第 34 卷 第 8 期 计 算 机 工 程 2008 年 4 月
Vol.34 No.8 Computer Engineering April 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)08—0229—03 文献标识码：A 中图分类号：TP18

一种基于 Web 的通用本体学习框架

刘柏嵩
(宁波大学网络中心，宁波 315211)

摘 要：提出一种通用的多策略本体学习框架，通过对 Web 上各专业领域文档集进行挖掘来实现本体自动构建。讨论本体学习中本体概念的抽取、概念之间语义关系的抽取和分类体系的自动构建等关键技术，通过实验对算法进行测试和评价。由于集成了多种机器学习算法，该方法在概念抽取和语义关系学习方面具有更高的准确性，采用通用本体 WordNet 和 HowNet 作为语料库，可适用于不同的专业领域。通过按需获取 Web 文档，该方法能实时生成本体。

关键词：本体；本体学习；本体评价；语义 Web

Web-based General Ontology Learning Framework

LIU Bai-song
(Network Center, Ningbo University, Ningbo 315211)

【Abstract】This paper proposes a General Ontology Learning Framework(GOLF), dicusses the key technologies of ontology learning such as domain concepts extraction, semantic relationships between concepts and taxonomy automatic construction and ontology evaluation methods. By integrating many machine learning algorithms, this approach suffers less ambiguity and can identify domain concepts and relations more accurately. By using generalized corpus WordNet and HowNet, the method is applicable across different domains. By obtaining source documents from the Web on demand, the method can produce up-to-date ontologies.

【Key words】ontology; ontology learning; ontology evaluation; semantic Web

1 概述

语义Web和语义网格(semantic grid)的实现很大程度上依赖于大量本体(ontology)的建立。本体已广泛地应用到很多领域，如信息检索、机器翻译、知识管理、电子商务和信息集成等[1-2]。相对于Internet上的海量信息而言，目前只有少量手工构建的通用本体，如WordNet和Cyc。本体构建是一个非常复杂的过程，需要多个领域的专家参与。虽然目前本体工程工具已经较为成熟，但本体的手工构造仍是一项繁琐而辛苦的任务，并最终导致所谓的知识获取瓶颈[1,3]。

为此，研究人员提出了本体学习(ontology learning)这一涉及人工智能中信息获取、机器学习、自然语言处理等多领域交叉的研究课题。本体学习技术是当前计算机领域的热点之一[1,4]，它旨在开发能够实现本体自动构建的机器学习技术来协助知识工程师构建本体，从而减少本体知识获取过程的成本。目前已经提出了很多本体学习方法，但大部分都不理想。就基于半结构化数据的本体学习来说，现有的方法往往是将其按照纯文本对待，没有充分地利用其隐含的结构信息。针对同一个学习目标，本体学习技术中的任意一种方法都有自己的适用范围，无法保证在所有情况下都得到好的学习结果。

鉴于上述问题，本文提出一种基于 Web 的多策略本体学习方法，将各种方法进行综合从而获得更好的学习结果。在现有的开源工具基础上，开发出一个稳定的、整合的、能够完成多种学习任务，且能处理中文源的本体学习工具——通用本体学习框架(General Ontology Learning Framework, GOLF)。

2 总体架构

GOLF 系统框架如图 1 所示，包括如下内容：文档和语料库预处理，抽取候选概念，术语选择，语义关系抽取，分类体系构建，本体构建和本体评价等。其主要目标包括从 Web 文档中自动获取领域术语及其相互关系，采用机器学习方法来确定概念对之间的语义关系，在获取的概念及其语义关系的基础上构建本体。

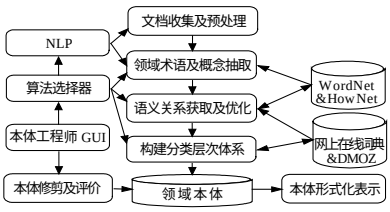


图 1 GOLF 系统框架

3 本体概念的抽取

概念的自动提取是构建本体的重要工作之一[5]。GOLF 系统结合语言学分析技术和统计分析方法，在概念抽取模块中集成了一个概念抽取算法库，包括Relative Term Frequency (RTF), TFIDF (Term Frequency Inverted Document Frequency),

基金项目：国家“973”计划基金资助项目(2003CB317000)；浙江省自然科学基金资助项目(Y105625)；浙江省社科规划基金资助项目(NM05GL02)
作者简介：刘柏嵩(1971－)，男，研究员、博士，主研方向：人工智能，语义 Web，数字图书馆
收稿日期：2007-05-23 **E-mail：**lbs@nbu.edu.cn

Entropy等学习方法，通过集成能够得到比单个学习器更强的泛化能力。对于复合概念的抽取，GOLF系统采用了C-value/NC-value方法，该方法结合了词汇句法模式和词频统计方法，并在自动抽取概念时能利用上下文信息。

在进行概念抽取时集成多种学习算法，能有效提高抽取质量。集成学习的基本思想使所有的学习算法都在某些方面有所偏置，而通过对多个不同算法的平均，可以有效地消除这些偏置。如何将多种不同的学习算法组合成一个集成系统是当前的难点问题[5]，本文采用组合概率分布模型定义了集成学习算法的组合框架，假设给定n个学习算法：

$$P(C \mid w, C_1^n) = f((P_i(C \mid w, C_1^n))_{i=1,2,\cdots,n}) \quad (1)$$

其中，C_i表示第i个学习算法的分类输出；f是一个组合函数。

对各学习算法的类别概率分布进行线性插值：

$$P(C \mid w, C_1^n) = \sum_{i=1}^{n} P(C \mid w, i, C_i) \cdot P(i \mid w) = \sum_{i=1}^{n} P_i(C \mid w, C_i) \cdot \lambda_i(w) \quad (2)$$

其中，权值$\lambda_i(w)$表征了对于词w的上下文，第i个学习算法在组合中的重要程度；$P_i(C|w,C_i)$表示在给定第i个学习算法对于词w的输出是C_i的情况下，正确分类结果是C的概率的估计。

4 语义关系学习

概念之间的各种语义关系是本体学习中的重要部分，语义关系可以分为分类关系和非分类关系。

4.1 分类关系学习

分类关系主要考虑上下位关系(is-a)和部分整体关系(part-of)。GOLF系统采用语义词典WordNet, HowNet及改进的Hearst模式匹配的方法来抽取分类关系。Hearst模式是一种基于模式的信息抽取方法[6-7]，它通过扫描自然语言文本中出现的特定模式来抽取相关信息，通过对Hearst模式的改进，得到了以下模式：

(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概念；NP1-NP4 匹配相关的下位概念；HOLONYM 模式用于匹配包含有整体关系的概念；NOUN 匹配相关的概念。

在此基础上，GOLF 采用混合的方法，即采用基于符号的 Hearst 模式方法和基于统计的聚类方法来进行核心本体(taxonomy)的构建。

4.2 非分类关系学习

在非分类关系学习中首先要发现表达非分类关系的语言学模式[8-9]。GOLF采用基于关联规则的$VCC(n)$事务方法进行非分类关系学习，$VCC(n)$事务方法假定：如果概念c_1和c_2间具有非分类关系v，当且仅当c_1和c_2都出现在带有动词v的n个词内(即c_1和c_2都出现在动词v的周围)，可以用一个条件概率来表示动词和概念对间的关联度。

$$P(c_1 \wedge c_2 / v) = \frac{|\{t_i \mid v, c_1, c_2 \in t_i\}|}{|\{t_i \mid v \in t_i\}|} \quad (3)$$

$$P(v / c_1 \wedge c_2) = \frac{|\{t_i \mid v, c_1, c_2 \in t_i\}|}{|\{t_i \mid c_1, c_2 \in t_i\}|} \quad (4)$$

其中，$|.|$表示集的基数；t_i表示$VCC(n)$-transactions。式(3)表示概念对与给定动词的可能关联度，式(4)表示动词与给定概念对的可能关联度。在某些情况下，动词与概念对中的每个概念同时出现的频率很高，但该动词并不表示概念对间的关系，如概念“ city ”和“ island ”对应的词项都经常与动词“ reach ”同时出现。因此，P(City∧Island/reach)和P(reach/ City∧Island)的值很高，甚至超过真正表示概念对间关系的动词的条件概率，如“ located ”。为此进行改进，采用启发式AE(Above Expectation)方法：

$$AE(c_1 \wedge c_2 / v) = \frac{P(c_1 \wedge c_2 / v)}{P(c_1 / v) \cdot P(c_2 / v)} \quad (5)$$

$$AE(v / c_1 \wedge c_2) = \frac{P(v / c_1 \wedge c_2)}{P(v / c_1) \cdot P(v / c_2)} \quad (6)$$

若 AE 的值大于阈值，则该动词确定为概念对间的关系。$VCC(n)$ 事务可以表示为一个三元组(Concept, Relation, Concept)，在具体实现中可以转化为三元组(Noun, Verb, Noun)或(Subject, Verb, Object)，该模型类似于一个 RDF 模型。系统按照以下正则表达式来识别 Noun 和 Verb：

Noun : (DT)?(JJ)*(NN|NNS|NNP|NNPS)+

Verb : (VB|VBD|VBN|VBZ)+

5 实验及分析

为了检验本文方法的有效性，这里选择高等教育领域和高校科研管理领域的中英文语料进行实验。其中新闻领域的英文语料(news_corpus)来自 http://english.sohu.com/，高等教育领域的英文语料(edu_corpus)来自 http://www.harvard. edu/，高校科研管理领域的中文语料(research_corpus)来自 http://www.sro.shu.edu.cn/。

以下主要针对 GOLF 系统的概念抽取和语义关系抽取两方面进行评价。对于实验的评价方法，采用在 IE 领域广泛使用的准确率(precision)、召回率(recall)。同时基于相同的实验语料，将 GOLF 的运行结果与 Text2Onto(不支持中文)相比较。采用 Text2Onto 进行比较主要基于 2 点：(1)Text2Onto 是一个开源软件，能够从 http://ontoware.org/projects/text2onto/免费下载使用；(2)Text2Onto 是目前最为著名的本体学习系统。部分实验结果如表 1~表 4 所示。

表 1 GOLF 与 Text2Onto 对于 news_corpus 的概念抽取结果

系统名	文档数	抽取概念总数	人工分析正确的概念数	准确率/(%)
GOLF	50	2 076	2 026	97.6
Text2Onto	50	2 015	1 912	94.9

表 2 GOLF 与 Text2Onto 对于 edu_corpus 的概念抽取结果

系统名	文档数	抽取概念总数	正确的概念数	准确率/(%)
GOLF	157	8 256	7 620	92.3
Text2Onto	157	7 013	6 150	87.7

表 3 GOLF 与 Text2Onto 对于 news_corpus 的语义关系抽取结果

系统名	文档数	抽取分类关系总数	正确的分类关系数	分类关系准确率/(%)	抽取非分类关系总数	正确的非分类关系数	非分类关系准确率/(%)
GOLF	50	723	504	69.7	37	29	78.4
Text2Onto	50	697	483	69.3	18	16	88.9

表 4　GOLF 与 Text2Onto 对于 edu_corpus 的语义关系抽取结果

系统名	文档数	抽取分类关系总数	正确的分类关系数	分类关系准确率/(%)	抽取非分类关系总数	正确的非分类关系数	非分类关系准确率/(%)
GOLF	157	2 811	2 148	76.4	147	121	82.3
Text2Onto	157	2 739	2 035	74.3	92	84	91.3

从实验结果可知，在新闻(news)、科研(research)和教育(edu)这 3 个所选择领域中，GOLF 的整体性能都优于 Text2Onto 系统。

6　结束语

本文提出一种基于 Web 的通用多策略本体学习框架 GOLF，与目前其他同类系统相比，有如下几点改进：

(1)采用多策略学习方法，通过对多个不同算法的平均，有效地消除单个学习方法的偏置。将多种不同的学习算法组合成一个集成系统，系统的性能得到明显改善。各学习算法的组合框架采用概率组合分布，根据不同的语料特征为每个算法设定不同的权值，增强了对不同领域语料的适应性。

(2)能够较好地处理各语种。与同类系统相比，对中文的处理能力明显加强。

(3)引入语料知识库思想，在进行领域本体学习时借助 WordNet 和 HowNet 等语义词典，提高 GOLF 系统的准确率和召回率。

(4)基本实现了跨领域、跨语种的全自动无监督本体学习功能，且性能良好。

参考文献

[1] Buitelaar P, Handschuh S, Magnini B. Towards Evaluation of Text-based Methods in the Semantic Web and Knowledge Discovery Life Cycle[C]//Proc. of ECAI Workshop on Ontology Learning and Population. Valencia, Spain: [s. n.], 2004-08.

[2] Cimiano P, Staab S, Tane J. Automatic Acquisition of Taxonomies from Text: FCA Meets NLP[C]//Proc. of ECML/PKDD Workshop on Adaptive Text Extraction and Mining. Cavtat-Dubrovnik, Croatia: [s. n.], 2003.

[3] Faure D, Nedellec C. A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology[C]//Proc. of the LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications. Granada, Spain: [s. n.], 1998.

[4] Buitelaar P, Olejnik D, Sintek M. OntoLT: A Protégé Plug-in for Ontology Extraction from Text[C]//Proceedings of the International Semantic Web Conference. Sanibel Island, USA: [s. n.], 2003.

[5] Quan T T. Automatic Generation of Ontology for Scholarly Semantic Web[C]//Proc. of ISWC'04. Hiroshima, Japan: [s. n.], 2004.

[6] Specia L, Enrico M. A Hybrid Approach for Extracting Semantic Relations from Texts[C]//Proc. of OLP'06. Sydney, Australia: [s. n.], 2006.

[7] Sabou M. Learning Web Service Ontologies: An Automatic Extraction Method and Its Evaluation[C]//Proc. of ISWC'05. Osaka, Japan: IEEE Computer Society, 2005.

[8] Maedche A, Staab S. Ontology Learning for the Semantic Web[J]. IEEE Intelligent Systems, 2001, 16(2): 72-79.

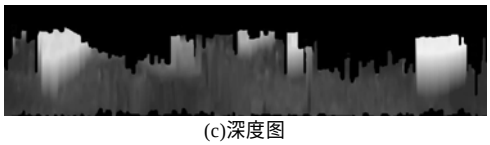
[9] Velardi P. Ontology Learning from Text: Methods, Evaluation and Applications[M]. Amsterdam, Holland: IOS Press, 2005.

(上接第 228 页)

95%的点误差小于 15 mm，只比标定误差略大。移动机器人能越过 25 mm 的障碍物，因此，以 25 mm 为阈值，得到如图 10 所示的场景信息，其中，深色为机器人可通行区域；浅色为检测到的障碍物区域。

(a)上镜面展开图

(b)下镜面展开图



(c)深度图

图 9　实验结果

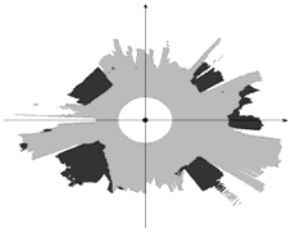


图 10　全向场景信息

6　结束语

本文提出一种 3 步立体匹配算法，结合特征匹配与全局匹配的优点，用全向立体视觉系统获取了全向深度图。该方法基本解决了系统面临的 3 个主要难点，实验结果令人满意。尽管此方法使用了一些经验参数和阈值，但是匹配结果并不会因为这些参数的微小变化而改变。在实验中也有少数极线匹配到的部分太少，今后将考虑利用极线间的一致性约束加以改进。

参考文献

[1] Cabral E L L, Souza J J C, Hunold M C. Omnidirectional Stereo Vision with a Hyperbolic Double Lobed Mirror[C]//Proceedings of International Conf. on Pattern Recognition, Cambridge, UK: [s. n.], 2004.

[2] 苏连成, 朱　枫. 一种新的全向立体视觉系统的设计[J]. 自动化学报, 2006, 32(1): 67-72.

[3] Sara R. The Class of Stable Matchings for Computational Stereo[R]. Prague, Czech: Czech Technical University, Technical Report: CTU-CMP- 1999-22, 2001.

[4] Strobl K. Camera Calibration Toolbox for Matlab[Z]. (2005-02-07). http://www.vision.caltech.edu/bouguetj/calib_doc.

[5] Luo Chuangjiang, Su Liancheng, Zhu Feng. A Novel Omni-directional Stereo Vision System via a Single Camera[M]. [S. l.]: Brill Academic Publishers, 2007-06.

[6] Lee Y, Kim D, Chung M. Feature Matching in Omnidirectional Images with a Large Sensor Motion for Map Generation of a Mobile Robot[J]. Pattern Recognition Letters, 2004, 25(4): 413-427.