文章编号:1001-5132（2008）01-0062-06

一种基于 Web 的分类体系学习算法

刘柏嵩[1]，贺赛龙[2]

（1.宁波大学 网络中心，浙江 宁波 315211；2.宁波大学 学报编辑部，浙江 宁波 315211）

摘要：领域分类结构的抽取已成为本体工程和本体学习的关键部分，提出一种新的分类结构学习算法，将 Web 作为知识获取的语料库，运用迭代方法抽取相关语言学模式，再利用语言学模式抽取分类结构，并采用改进的互信息方法对结果进行评价和过滤，最后通过实验对该分类学习算法的性能进行评价. 实验表明：算法具有良好的跨领域性，在准确率和召回率方面也有改善.

关键词：本体；分类学习；语言学模式；机器学习

中图分类号：TP316 文献标识码：A

本体是实现语义 Web 的基石，目前大多数是由专家人工构建本体，既费时又费力. 因此(半)自动化的本体学习方法成为实现语义 Web 的急迫和重要研究领域. 本文主要讨论本体工程中的关键部分——分类结构(Taxonomy)，它是特定领域中的核心本体，它将该领域的词汇从根节点到叶子节点，用渐增的特征性联系起来. 分类学习的目标是自动建立特定领域的本体结构，用于表述该领域内特定的概念和语义关系.

分类关系的抽取目前主要有 3 种类型的方法：(1)基于人工或自动创建的正则表达式的方法[1-3]；(2)基于统计和词汇特征向量的方法[4-8]；(3)基于词典分析的方法[9-12]. 但这些方法都存在一些缺陷：基于正则表达式的方法很难兼顾准确率和召回率；统计和机器学习方法大多都基于术语上下文特征的分析和比较，因为基于统计方法学习到的分类关系包含大量噪音数据和异构数据，且计算开销很大，因此通过这些方法获取的分类结构很难进行人工评价；词典分析的明显不足是存在迂回定义和定义过于抽象等问题. 此外，多数方法都只注重分类学习的单一阶段，几乎没有知识抽取价值链(词汇—关系—分类)的完整方法. 另外训练集的有效性不是通用的，需要专业注解者参与. 最后，系统性能通常需要由领域专家进行人工评价，效率很低.

为克服单一学习方法的局限，本文采用一种基于 Web 的结合语言学技术和统计分析的混合分类学习方法. 该方法将 Web 作为知识获取的语料库，依据领域主题检索最具领域代表性的网站，利用迭代方法抽取相关语言学模式，再利用语言学模式抽取分类关系，并采用改进的互信息方法对结果进行评价和过滤，最后通过实验对该分类学习算法的性能进行评价.

1 相关定义

分类关系一般包括 IS-A 关系、PART-OF 关系

收稿日期：2007-10-30. 宁波大学学报（理工版）网址：http://3xb.nbu.edu.cn
基金项目：浙江省自然科学基金（Y105625）；浙江省哲社规划课题（NM05GL02）；浙江省教育厅科研项目（20071008）；宁波大学人才工程项目（XJ0719003）.
第一作者：刘柏嵩（1971－），男，安徽安庆人，博士/研究馆员，主要研究方向：人工智能及计算机网络. E-mail: lbs@nbu.edu.cn

和 INSTANCE-OF 关系，并作如下定义.

定义 **1**　层级关系. 根据概念间的包含关系，可将概念区分为上位概念和下位概念. 上位概念称为大概念，下位概念称为小概念. 按同一标准或同一维度划分并处于同一层面的概念称为并列概念. 它主要指属种关系，即概念外延的包含关系，小概念(种)的外延是大概念(属)外延的一部分. 小概念除了具有大概念的一切特征外，还具有本身独有的区别特征. 比如(属)－树；(种)－乔木、灌木.

定义 **2**　核心本体. 其被定义为某种结构 $H=(h,\prec)$，包括一组概念集 h 和 h 的偏序 $\prec$，称之为概念层次或分类体系.

定义 **3**　父结点. 在核心本体 O, $H=(h,\prec)$ 中，如果 $x, y\in h$, $x\prec y$，且不存在概念 $z\neq x, y$，满足 $x\prec z$ 且 $z\prec y$，则概念 y 称为概念 x 的父结点.

定义 **4**　超类. 对任意的类 C_1 和 C_2，如果 $\forall i$: 是实例 $(i, C_2)\rightarrow$ 是实例 (i, C_1)，即 $\mathrm{domain}(C_2)\subseteq \mathrm{domain}(C_1)$，则称 C_2 为 C_1 的子类，记为是子类 (C_2, C_1)；相应地，C_1 称为 C_2 的超类，记为是超类 (C_1, C_2).

定义 **5**　类的等价. 对任意的类 C_1 和 C_2，类 C_1 等价于 C_2，记为 $\mathrm{eqv}(C_1, C_2)$ 当且仅当是子类 $(C_2, C_1)\wedge$ 是子类 (C_1, C_2)，或者是超类 $(C_1, C_2)\wedge$ 是超类 (C_2, C_1).

定义 **6**　上位节点. 给定一分类系统，其中 $(C, C_1, \cdots, C_n)$ 为其节点. 如果 $\forall i$: 是成员 $(i; C_i)\rightarrow$ 是成员 $(i; C)$，则 C 称为 $(C_1, \cdots, C_n)$ 的上位节点，记为是上位节点 $(C, C_1, \cdots, C_n)$.

2　基于 Web 的混合分类学习算法

在处理大规模语料时，如果单纯采用统计方法会使得计算开销非常大. 考虑到基于词汇句法模式学习方法的特点，在计算效率和准确率上都有明显优势，但在模式学习的过程中经常会出现数据稀疏等问题. 本文提出一种新的基于 Web 的混合分类学习算法 WHTLA(Web-based Hybrid Taxonomy Learning Algorithm)，采用 Web 作为学习语料库来减少数据稀疏问题，并利用学习到的模式进行分类学习. 另外对于学习结果的评价是整个学习过程中非常重要的一步，本文采用基于 Web 搜索引擎的统计方法来实现自动评价，克服了当前大多数方法都必需由人工专家评价的缺陷.

学习算法的具体步骤如下：

输入：领域标注语料库和一个初始语义关系，如上下位关系；输出：领域相关的分类结构.

步骤 1　收集由初始语义关系关联的关系术语对. 即从标注语料库中抽取初始语义关系关联的术语对，如从“ anatomy IS-A subject ”中抽取关系术语对(anatomy，subject).

步骤 2　发现包含概念相关术语(即前面步骤中抽取的术语对)的句子. 这些句子都经过词元解析(lemmatized)并从中识别名词短语. 因此，句子可以作为词汇句法表达式，通过简化处理得到句子的更泛化的表达式，句子间的比较也更容易. 如可利用步骤 3 中抽取的 HYPERNYM(subject，anatomy)，从语料库中发现句子“ The teacher teach the subjects, such as anatomy, physiology, biochemisttry, pathology, microbiology. ”该句子可以简化为词汇句法表达式：NP teach NP such as LIST，其中 NP 表示名词短语，LIST 表示连续的名词短语.

步骤 3　对步骤 2 中抽取的词汇句法表达式进行归纳，发现它们之间共同部分. 该共同部分是通过相似度计算和聚类得到的，可将它作为候选词汇句法模式.

步骤 4　由人工专家对候选词汇句法模式进行检验.

步骤 5　利用新词汇句法模式抽取更多的候选术语对.

步骤 6　由人工专家对候选术语对进行检验，返回到步骤 2，直到获取了足够的词汇句法模式.

步骤 7　利用获得的词汇句法模式从语料库中

抽取分类关系.

步骤 8　通过对搜索引擎查询的统计，对抽取的分类关系进行评价并过滤掉置信度较低的分类关系. 利用搜索引擎命中次数，计算分类结构中术语间的互信息. 例如利用词汇句法模式抽取到“ Ningbo IS-A city ”，如果“ Ningbo ”和“ city ”间的互信息值越高，则说明“ Ningbo ”越有可能是“ city ”的下位词.

该算法的核心部分主要是步骤 3 和步骤 8，即以 web 作为语料库，通过相似度计算和聚类，归纳出语言学模式. 接着利用语言学模式抽取候选分类关系，最后采用改进的互信息方法对候选概念进行置信度计算和过滤.

3　算法实现

3.1　Web 文档收集和预处理

首先利用网络蜘蛛搜集特定领域的网页，形成一个领域相关的 web 文档集. 由于 web 数据资源的异质性和存在大量无用的数据，需要对其进行大量的清理，才能进行后续的操作，如将 web 文档转化成纯文本格式，删除图像、音频、视频和 flash 等数据. 接着进行词性标注处理，形成一个大规模的标注语料库.

3.2　识别关系术语对

通常假定动词指示概念间的语义关系，因此语义关系组Concept，Relation，Concept可以词汇化为Noun_1，Verb，Noun_2，其中Noun_1和Noun_2是文本中的名词术语，Verb是文本中的动词术语，Noun_1是Verb的主题(主语)，Noun_2是Verb的对象. 因此就能非常方便地在标注语料库中识别关系术语对，即识别Noun_1，Noun_2，其中名词和动词术语的正则表达式分别为：

Noun: (DET)?(JJ)*(NN|NNS|NNP|NNPS)$^+$，

Verb: (VB|VBD|VBN|VBZ)$^+$，

其中 JJ 表示形容词，NN，NNS，NNP 和 NNPS 表示名词，DET 表示冠词，VB，VBD，VBN 和 VBZ 表示动词.

由于基于相同的关系(动词)可能会从语料库中抽取大量关系术语对，因此可通过对出现术语对频数的统计，设定了 1 个阈值，将频数过低的术语对过滤掉，提高可信度.

3.3　生成词汇句法模式

在步骤 2 中通过 HYPERNYM(subject，anatomy)获得了表达式 NP teach NP such as LIST. 类似地，通过 HYPERNYM(application，Word)从句子“ This includes related applications such as Word, Excel, Powerpoint, FrontPage, and can choice the language by yourself. ”中可以获得以下词汇句法表达式：NP such as LIST can choice NP. 将以上表达式抽象为：

$$A = A_1A_2\cdots A_j\cdots A_k\cdots A_n \text{ with } \begin{cases}\text{RELATION}(A_j, A_k),\\ k > j+1.\end{cases}$$

$$B = B_1B_2\cdots B_{j'}\cdots B_{k'}\cdots B_{n'} \text{ with } \begin{cases}\text{RELATION}(B_{j'}, B_{k'}),\\ k > j+1.\end{cases}$$

假设 $\text{Sim}(A, B)$ 为词汇句法表达式 A 和 B 间的相似度函数，并设定

$$\begin{cases}\text{Win}_1(A) = A_1A_2\cdots A_{j-1},\\ \text{Win}_2(A) = A_{j+1}\cdots A_{k-1},\\ \text{Win}_3(A) = A_{k+1}\cdots A_n,\end{cases}$$

和

$$\begin{cases}\text{Win}_1(B) = B_1B_2\cdots B_{j'-1},\\ \text{Win}_2(B) = A_{j'+1}\cdots A_{k'-1},\\ \text{Win}_3(B) = A_{k'+1}\cdots A_{n'}.\end{cases}$$

则 $\text{Sim}(A, B) = \sum_{i=1}^{3}\text{Sim}(\text{Win}_i(A), \text{Win}_i(B))$.

将 $\text{Sim}(A, B)$ 的值预定义为最长的共同字符串，通过两两比较可得出的候选词汇句法模式为：NP such as LIST. 最后通过多次迭代，主要获得以下词汇句法模式：

(1) HYPERNYM(,)? such as (NP3|NP2|NP1)；

(2) NP4 and other HYPERNYM；

(3) NP4 or other HYPERNYM；

(4) HYPERNYM, especially (NP3|NP2|NP1)；

(5) HYPERNYM, including (NP3|NP2|NP1)；

(6) such HYPERNYM as (NP3|NP2|NP1)；

(7) HYPERNYM like (NP3|NP2|NP1)；

(8) (NP3|NP2|NP1) is HYPERNYM；

(9) (NP3|NP2|NP1), another HYPERNYM；

(10)HOLONYM's (NOUN)；

(11)(NOUN) of DET [JJ | NN]* HOLONYM；

(12)(NOUN) in DET [JJ | NN]* HOLONYM；

(13)(NOUN) of　HOLONYM；

(14)(NOUN) in HOLONYM；

(15)NOUN = (NN|NNS|NP|NPS)；

(16)NP1 = (DET)?(JJ |JJR |JJS)*(NOUN)* NOUN；

(17)NP2 = NP1 (and|or) NP1；

(18)NP3 = NP1(, NP1)+ (and|or) NP1；

(19)NP4 = NP1(, NP1)*.

其中 HYPERNYM 模式用于匹配包含有上位关系的词；NP1～NP4 匹配相关的下位词；HOLONYM 模式用于匹配包含有整体关系的词；NOUN 匹配相关的名词.

3.4　词汇句法模式评价

在获得了词汇句法模式后，需要对这些模式进行评价，因为这些模式质量的好坏对后面的分类学习是至关重要的，可采用下面的模式置信度公式对产生的词汇句法模式进行评价.

$$\mathrm{Conf}(P)=\frac{P_{\mathrm{positive}}}{P_{\mathrm{positive}}+P_{\mathrm{negative}}},$$

其中，P_{positive} 表示模式 P 匹配正例的数量；P_{negative} 表示模式 P 匹配反例的数量. 同时也可通过估计词汇句法模式的覆盖率来评价模式，采用 Riloff 提出的评价算法 $\mathrm{Conf}_{R\log F}(P)$ 为：

$$\mathrm{Conf}_{R\log F}(P)=\mathrm{Conf}(P)\cdot\log_2(P_{\mathrm{positive}}).$$

3.5　利用模式进行分类学习

利用模式进行分类学习主要通过扫描自然语言文本，进行模式匹配来抽取相关信息，基本思路为：考察在每次尝试开始时，正文中正在扫描的字符 y_i，它不但与前面的 $m-1$ 个字符 $y_{i-m+1}y_{i-m+2}\cdots y_{i-1}$ 有关，而且还与其后的 $m-1$ 个字符 $y_{i+1}y_{i+2}\cdots y_{i+m-1}$ 有关系. 当前扫描的字符与其前后各 $m-1$ 个字符构成大小为 $1+(m-1)+(m-1)=2m-1$ 的窗口. 在这个窗口中，首先使用后缀自动机 SA 从窗口中间位置反向扫描模式前缀，然后使用正向有限状态自动机 FFA(Forward Finite Automata)从中间位置往后扫描相应的模式后缀. 这样就能够完全将窗口中心位置附近与之有关的有用信息全部扫描出来，在窗口移动前查找出正文中包含窗口中心位置字符的所有模式出现，减小窗口移动时的信息损失.

3.6　对候选分类关系的评价

为了提高抽取的准确率，对抽取到的候选分类关系进行评价，计算候选分类关系的置信度并过滤掉置信度较低的分类关系. 在自然语言处理领域的研究表明：在大规模语料库中，统计术语的同现率能显示术语间的关联度. 本文采用互信息方法来计算分类关系的置信度 $\mathrm{Conf}(T)$，公式为：

$$\mathrm{Conf}(T)=MI(X;Y)=\log_2\left(\frac{P(x,y)}{P(x)P(y)}\right),$$

其中，$P(x,y)$ 表示 x 和 y 同时出现的概率；$P(x)$ 和 $P(y)$ 分别表示 x 和 y 单独出现的概率. 由于函数 $Y=\log_2(X)$ 是单调递增的，因此 $\mathrm{Conf}(T)$ 可以简化为：

$$\mathrm{Conf}(T)=\frac{P(x,y)}{(P(x)P(y))}.$$

我们采用搜索引擎命中数(即对特定查询的有效返回结果数)来计算术语在领域相关网页中同现频率，并将 $\mathrm{Conf}(T)$ 改进为：

$$\mathrm{Conf}(T)=\frac{|\mathrm{hits}(x+y)|}{|\mathrm{hits}(x)|},$$

其中，x 表示分类结构中的下位词，y 表示关系短语，如假设 $x=$“Ningbo”，$y=$“city of x”，则 $x+y=$“city of Ningbo”，$|\mathrm{hits}(x+y)|$ 表示短语 $x+y$ 在搜索引擎中的返回数，$|\mathrm{hits}(x)|$ 表示术语 x 在

搜索引擎中的返回数.

4 实验分析

4.1 实验语料

为检验本文提出方法的有效性，选择新闻领域和高等教育领域的语料进行实验. 其中新闻领域的语料(news_corpus)来自 http://english.sohu.com/，该语料选取了搜狐新闻网站中 120 个文档；高等教育领域的语料(edu_corpus)来自 http://www.harvard.edu/，该语料库包含哈佛大学网站中的 150 个文档.

4.2 实验评估方法

对于实验的评估方法，我们采用在 IE 领域广泛使用的准确率(*Precision*)、召回率(*Recall*)和 *F* 指数(*F-measure*). 准确率指正确抽取的概念占所有抽取概念的百分比，召回率指抽取的概念占语料库中所有概念的百分比，*F* 指数指召回率和准确率的加权几何平均值，具体计算公式如下：

$$Precision = correct_{\text{extracted}} / all_{\text{extracted}},$$

$$Recall = correct_{\text{extracted}} / all_{\text{corpus}},$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}.$$

同时基于相同的实验语料，本文将基于Web的混合分类学习算法的运行结果，与Hearst模式方法和决策树方法相比较. 与Hearst模式方法和决策树方法进行比较的主要原因是：(1)Hearst模式[10]方法是基于符号学习的典型代表；(2)决策树方法[11]是机器学习领域的一个经典算法，同时分类结构也类似于树型结构.

4.3 实验结果

本实验条件为 Windows XP 操作系统，CPU 主频 3GHz，内存 1G，算法都使用 java 语言实现. 本文分别在新闻和高等教育 2 个不同领域对混合算法、Hearst 模式方法和决策树方法进行实验，具体实验结果如图 1～图 6 所示. 结果表明：混合算法 WHTLA 在不同领域、不同指标上都高于其他算法. 这是由于该混合算法以 Web 为学习语料库，建立了一个较全面的模式库，确保了在召回率方面的性能. 同时混合算法还对学习到的结果进行了基于 Web 搜索引擎的统计分析，进一步提高了准确率.

5 结论

本体学习是当前计算机科学领域的一个研究

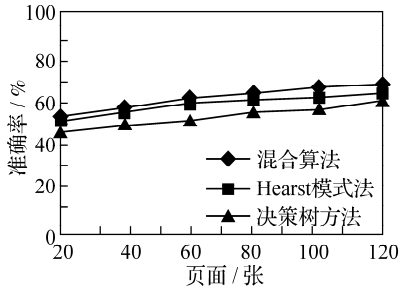

图 1 在新闻领域的准确率对比

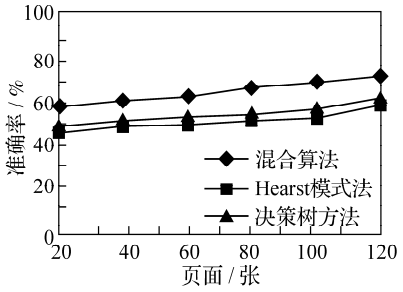

图 2 在新闻领域的召回率对比

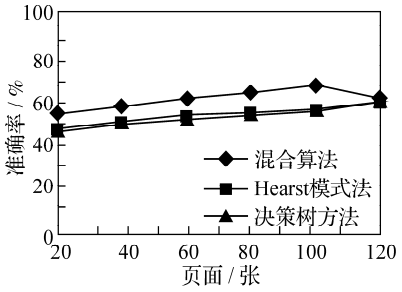

图 3 在新闻领域的 *F* 指数对比

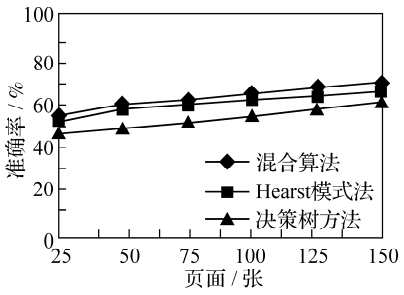

图 4 在高等教育领域的准确率对比

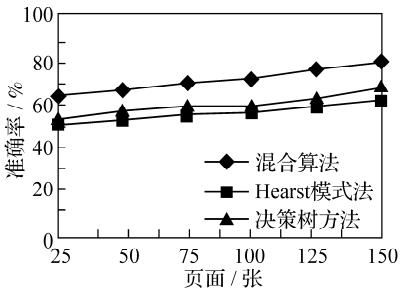

图 5 在高等教育领域的召回率对比

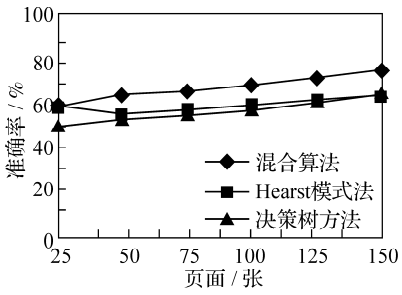

图 6 在高等教育领域的 *F* 指数对比

热点，分类结构是本体的骨架. 本文介绍了一种新的基于 Web 的分类学习算法 WHTLA，该算法首先利用信息抽取技术从 Web 中进行模式学习，以克服数据稀疏问题. 然后利用学习到的模式进行分类关系学习，最后采用基于搜索引擎的统计对学习结果进行评价，克服了必需由人工专家评价的局限. 该算法的跨领域性得到充分证明，在不同的语料库上都表现出良好的性能. 通过实验在与其他分类学习算法的比较中，也证实了该算法在准确率和召回率上都有所提高.

参考文献：

[1] Hearst M. Automatic acquisition of hyponyms from large text corpora[C]//Proc of 14th COLING. France: Nantes, 1992.

[2] Oakes M P. Using hearst's rules for the automatic acquisition of hyponyms for mining a pharmaceutical corpus[M]. RANLP Text Mining Workshop, 2005.

[3] Snow R, Jurafsky D, Ng A Y. Learning syntactic patters for automatic hypernym discovery[C]//Proc Conf Neural Information Processing Systems, NIPS 17, 2005.

[4] Widdows D. Unsupervised methods for developing taxonomies by combining syntactic and statistical information[C]//HLT-NAACL 2003. USA: Edmonton, 2003.

[5] Paola Velardi, Alessandro Cucchiarelli, Michaël Petit. A taxonomy learning method and its application to characterize a scientific web community[J]. Knowledge and Data Engineering, 2007, 19(2):180-191.

[6] 刘柏嵩，高济. 通用本体学习框架研究[J]. 东南大学学报, 2006, 22(3):381-384.

[7] 刘柏嵩. 基于 Web 的通用本体学习研究[D]. 杭州: 浙江大学, 2007.

[8] Cimiano P, Hotho A, Staab S. Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text[C]//16th ECAI. Spain: Valencia, 2004.

[9] Marta Sabou. Learning web service ontologies: an automatic extraction method and its evaluation[C]//ISWC, 2005.

[10] Hearst M A. Automated discovery of wordnet relations [M]. Cambridge: MIT Press, 1998.

[11] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986, 62(1):81-106.

[12] 杜小勇，李曼，王珊. 本体学习研究综述[J]. 软件学报, 2006, 17(9):1 837-1 847.

A Web-based Taxonomy Learning Algorithm

LIU Bai-song[1], HE Sai-long[2]

(1.The Network Center, Ningbo University, Ningbo 315211, China; 2.Editorial Office, Ningbo University, Ningbo 315211, China)

Abstract: Extraction of domain-specific taxonomies has been increasingly needed in ontology engineering. In this paper, a novel learning algorithm of taxonomy is presented. It uses the iterative methods to extract linguistic patterns, by which the taxonomic components are obtained. The improved mutual information method is adopted for evaluating and filtering the results. In the end, the performance of the taxonomy learning algorithm is assessed by the experiment and shows the good inter-disciplinary applications and improvement in terms of accuracy and recall-rate.

Key words: ontology; taxonomy learning; linguistics pattern; machine learning

CLC number: TP316　　**Document code:** A

（责任编辑　章践立）