联规则方法本质上都是属于统计方法的范畴。

基于概念聚类的统计方法就是通过概念间的相似度或其它准则对概念进行聚类，同一类簇中的概念具有相近似的关系，它既可用来发现概念间的继承关系和其它关系，当使用层次聚类方法进行概念聚类时得到的结果就是概念间继承关系。概念间相似度可以有多种度量，比如余弦距离、几何距离、相对信息熵、互信息、语义距离等等，相似度的度量一般是通过统计信息计算而来的。通过概念层次聚类方法获取概念间继承关系的研究有很多，例如，Fisher[21]提出的一种基于矢量的聚类方法，Emde 等人[22]提出的基于 FOL 的聚类方法。The Mo ’ KWorkbench 采用无监督机器学习的聚类方法从文本中获得概念层次。这些方法不足之处是只能得到概念间严格的层次关系，然而在本体中一个概念却可以有多个父概念。Faure 等人采用宽度优先的方法对概念进行逐层聚类，该方法的特点在于它在进行每层聚类的时候都要考虑除当前簇的父簇外的所有簇，而不管这些簇所在的层次，能够较好的解决一个概念有多个父概念的问题[23]。形式概念分析(FCA, Formal Concept Analysis)是应用数学的一个分支，它建立在概念和概念层次的数学化基础之上。FCA 使用二元关系来表达领域中的形式背景(Formal Context)，从中提取概念层次结构，即概念格，从数据集中生成概念格的过程实际上是一种概念聚类的过程。Cimiano 等人提出了基于形式概念分析的概念聚类算法，并与层次合并聚类算法和二分 K 均值聚类算法做了比较，在给定的数据集上实验结果，要优于后两种算法[24]。

基于关联规则的统计方法的基本思想是：如果两个概念经常出现在同一文档中，则这两个概念之间必定存在关系。TextToOnto 中采用关联规则学习算法来发现概念之间的非分类关系。Stephens[25]首先采用一种词汇关联度的方法来提取含有潜在关系的基因对，然后利用同义词辞典来给出基因对之间的关系。

此外，Sanderson 等人[26]提出了术语包含的概率统计方法，基本思想是：对于两个术语 *t1* 和 *t2*，如果 *t2* 出现的文档集合是 *t1* 出现的文档集合的子集，那么 *t1* 包含 *t2*，然后利用这种包含关系来获取术语间继承关系。Fotzo 等人[27]作了更进一步的扩展，认为如果术语两个术语 *t1* 和 *t2* 满足以下条件：$P(t1|t2)>\theta$ 且 $P(t2|t1) < P(t1|t2)$，其中 θ 为阈值，$P(t1|t2)$ 和 $P(t2|t1)$可以根据根据条件概率和极大似然估计法利用术语在文档中出现的频率进行计算，进而抽取术语间的泛化/特化关系（即概念间继承关系）。

总体而言，基于统计的方法具有语言依赖性低、普适性强等特点，是目前研究的主流，但最大的缺点就是容易产生数据稀疏现象。

除了上面介绍的三种基于语境、基于语言学和基于统计的方法外，还可以采用基于词典的方法获取概念间继承关系，它往往从一些现有的词汇词典中定义的同义词、近义词和反义词等知识来获取本体中概念间的关系。例如，Nakaya 等人[28]使用 WordNet 来获取概念间的继承关系。混和方法往往是上述若干种方法的综合使用，以期得到更好的结果。

3 概念间一般关系获取方法

在领域本体中除了存在继承关系外，还存在其它一般性关系，例如属性关系、部分整体关系、因果关系及其它领域关系等。与继承关系获取研究相比，概念间一般关系获取研究则要少得多。下面对现有的研究做简单介绍。

3.1 属性关系获取

概念内涵是属于这个概念的所有对象所共有的属性集，因此属性关系对于概念理解至关重要。

Guarino 认为属性分为两类[29]：关系型属性和非关系型属性。关系型属性包括性质（比如颜色、位置等）和关系社会角色（比如儿子、配偶等）。非关系型属性包括部件（如车轮和发动机等）。Pustejovsky [30]在通用词法理论中也对属性作了分类，认为属性分为四类：外观角色、构成角色、目的角色和施事角色。外观角色指明对象是什么类型，以书为例，书有属性形状和颜色等。构成角色对象的由什么材料或部件构成，例如纸张、章节是书的构成角色属性。目的角色对象的目的，例如读是书的目的角色属性。施事角色指明该对象如何被创建，例如写是书的施事角色属性。

目前，国内外关于属性关系学习的相关研究并不多见，从公开可查阅的文献中，仅有英国 Essex 大学的 Poesio 和 Almuhareb 对属性学习作了研究[31]。他们首先采用基于模版的方法从从 Web 网页中抽取含有候选属性的实验数据。结合上面两种属性分类知识，认为实验数据分为六类：性质、部件、相关对象、活动、施事者和非属性，并使用语形信息、属性模型、问题模型和属性使用模型四类信息构造了分类器对实验数据进行自动分类。总体而言，Poesio 和 Almuhareb 在英文属性学习的研究中取得了一定的成果。但是其不足之处在于上述关于属性分类是广义上的，其范畴大于本体中关于属性的理解，因此不能完全满足属性关系学习的需要。

3.2 部分整体关系获取

部分整体关系的获取主要采用模式匹配方法。Charniak 等描述了在大量语料中发现部分与整体关系的模式[32]。Text2Onto 系统中开发的 JAPE 模式，通过计算部分与整体关系的模式共现率指示概念术语之间部分与整体关系的概率。文献[33]也对概念间的部分整体关系做了研究。另外可借助通用本体如 WordNet 中的语义关系推理概念术语间的部分与整体关系。董振东先生设计的中文知识库——知网中，也包含有大量的部分整体关系。

3.3 其它关系获取

其它领域关系一般是专门存在于特定领域中的关系。例如在商品销售中的购买者与商产之间的购买关系，在新闻报道中的原因事件关系等等。其它领域关系可通过语言结构分析法、关联规则的方法获取。例如在 OntoLT 系统中定义了“SubjToClass_PredToSlot_DobjToRange”的规则发现领域关系，其思想是将主语映射为类，谓语映像为该类的相应属性，直接宾语映像为属性的范围值。TextToOnto 系统中通过次范畴框架获取的方式发现领域关系，比如由“爱（男人,女人）”、“爱（小孩,母亲）”、“爱（小孩,父亲）”可推导出人与人之间存在爱的关系，即“爱（人，人）”。国内的方卫东等[15]利用关联规则方法获取领域关系。

4 公理获取方法

目前查到的只有 Shamsfard 等人[34]提出的基于模板的抽取方法。该方法在对句子结构分析的基础上，应用预先定义的蕴涵本体公理的模板，如果与模板匹配，则得到相应的本体公理。该方法是公理获取方法的一个积极尝试，但局限性很明显，它不仅需要人工预先制定模板，而且无法获取隐式的公理。

5 结束语

面向文本的本体学习是目前本体学习研究的重点。本文对国内外面向文本的领域概念学习方法、概念间继承关系学习方法、属性关系学习方法、整体部分关系学习方法以及公理获取方法等各种本体学习方法的进行分析与评述，指出了各种方法的优缺点与适用范围。

参考文献

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

[2] Bourigault ， D. Surface Grarnmatieal Analysis for the Extraetion of Terminologieal Noun Phrase. Proeeedings of COLING92,pp977-981.

[3] Justeson,J.S. and Katz,S.M. Teehnlcal Terminology: Some Linguistic Properties and an Algorithm for Identifieation in Text. Natural Language Engineering,1995,l(1)9-27.

[4] A. Smeaton, Quigley. Experiments on using semantics distances between words in image caption retrieval. In Proc. of 19th International Conrerence on Research and Development in Information Retrieval, Zurieh, Switzerland,1996.

[5] Shamsfard M., Barforoush A.A. Learning ontologies from natural language texts.

International Journal of Human Computer Studies ,2004,60 (1) :17-63.

[6] Morin E. A utomatic acquisition of semantic relations between term s from technical co rpo ra [A]. P roc 5th Int Congress on Term ino logy and Know ledge Eng (TKE'99) [C]. V ienna: TermN et, 1999.

[7] Maedche A., Staab S. Discovering conceptual relations from text. In: Horn W, ed. Proc. of the ECAI 2000. Amsterdam: IOS Press, 2000:321-325.

[8] Salton.G, Yng.C.S and Yu.C.T. A Theory of Term Importance in Automatic Text Analysis. Journal of the Ameriean Soeiety for Information Seience,1975,26(l):33-44.

[9] Damerau F.J. Evaluating Domain-oriented Muiti-Word Terms from Text.Information Proeessing and Management, 1990, 29(4):433-447.

[10] Cohen.J.D. Highlights: Lansuage- and Domain-Independent Automatic Indexing Terms for Abstracting. Journal of the American Soeiety for Information Seienee, 199546(3):162-174.

[11] K.T. Frantzi, S. Ananiadou.The C-Value/NC-Value domain independent method for multi-word term extraction. Journal of Natural Language Processing, 6(3):145-179,1999.

[12] Patrick P. and Dekang Lin. A Statistical Corpus-Based Term Extractor. Canadian Conference On AI2001,pp36-46.

[13] 陈文亮,朱靖波,姚天顺.基于 Bootstrapping 的领域词汇自动获取. 第 7 届全国计算语言学联合学术会议论文集(JSCL 2003).北京:清华大学出版社,2003.67−72.

[14] 梁健,吴丹. 种子概念方法及其在基于文本的本体学习中的应用[J]. 图书情报工作,2006,50(9):18-21.

[15] 方卫东,袁华,刘卫红. 基于 Web 挖掘的领域本体自动学习. 清华大学学报(自然科学版), 2005,45(S1): 1729-1733.

[16] Paola V, Michele M, Roberto B. Identification of Relevant Terms to Support the Construction of Domain Ontologies[C]. ACL-EACL Workshop on Human Language Technologies, Toulouse, France, July 2001.

[17] 朱毅华. 智能搜索引擎中的同义词识别算法研究:[学位论文]. 南京:南京农业大学,2001.

[18] 刘群,李素建.基于《知网》的词汇语义相似度计算 [J].Computational Linguistics and Chinese Language Processing,2002,(2):59-76.

[19] Marti A. Hearst, Automatic Acquisition of Hyponyms from Large Text Corpora. Proceedings of the 14th International Conference on Computational Linguistics. Nantes France. 1992, pp 539-545.

[20] P. Buitelaar, D. Olejnik, M. Hutanu, A. Schutz, T. Declerck, M. Sintek Towards Ontology Engineering Based on Linguistic Analysis. In: Proceedings of LREC, 2004.

[21] D. Fisher, Knowledge acquisition via incremental conceptual clustering. Machine Learning 2, pp. 139-172, 1987.

[22] Emde W, Wettschereck D. Relational instance-based learning. In: Saitta L, ed. Proc. of the ICML'96. San Francisco: Morgan Kaufmann Publishers, 1996. 122−130.

[23] Faure D, Nedellec C. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In: Velardi P, ed. Proc. of the LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications. Granada: LREC, 1998. 5-12.

[24] Cimiano P,Hotho A,Staab S.Learning concept hierarchies from text corpora using formal concept analysi-s.Journal of Artificial Intelligence Re-search,2005,24:305 ～339.

[25] Matthew Stephens, Mathew J. Palakal, Snehasis Mukhopadhyay, Rajeev R. Raje, Javed Mostafa, Detecting Gene Relations from MEDLINE Abstracts. Pacific Symposium on Biocomputing. Hawaii. 2001, pp 483-496.

[26] Sanderson M,Croft W B.Deriving concept hierarchies from text[C]. Proceedings of SIGIR,1999::206-213.

[27] H.N. Fotzo, P. Gallinari, Learning Generalization/Specialization Relations

between Concepts - Application for Automatically Building Thematic Document Hierarchies, In: Proceedings of RIAO, 2004.

[28] Nakaya N, Kurematsu M, Yamaguchi T. A domain ontology development environment using a MRD and text corpus. In: Proc. of the Joint Conf. on Knowledge Based Software Engineering. 2002.

[29] Guarino, N. Concepts, attributes and arbitrary relations: some linguistic and ontological criteria for structuring knowledge base. Data and Knowledge Engineering, 8, (pp. 249–261).

[30] Pustejovsky, J. (1995). The generative lexicon. MIT Press.

[31] Poesio, M. and Almuhareb, A. (2005). Identifying Concept Attributes Using a Classifier. In Proc. ACLSIGLEX Workshop on Deep Lexical Acquisition. Ann Arbor, USA, June.

[32] E. Charniak, M. Berland. Finding parts in very large corpora. In: Proceedings of the 37th Annual Meeting of the ACL, pp. 57-64, 1999.

[33] Willem Robert van Hage,Hap Kolb, and Guus Schreiber. A Method for Learning Part-Whole Relations. Proceedings of the 5th International Semantic Web Conference Athens, Georgia, USA, November 5-9,2006.

[34] Maedche A., G. Neumann, S. Staab. Bootstrapping an Ontology- Based Information Extraction System, Studies in Fuzziness and Soft Computing, Intelligent Exploration of the Web, Springer, 2002.