一种基于 WWW 的 Ontology 属性值自动提取方法*

赵庆亮
北京大学计算语言研究所 北京 100871
E-mail: zhaoqingliang@pku.edu.cn

穗志方
北京大学计算语言研究所 北京大学 100871
E-mail: szf@pku.edu.cn

摘 要：属性值是描述 Ontology 中类的重要信息，但是当前关于属性值的自动提取的研究并不多。本文提出一种基于 WWW 的 Ontology 属性值自动提取方法。论文首先提出了一种在小规模属性值种子集的基础上，包含属性值的句子的选择与属性值提取互动的方法。这种方法利用互联网信息的冗余性，自动抽取并扩充目标属性值集合。然后，为避免人工构造属性值种子集，提出种子集自动生成的方法。我们设计实验来计算提取结果的正确率和召回率，此外，我们还通过将填充后的 Ontology 信息用于网页正文提取任务来展示 Ontology 自动扩充结果的有效性。

关键字：本体 万维网 搜索引擎 互动方法 属性值提取

To extract ontology attribute values automatically based on WWW

Zhao Qingliang
Institute of Computational Linguistics, Peking University, Beijing 100084
E-mail: zhaoqingliang@pku.edu.cn

Sui Zhifang
Institute of Computational Linguistics, Peking University, Beijing 100084
E-mail: szf@pku.edu.cn

Abstract: Attributes value is among the most important information to describe Ontology. However, few researches have been done about attribute values extraction so far. This paper proposes a method of extracting Ontology attribute values automatically based on WWW. Firstly, a method based on a seeds set is described about interaction between related sentences selection including attribute values and attribute values extraction, so that we can extract and expand the target attribute value set by the redundancy of WWW. Secondly, we construct the seeds set with an automatic method instead of by hand. Experiments are done to compute the precision and recall. Also automatically enriched Ontology information is used in webpage content extraction to show its usage.

Keywords: Ontology; WWW; Search Engine; interactive method; Attribute value extraction

1 引言

Ontology 是一种能在语义层次上描述知识的概念模型，其目的在于以一种通用的方式来获取领域中的知识，提供对领域中概念的共同一致的理解，从而实现知识在不同的应用系统之间的共享和重利用 [1] 。早期的 Ontology 的构建工作是通过人工完成的，耗费大量的人力、物力和财力，时间周期也很长，在很大程度上影响了 Ontology 的应用。近 30 年来，研究人员将精力集中在 Ontology 的自动、半自动构建上，取得了很多的成果 [2][3][4] 。

与 Ontology 的自动构建相关的研究大致可以分为下面几个方面。术语抽取，术语是构成 Ontology 的关键要素，因此术语自动提取是 Ontology 自动扩充的重要部分[5][6]；概念提取，Ontology 中对术语的完整描述是通过对其定义和属性的描述构成的，因此还应提取术语的定义及属性等与概念相关的信息 [7]；概念间关系提取，Ontology 是描述概念及概念之间的关系，概念

*基金资助：国家自然科学基金 60503071，863 项目 2006AA01Z144，973 项目 2004CB318102

间关系提取也是 Ontology 自动构建研究的一个重要方面[8]；术语自动分类，将相似的术语自动聚类，从而形成类的层次结构 [9]等。

Ontology 中一个类是通过这个类的属性以及这个类与其他类的关系来描述的，类的属性是承载一个类的信息最基本、最直接的载体。例如我们要了解某个型号的计算机（Ontology 中的一个类），最直接的办法是了解这个类的属性："CPU"、"Memory"、"Hard Disk" 等；如果还需要更深入的了解，还需要知道这个类与其他型号计算机（其他类）的关系，例如是由哪个型号的计算机（哪个类）改进而来的（继承关系）。而现在绝大部分的研究集中在研究类与类之间的关系上，而最基本的属性值自动提取却没有受到足够的重视。同时，属性值自动提取对于新术语发现、术语关系的发现以及术语的自动聚类都有着指导性的意义。如果两个类的属性值十分相似，那么这两个类之间则有可能存在某种关系。例如，医学类"慢性胃炎"和它的子类"慢性浅表性胃炎"的属性"症状"的属性值都有"上腹痛"、"嗳气"、"腹胀"、"食欲不振"、"反酸"、"恶心呕吐"等症状，重合率非常高，恰恰印证了它们之间存在上下位关系；在术语的自动分类任务中，我们也可以将它们的属性值是否类似作为判断能否聚类的指标。

本文提出一种基于 WWW 的 Ontology 属性值自动提取方法。论文首先提出了一种在小规模属性值种子集的基础上，包含属性值的句子选择与属性值提取互动的方法，利用互联网信息的冗余性，自动抽取并扩充目标属性值集合；进一步，为避免人工构造属性值种子集，提出种子集自动生成的方法。本文的结构如下：第 2 部分提出我们的基本假设；第 3 部分介绍我们的核心方法；第 4 部分介绍实验设置及对实验结果分析评价；最后对本文工作进行总结，并指出未来工作方向。

2 基于 WWW 的 Ontology 属性值自动提取的基本思想

基于单一文本进行 Ontology 自动构建，存在信息量有限、更新慢的缺陷，而 WWW 恰恰可以弥补这一缺陷。基于 WWW 获得的海量的网页信息存在很大的冗余性，虽然对于人而言，网页信息冗余会影响信息获取的效率，但这种冗余性对计算机自动获取知识却是很有帮助的。具体来说，这种网页信息的冗余性可以在两个方面对 Ontology 属性值自动提取提供帮助。

辅助计算机判断网页信息的可靠性和权威性。例如我们以"感冒 症状"为关键字在 Google 中进行检索，在最相关的 10 个网页中就有 5 个论述的感冒症状是一致的，占到 50%。据此，我们可以对相关网页中的候选属性值短语进行统计，那些频繁出现的短语很可能就是目标属性值。

选择简单的语法结构即可保证信息的完备性。在上边构造的检索中，取出最相关的 100 个网页，我们发现虽然网页中论述的感冒症状是相同的，但是它们分布在各种语法结构中。所以即使抛弃某些部分，对信息的完备性也不会产生大的影响。例如我们在本文中，只考虑比较容易处理和能够保证效果语法结构——并列结构，实验证明，这种选择对结果的完备性没有太大影响。

自然语言处理技术用于海量信息处理最大的困难在于，当处理复杂的语言结构时其正确率和处理速度都不能够满足实际需求。而本文通过利用了网页信息的冗余性，能够避开网页中语言的不规整现象，从而使得自然语言处理技术有可能在海量网页处理中得到较好的应用。

3 关键技术

3.1 整体框架

基于 WWW 的大规模 Ontology 的自动扩充方法，其输入是 Ontology 中的类名称、属性名称及属性值种子集，经相关句子提取和属性值提取，得到候选属性值集合，整体框架如下所示：

图 1：整体框架图

其中涉及到的关键技术包括：包含属性值的句子选择、候选属性值的提取及评价、种子集合自动生成等，以下分别予以介绍。

3.2 弱指导下相关句子选择与属性值提取互动的方法

我们的任务是对于给定的目标类及种子属性集合，从搜索引擎检索到的相关网页集合中提取出该目标类的属性值。网页作为处理的对象，颗粒度过大，网页中大部分内容会跟我们的目标无关。例如：对疾病类“感冒”的属性“症状”进行填充的过程中，我们使用关键字“感冒 症状”检索到网页中，在最相关的网页中，真正描述症状的字符数仅占整篇文章的 10%左右。所以在属性值提取时，将句子作为处理的基本单位，这样既能够保留必要的上下文信息，又能够保证颗粒度的足够细致。这样从网页中提取指定类的属性值的任务便被分解为两个步骤：1. 从网页中寻找包含属性值的句子；2. 在相关句子中寻找属性值。

我们已经提到抛弃搜索结果中不利于处理的某些网页（或网页中的某些部分）对于信息的完备性没有太大的影响。这里我们在筛选句子时只取满足以下两个条件的句子：

表 1 相关句子选择的两个条件

条件 1	句子中包含并列结构
条件 2	并列结构中出现种子属性值

之所以采用并列结构是由于其构成最简单，所以判断是否为并列结构非常准确；而且处于并列关系的两个概念具有相同的特性，对于本文的意义在于，如果并列结构中的一个概念是当前类的属性值，那么跟它具有并列关系的其它概念也非常可能都是目标类的属性值。

为了实现在给定尽量少的属性值的情况下，能够获得足够的相关句子，并最终获得准确完备的属性值集合。我们提出句子选择与属性值提取互动的策略。具体算法如下：把在句子中找到的新的属性值添加到种子集合当中，这样就可以找到更多的句子，而从更多的句子中又可以提取出来更多的属性值。同时我们设计了评价候选属性值的方法，只有置信度高于一定阈值的候选属性值才会被采纳。两个过程互为监督，这样能够最大限度的保证结果的完备性和准确性，整个过程结束于不再产生新的候选属性值。整个过程如下图所示：

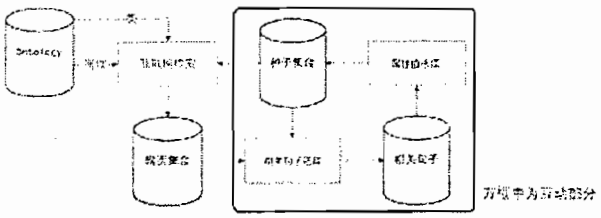


图 2 弱指导下互监督的属性值自动提取

对于候选属性值的评价，一个基本的想法是使用 χ^2 计算并列结构中每个短语的权重，如果权重高于设定的阈值则将其加入种子集中。然而与置信度高的种子并列的短语应该具有高置信度；而与置信度低的种子并列的短语则应该具有低置信度。所以这里当一个短语

和种子属性同时出现在一个并列结构中时，还应当加上种子属性值的权重，计算公式如下：

使用χ^2作为初始权重：

$$weight_{i,j}=\begin{cases}\dfrac{freq_{i,j}-m_{i,j}}{m_{i,j}}, freq_{i,j}>m_{i,j} \\ 0,\text{其他}\end{cases} \quad \text{其中 } m_{i,j}=\frac{\sum_i freq_{i,j}\sum_j freq_{i,j}}{\sum_{ij} freq_{i,j}}$$

迭代公式：

$$weight_{phrase}=weight_{phrase}+\sum_0^m weight_{phrase_m}，\ phrase_m\text{是与目标短语共现的种子短语}$$

公式 1 权重计算公式

算法 1：弱指导下相关句子选择与属性值提取互动的方法

3.3 无指导下的属性值自动提取方法

3.2 中提出的方法解决了当目标类的某个属性有种子属性集的情况下对这个属性的填充。但这种方法面临的一个困难是，当我们需要对大规模 Ontology 的自动构建时，不可能给每个类的每个属性都指定一个属性值种子集。为了解决这个问题，我们进一步提出了无指导的属性值提取方法。从而可以实现在填充整个 Ontology 的过程中，人工只需要为一个类的某个属性指定一个属性值种子集合，就可以完成其他所有类该属性的属性值的自动填充。

在 3.2 对一个类进行填充的过程中，我们得到一个句子的集合，如果我们能够在这些句子找出一定的模式，我们便可以利用这些模式来判断一个句子是否为描述某一类属性的句子，从而替代了表 1 中的两个判断条件。通过实验，我们发现情况不理想，即根据这种方法无法得到完全正确的属性值集合。但是，在获取的候选属性值集合中，那些置信度高的候选属性值，是目标属性值的可能性非常大。因此，我们只选这些高置信度候选属性值，将它们作为种子集，便可以按照 3.2 的方法，完成属性值自动提取。无指导下的属性值自动提取方法的示意图如下：

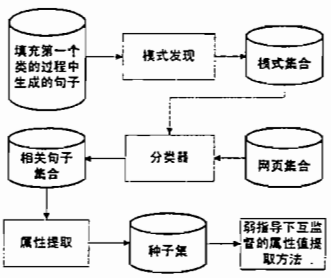


图 3 无指导下种子集的自动生成

依据 3.2 中的过程我们可以获得相关句子的一个集合，记作 TrainingSentenceSet。我们将 SentenceSet 中每一个句子中的并列结构替换为变量，例如：“…症状有：咳嗽、发烧等…” 替换为：“…症状有：$等…”，得到的结果便是这个句子和具体情况无关的部分，即反应当前属性的句子模式，然后将它们作为训练样例训练得到一个分类器。本文中选择的特征模板如下：

表 2 句子模式分类器特征模板

ID	特征模板	ID	特征模板	ID	特征模板
1	词： $word_i$	3	Bigram： $word_{i-1}+word_i$	5	Trigram： $word_{i-1}+word_i+word_{i+1}$
2	词性： pos_i	4	词性 Bigram： $pos_{i-1}+pos_i$	6	词性 Trigram： $pos_{i-1}+pos_i+pos_{i+1}$

经过一系列实验，本文最终选择的特征模板为 1、2、3、4，选用的特征模板不多，因为描

述属性的句式有限，如果选用太多模板会导致数据稀疏，对问题的处理产生负面影响。使用最大熵算法训练出一个分类器。我们仍然沿用 3.2 中的方法对候选属性值的进行评价，这样我们便获得了种子集合，可以 3.2 中进行有指导的属性填充过程。

4 实验设置与结果分析

4.1 实验设置

本文实验所用的本体为现代医学领域 Ontology，该本体基于美国国立医学图书馆编撰的《医学主题词表》（MeSH）作为知识描述体系的基础，MeSH 包括了医学领域中相对比较完整的主题词及其上下位关系，将其中的主题词作为现代医学 Ontology 的知识元，将其上下位关系作为 Ontology 的树状结构。对于每一种疾病的描述包括：名称、英文名、释义、代码与约束；临床类描述：症状与体征、实验室与其他辅助检查、发病部位、病因病机与病理等。

本实验采用北京大学计算语言学研究所的汉语词语切分与词性标注软件进行分词和词性标注。对于自动填充的结果，我们人工判断填充结果的对错，进一步通过 F 值来衡量：

$$F-value=\frac{2*precision*recall}{precision+recall}$$

公式 2：F 值计算公式

其中 precision 为填充结果的正确率，recall 为填充结果的召回率。

4.2 网页预处理

本文使用 Google API 作为获得原始网页的工具。我们选取检索结果中相关性排名前 100 个网页进行信息提取，在使用网页之前还需要对原始网页进行去燥处理，这项工作也称为网页正文提取。去燥主要基于以下两个假设：A. 互联网上的绝大部分网页是使用 Table 或者 Div 来分块的；B. 网页正文中链接数量占正文总量的比重较小；而导航、广告、索引部分中链接的比重较大。基于上述两个假设，首先构造网页的 Dom 树，统计叶子 Table 标签和叶子 Div 标签下的链接的比重，如果链接的比重超过 50%，则认为该块是噪音去掉。

4.3 实验结果及分析

4.3.1. 弱指导下互监督的属性值提取

- 使用种子集{“咳嗽”}填充类“感冒”的属性症状的结果(前 30 个)：

属性值候选	权重	属性值候选	权重	属性值候选	权重	属性值候选	权重	属性值候选	权重
头痛	0.95	喷嚏	0.29	恶心	0.14	咯痰	0.09	口不渴	0.07
咳嗽	0.68	恶寒	0.21	流鼻涕	0.13	盗汗	0.08	咳痰清稀	0.07
流涕	0.65	无汗	0.19	怕冷	0.13	咽痒	0.08	不发热	0.07
鼻塞	0.50	流泪	0.17	声音嘶哑	0.12	咽部不红肿	0.07	全身酸痛	0.07
打喷嚏	0.44	咽痛	0.17	口渴	0.10	轻微发热	0.07	乏力	0.06
发热	0.42	咽喉疼痛	0.16	胸痛	0.09	流清水鼻涕	0.07	无痰	0.06

从以上结果可以看出来，只输入类“感冒”和它的一个“症状”“咳嗽”的情况下，获得了与“感冒”相关的其它症状。尤其值得说明的一点是，我们只进行分词和词性标注的基本操作的情况下，可以提取出感冒相关的术语，例如“鼻塞声重”、“恶风寒”、“全身酸楚”等。

- 对多个类的“症状”属性进行自动填充统计正确率和覆盖率

类	种子集	权重排名前 15 的填充结果			类	种子集	权重排名前 15 的填充结果		
		Precision	Recall	F-value			Precision	Recall	F-value
感冒	{咳嗽}	73.3%	100%	0.84	浅表性胃炎	{腹胀}	80%	88.2%	0.84
慢性咽炎	{咽痛}	67%	100%	0.82	病毒性心肌炎	{心悸}	73.3%	88.2%	0.80
肺炎	{咳嗽}	80%	73.3%	0.77	哮喘	{喘息}	67%	100%	0.82
过敏性鼻炎	{打喷嚏}	73.3%	100%	0.84	尿毒症	{恶心}	73.3%	83.3%	0.78
非典型性肺炎	{发热}	87.8%	83.3%	0.86	结膜炎	{眼红}	53.3%	83.3%	0.65

F 值平均值：0.802。从结果上我们可以看出自动提取结果还是令人满意的。召回率上的表现要优于在准确率上的表现。

4.3.2. 无指导下的属性值提取

- 在给定“感冒”的一个症状“咳嗽”时填充类“慢性咽炎”的属性“症状”的结果：

属性值候选	权重	属性值候选	权重	属性值候选	权重	属性值候选	权重	属性值候选	权重
头痛	0.98	发痒	0.48	头晕	0.31	消化不良	0.26	心脏病	0.20
干燥	0.94	症状	0.44	四肢酸痛	0.30	支气管炎	0.26	咽痒	0.20
咽痛	0.82	微痛	0.43	食欲不振	0.28	刺激性咳嗽	0.25	刺激咳嗽	0.19
灼热	0.82	灼热感	0.35	变薄	0.28	烟熏感	0.24	干咳	0.18
痒	0.58	发胀	0.34	萎缩	0.28	疼	0.24	低热	0.18
异物感	0.53	咳嗽	0.34	声音嘶哑	0.27	干燥感	0.23	鼻塞	0.17

从这个例子上我们可以看出，在不给定目标类属性值的情况下，这种无指导的方法也能够取得较好的效果。当然这只是一个直观的具体的例子，更客观的评价还有待于更大规模的测试。

- 对多个类的“症状”属性进行自动填充统计正确率和覆盖率

类	种子集	权重排名前 15 的填充结果			类	种子集	权重排名前 15 的填充结果		
		Precision	Recall	F-value			Precision	Recall	F-value
慢性咽炎	N/A	67%	93.3%	0.78	病毒性心肌炎	N/A	73.3%	88.2%	0.80
肺炎	N/A	80%	73.3%	0.77	哮喘	N/A	67%	100%	0.82
过敏性鼻炎	N/A	73.3%	100%	0.84	尿毒症	N/A	73.3%	83.3%	0.78
非典型性肺炎	N/A	87.8%	83.3%	0.86	结膜炎	N/A	53.3%	83.3%	0.65
浅表性胃炎	N/A	80%	88.2%	0.84	**平均值**		**73%**	**88.1%**	**0.793**

F 值平均值：0.793。可以看出由于没有种子集，最后的填充结果有所下降，但并没有受到特别大的冲击，可见候选属性值评估方法具有一定的鲁棒性。

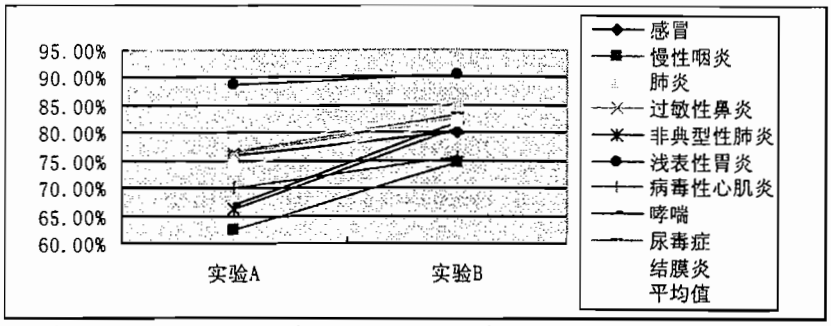
4.3.3. 用于网页正文信息提取对比实验

对自动构建的 Ontology 进行评价是一个相对困难的工作，原因在于评判标准缺乏统一性和客观性，所以我们设计了一种任务导向的实验，对网页正文信息提取任务标准数据分别使用下面三种方法进行评测，由于辨别是否为网页正文信息标准相对客观，从而可以获得较为可信的测评结果，同时也从一个侧面印证了 Ontology 属性值填充的意义。以“感冒”作为关键字使用搜索引擎进行检索，使用“感冒”类的属性“症状”对于检索出来的前 50 个网页进行正文信息提取。

A．Baseline：使用链接分析方法（见 3.2“网页正文信息提取”）

B．对比实验 1：使用未进行属性填充的 Ontology 辅助 Baseline 的网页正文信息提取

实验结果如下图所示：



可以看出，实验 B 比较实验 A 提高了 7.9%，但是效果比较有限；而实验 C 对比实验 A 提高了 20.5%，效果有了很大的提高。

5 结论

本文提出一种基于 WWW 的大规模 Ontology 的属性值自动填充方法。论文首先提出了一种在小规模种子集的基础上，相关句子的选择与属性值的填充互动的方法，利用互联网上网页的重复性和海量性，自动凸现目标属性值；进一步，为避免人工构造种子集，提出种子集自动生成的方法。同时，属性的自动提取对于 Ontology 自动构建中新术语发现、Ontology 自动构建中术语关系的发现以及术语的自动聚类都有着指导性的意义。

参 考 文 献

[1] 刘耀, 领域 Ontology 自动构建研究, 北京大学博士后出站报告, 2007.

[2] Maedche A. Ontology Learning for the Semantic Web. Boston: Kluwer Academic Publishers, 2002.

[3] P. Cimiano, A. Hotho and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. J. Artificial Intelligence Research Vol. 24, pp. 305 - 339, 2005.

[4] Kavalec M, Svátek V. A study on automated relation labelling in ontology learning. In: Buitelaar P, Cimiano P, Magnini B, eds. Ontology Learning from Text: Methods, Evaluation and Applications. Amsterdam: IOS Press, 2005.

[5] SUI Zhifang, CHEN Yirong, HU Junfeng, WU Yunfang, YU Shiwen. The Research on the Automatic Term Extraction in the Domain of Information Science and Technology. 第二届东亚术语论坛，2002 年 12 月.

[6] Delphine Bernhard. Multilingual Term Extraction from Domain-specific Corpora Using Morphological Structure. The Association for Computational Linguistics, Trento Italy, 2006.

[7] Agirre, E., Ansa, O., Hovy, E., and Martinez, D. 2000. Enriching very large ontologies using the www. In Proceedings of the Ontology Learning Workshop, ECAI 2000. Berlin, Germany.

[8] SATOSHI SATO and YASUHIRO SASAKI. Automatic collection of related terms from the web. IPSJ SIG Notes, 2003(4):57–64, 20030120.

[9] 昝红英、胡俊峰、穗志方、俞士汶. 信息科学与技术领域中的术语分类研究. The 5th East Asia Forum of Terminology Proceedings，中国海口，2002 年 12 月，p191-197.

[10] Paul Buitelaar, Philipp Cimiano, Marko Grobelnik, Michael Sintek. Ontology Learning from Text. Tutorial at ECML/PKDD 2005.

[11] P. Cimiano and S. Staab, Learning Concept Hierarchies from Text with a Guided Agglomerative Clustering Algorithm. In: Proceedings of the ICML 2005 Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods. 2005.