基于Web弱指导的本体概念实例及属性的同步提取*

康为 穗志方

北京大学计算语言研究所

北京大学计算语言学教育部重点实验室 北京 100871

E-mail: kangwei@pku.edu.cn szf@pku.edu.cn

摘 要：本文提出了一种基于Web弱指导的本体概念实例和属性的同步提取方法，利用小规模的种子实例和属性集，本文从Web上自动获取实例和属性共现的上下文模式，并利用种子实例和属性的关联性来评价这些模式。进一步，根据上下文模式提取候选概念实例和属性后，本文提出两种方法来评价提取的候选实例和属性。第一，利用概念实例和属性的关联性来互相评价对方的准确度；第二，利用候选实例或候选属性与种子实例或属性在上下文模式分布上的相似度来评价准确度。实验结果表明，本文的方法能够有效地辅助本体的自动构建。

关键词：Web，概念实例提取，属性提取，弱指导，上下文模式

Weakly-Supervised Extraction of Ontology Concept Instances and Concept Attributes from the Web

Kang Wei, Sui Zhifang

Institute of Computational Linguistics, Peking University, Beijing, 100871

E-mail: kangwei@pku.edu.cn szf@pku.edu.cn

Abstract: In this paper, we propose a weakly-supervised method of extracting Ontology concept instances and attributes from the Web. We automatically acquire the co-occurrence patterns of the concept instances and attributes from the Web, and we evaluate these patterns based on the assumption that concept instances are relevant to their attributes. Furthermore, we extract the candidate concept instances and attributes. This paper proposes two ways to evaluate the accuracy of the candidate instances and attributes: the first measure is based on the correlation between concept instances and attributes, and the second one is based on the distribution similarity on the context patterns between the candidate instances (or attributes) and the seed instances (or attributes). Experiments show that our method can effectively aid the automatic construction of Ontology.

Keywords: Web, Domain concept instance extraction, Attributes extraction, Weakly-supervised, Contextual Pattern

1. 前言

Ontology可以看作是概念和概念之间关系组织起来的结构，而其中Instance-of和Attribute-of是最基本的两种关系。作为本体学习的重要部分，本体概念实例提取和属性提取的研究越来越受到重视。

Hearst提出了利用句法模式从文本中得到上下位关系的方法 [1]，利用的句法模式如"such NP_o as NP_i,...,NP_{n-i} (or|and) other NP_n"等，这种方法能够取得较高的准确率，但是基于单一文本的实例提取往往会出现数据稀疏的问题。[2]从Web上提取候选概念属性，并将判别属性看作分类问题，利用两个有指导的分类器来进行分类。 [3][4][8]利用无指导或弱指导的方法从非结构化的Web文本中提取概念实例，[5]使用无指导的方法从半结构化的HTML文档中提取属性和属性值对，[6]利用弱指导的方法从结构化的Web文档中提取概念属性，近年来随着Wikipedia的不断发展，基于Wikipedia的属性提取也受到关注[7]。

上述的研究大多关注于单独的概念实例提取或属性提取任务，而没有同时进行概念实例提取和属性提取，只有[8]利用弱指导方法从Web文档和搜索引擎查询日志中获取开放领域的概念的实例和属性。本文提出了一种基于Web弱指导的本体概念实例和属性的同步提取方法，利用

*本文承国家自然科学基金项目 60873156、国家 863 项目 2006AA01Z144、国家 973 项目 2004CB318102 的资助。

小规模的种子实例和属性集，从Web上自动获取实例和属性共现的上下文模式，并利用种子实例和属性的关联性来评价这些模式。进一步，本文提出两种方法来评价提取的候选实例和属性。第一，利用概念实例和属性的关联性来互相评价对方的准确度；第二，利用候选实例或候选属性与种子实例或属性在上下文模式分布上的相似度来评价准确度。实验结果表明，本文的方法能够有效地辅助本体的自动构建。

本文的组织结构如下：第二章主要介绍基于Web的本体概念实例和属性同步提取的的基本思想；第三章介绍实例和属性提取的关键技术；第四章介绍实验设置及对实验结果的分析评价；最后是本文工作的总结。

2. 基于Web的本体概念实例和属性同步提取的基本思想

2.1. 基本思想

本体是对某个领域中的概念的形式化的明确的表示。从语义上分析，实例表示的就是对象，而概念表示的则是对象的集合。譬如一个医学本体中，“疾病”是一个概念，而具体的疾病“感冒”、“高血压”等是“疾病”的实例，这些实例都有一些共同的特征和属性，譬如疾病的实例都有“症状”、“治疗”、“病因”等属性，这些属性是用来描述概念及其实例的本质特征。因此，概念实例和概念的属性之间具有密切的关系，具有相同属性集合的对象可以认为是同一概念的实例，而一个概念的实例具有相同的属性集合。在领域语料中，概念的实例往往和其属性描述同时出现。本文以Web作为语料，利用少量的种子实例和种子属性，提取实例和属性共现的上下文模式，并进一步利用上下文模式同时提取概念实例和属性。本文的工作主要围绕着以下三个问题进行解决：

1) 如何解决概念实例和属性提取中的数据稀疏问题？

基于上下文模式进行信息提取，往往会出现数据稀疏的问题，而以Web作为语料就可以较好的解决这一问题。由于Web上信息传播、共享的便捷性，Web上的信息具有很大的冗余性。对于人而言，这种冗余性会影响信息获取的效率，而对于基于模式的信息提取任务，由于Web信息的冗余性，在单一文本中会出现数据稀疏的模式，在Web中则可以出现很多次，因此Web信息的冗余性恰恰可以用来解决数据稀疏的问题[9]。进一步，本文通过自动构造查询请求，利用Google搜索引擎返回的排序靠前的检索结果作为语料来提取概念实例和属性。

2) 如何评估候选模式的准确性？

概念实例和属性往往出现在特定的上下文模式中，本文利用种子概念实例和属性构造如“IH_1AH_2”（I为种子实例，A为种子属性，H_1和H_2是上下文）的查询请求，在Google返回的检索结果中自动提取实例和属性共现的上下文模式，通过这种方式提取的模式有很多是不准确的，因此我们利用种子实例与属性的关联性来评价候选模式的准确性，越能体现种子实例和属性的模式越准确。例如，对于上下文模式P＝“I的A及”，种子实例“感冒”和种子属性“症状”在P上的关联性表现为，“感冒”和“症状”出现在Web时“感冒的症状及”在Web上出现的概率，综合所有这样的种子实例和属性的组合情况，就能够评价出模式P反映种子实例与属性关联性的程度。

3) 如何评价提取的概念实例和属性？

由于Web信息的复杂性，利用Web提取的候选概念实例和属性不可避免的有一些噪音，因此需要对候选实例和属性进行可靠性的评价。本文从两个方面对候选进行评价。第一，利用概念实例和属性之间的密切关系来相互评价候选实例和属性。譬如，对于概念“疾病”，我们有种子属性“症状”、“治疗”和“病因”，真正的疾病实例相比于不是疾病的候选实例与这些种子属性有更为密切的关系，我们用PMI-IR来衡量这种密切的关系[10]。第二，利用候选实例

（或候选属性）和种子实例（或种子属性）在上下文模式集合$\mathbb{R}'$上分布的相似度来评价候选实例（或候选属性）。譬如，对于疾病的种子实例“感冒”、“高血压”、“鼻炎”，候选实例“牙结石”比“科学正确”更接近种子实例的上下文分布。本文综合了以上两种评价方法，既考虑了概念实例和属性的密切关系，又把提取对象与其种子的联系也作为度量。

2.2. 整体框架

基于Web的本体概念实例和属性同步提取方法，其输入是少量的种子实例和种子属性，在Web上，利用种子实例和属性提取上下文模式并进行评价，利用模式提取候选实例和候选属性并进行评价，最终得到排序后的概念实例列表和属性列表。系统的框架如图1所示，它包括三个主要模块。

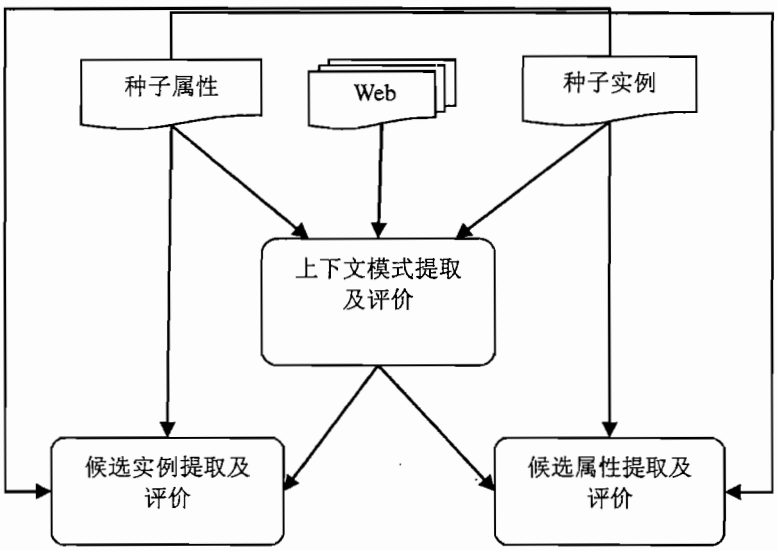

图 1 基于 Web 的本体概念实例和属性同步提取方法框架

1) 上下文模式的提取和评价模块：该模块提取概念实例和属性共现的上下文模式，并且利用Google评价提取的模式。2) 候选实例的提取和评价模块：在该模块中，利用种子属性和1）中提取的上下文模式构造查询请求，利用Google返回的结果提取候选实例，并用基于与种子属性PMI-IR和种子实例相似度的可靠性评价方法来评价候选实例，并扩充种子实例集合。3) 候选属性的提取和评价模块：该模块利用扩充的种子实例集合和1）中提取的上下文模式构造查询请求，利用Google的结果提取候选属性，并用与2）相似的方法评价候选属性。

3. 关键技术

3.1. 基于Web的上下文模式的提取和评价

❶ 上下文模式的提取

实例与属性的关系其实是一种“I具有A”的关系，如“感冒的症状有”、“高血压的治疗需要”等就体现了这种关系，因此我们试图提取出形如“IH_1AH_2”的上下文模式，其中I是概念实例，A是属性，H_1和H_2是在语料中出现频次高于阈值F、并且长度小于阈值L的上下文片段。本文利用小规模种子集，构造出给Google搜索引擎的查询请求，利用Google API获得每个查询排名前100的结果，把Google返回结果的网页标题和结果片段（Snippets）作为提取上下文模式的语料集，记为Corpus P。然后提取“IH_1AH_2”类型的模式，将得到的模式集合记为$\mathbb{R}$。

- 上下文模式的评价

我们从Corpus P中提取了上下文模式集合$\mathbb{R}$，并且可以得到$\mathbb{R}$中的每个模式γ在Corpus P中出现的频次，但是我们不能简单的使用这个频次来评价模式γ，因为γ在Corpus P中的出现的概率不能反映真实的情况，我们用γ在Web上出现的概率作为权重来评价γ，

$$score(\gamma)=\sum_{<i,a>\in(I\times A)}\frac{Hits(<i,a>,\gamma)}{Hits(<i,a>)} \quad (1)$$

其中，(<i,a>,γ)表示将模式“IH_1AH_2”中的I和A替换为具体的i和a，Hits(q)表示将q作为查询词在Google中检索得到的结果数目。我们将score(γ)小于给定阈值的模式排除，得到最终的模式集合$\mathbb{R}'$，再将$\mathbb{R}'$中的γ的权重归一化为score'(γ)

3.2. 概念实例提取及评价

我们利用2.1中提取的上下文模式集合$\mathbb{R}'$和种子属性构造Google的查询请求，并在Google返回的结果集合中提取候选的概念实例，同时本文提出了两种方法对概念的可靠性进行评价。

- 提取概念实例候选

基于2.1中提取的上下文模式和种子属性，我们用具体的种子属性a替换上下文模式“IH_1AH_2”中的“A”，并构造查询请求query=“$*H_1aH_2$”，我们通过在Google中检索query得到符合模式的结果，记做Corpus I。在Corpus I中依据模式集合$\mathbb{R}'$抽取概念实例候选，我们采用如下的策略：首先，我们以Corpus I中的句子为单位进行抽取，我们只选择所有以“$*H_1AH_2$”作为开头的句子，抽取其中匹配“*”的部分，记为集合$\mathbb{C}$。然后，进一步对集合$\mathbb{C}$中的字符串S进行处理，我们使用前缀和后缀停用词表去掉S中的噪音前缀和后缀，并只保留长度在2到10之间的字符串，经过上面的筛选，我们最终得到候选概念实例集合ξ。

- 候选概念实例可靠性评价

通过上下文模式提取出来的概念实例候选不可避免的会包含噪音，因此需要对候选实例进行置信度的评价。本文提出了两种方法来评价候选实例：

1)基于实例候选和种子属性PMI-IR的评价方法

一个合法的概念实例应该和种子属性的相关程度很大，因此我们用实例候选和种子属性的互信息来衡量实例候选的置信程度。本文中使用Google搜索引擎统计实例候选和种子的PMI-IR来计算实例候选的可靠性P(i)，如公式2、3所示：

$$P(i)=\frac{\sum_{a\in A}\left(\frac{pmi(i,a)}{\max_{pmi}}*P(a)\right)}{|A|} \quad (2)$$

$$pmi(i,a)=\log\frac{Hits(i,a)*N}{Hits(i)*Hits(a)} \quad (3)$$

其中，P(i)是候选实例的可靠性，P(a)是属性的可靠性，Hits(q)是以q作为Google的检索词得到的结果数目，(i,a)是以i和a同时作为Google检索的关键词，N是Web上所有文本的数目。

2)基于实例候选和种子实例相似度的评价方法

i) 根据2.1中提取的上下文模式集合$\mathbb{R}'$，为每个种子实例α构造特征向量。构造特征向量的方法为：对于$\mathbb{R}'$中的每个上下文模式γ =“IH_1AH_2”，将I和A替换为具体的种子实例和种子属性后作为查询请求在Google中检索，得到Google返回的Hits(γ)，则特征向量的每个特征值用公式4来计算：

$$p(\alpha|\gamma)=\frac{p(\alpha,\gamma)}{p(\gamma)}=\frac{Hits(\alpha,\gamma)}{score'(\gamma)*N} \quad (4)$$

其中，Hits(α,γ)是以α,γ共同作为Google的检索词得到的结果数目，score'(γ)是2.1中计算的模式γ的权值，N是Web上所有文本的数目。得到所有种子实例的特征向量后，把它们相加并做归一化得到一个参照特征向量vs。

ii)用i)中描述的方法为所有的候选实例I_{cand}构造特征向量vc。

iii)使用Jensen-Shannon divergence [11]计算候选实例特征向量vc和参照特征向量vs的相似度，并根据计算的相似度对候选实例进行排序。

3.3. 属性提取及评价

与概念实例提取类似，属性提取也分为提取候选属性和评价候选属性两个部分。

- 提取属性候选

在提取属性时，我们把实例提取之后置信度高的实例候选添加到实例种子集中。利用Google返回检索结果Corpus A，我们选择所有匹配上下文模式“IH_1*H_2”的句子，抽取其中匹配"*"的部分，记为集合$\mathbb{C}'$。对$\mathbb{C}'$中的字符串S只保留长度在2到8之间并且出现频次大于给定阈值的字符串，经过上面的筛选，我们最终得到候选概念实例集合ξ'

- 属性候选可靠性评价

1）基于候选属性与种子实例PMI-IR的评价

基于PMI-IR的属性候选可靠性评价与实例候选相似，有所不同的是种子实例集合添加了置信度小于1的实例，公式8描述了候选属性a置信度的评价：

$$P(a)=\frac{\sum_{i\in I}\left(\frac{pmi(a,i)}{\max_{pmi}}*P(i)\right)}{|I|} \quad (5)$$

其中pmi(a,i)与公式4中的pmi(i,a)相等。

2） 基于候选属性与种子属性相似度的评价

候选属性的可靠性也可以用其余种子属性的相似度来衡量，相似度计算的步骤也分为三个：

i)利用种子属性构造参照特征向量。ii)为候选属性构造特征向量。iii)计算候选属性特征向量和参照特征向量的相似度，并根据计算的相似度对候选实例进行排序。

4. 实验分析

4.1. 实验数据

本文使用Google API作为获得Web语料的工具，在构造查询请求后，我们把Google返回的检索结果的标题和上下文片段作为提取上下文模式、实例和属性的语料。本文在医学领域的概念“疾病”上进行了实验，我们使用的初始种子实例集为{感冒、高血压、鼻炎、颈椎病、肾结石}，初始的种子属性集为{症状、治疗、病因}。

本文中使用一个领域专家人工校订过的，基于美国国立医学图书馆编撰的《医学主题词表》（MESH）的现代医学领域Ontology作为实例和属性提取的黄金标准。作为标准的现代医学领域Ontology中包含3904个疾病实例，其中有148中常见疾病。我们使用准确率、覆盖率来评价实例提取，用准确率评价属性提取的结果。其中，对于准确率的评价采用了人工确认的方法，对每个自动提取的概念实例都一一进行人工确认。由于我们无法真正得到我们提取的实例和属性在整个

Web上的召回率，本文中采用了覆盖率作为替代的方法，这里的覆盖率是指我们提取的实例与作为黄金标准的本体中的实例的交集占黄金标准中实例的比例。

4.2. 概念实例提取的实验结果

我们在概念“疾病”上进行实例提取的实验，结果得到2200个候选实例，其中有315个疾病实例在黄金标准中出现，覆盖率达到了8.1%，有58个实例在黄金标准的常见病中出现，覆盖率达到了39.2%。表1中给出了实例提取的覆盖率。人工确认候选实例的准确率在前500个结果达到94%，前1000个结果的准确率也高达93%，自动提取实例的准确率在图表1中给出。

	S(标准中实例个数)	*E(自动提取实例个数)*	$\frac{S \cap E}{S}$
常见病(黄金标准)	148	2200	39.2%
所有疾病(黄金标准)	3904	2200	8.1%

表格 1 实例提取在黄金标准实例上的覆盖率

从结果中我们可以看出在Web上自动提取的概念实例已经覆盖了相当程度的常见病，同时相比与标准本体，也有315个实例被提取出来，而且标准本体在构造中并不能囊括所有的疾病实例，在我们提取的结果中有相当一部分实例经过人工确认是合法的疾病实例。

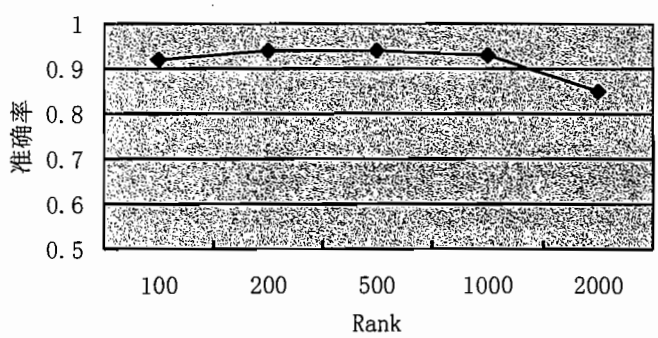

图2 实例自动提取的准确率

从图2可以发现，经过排序后的候选实例的前1000个结果准确率都达到了92%以上，而前2000个结果的准确率也维持在85%，准确率下降的原因在于经过我们的排序，置信度高的实例候选大多排在了前面，而置信度低的实例大多集中在后面。[4]利用人工选择的模式在规模为60,000,000的Web文档集合上提取概念“公司”和“国家”的实例，他们对提取的实例进行抽样并人工确认其准确性，在“公司”上抽取的实例有1,116个，准确率为90%。对比上述研究的实验结果，我们的方法利用较少的资源，在提取出更多实例的同时，准确率依然维持较高的水平。

4.3. 属性提取的实验结果

我们在概念“疾病”上进行实验，前30个结果的准确率都达到了60%以上，前20达到70%，前5个结果最高，达到了80%。[8]利用弱指导方法从Web文档和搜索查询日志中获取开放领域的概念的实例和属性，其提取的排名前20的候选属性的平均准确率达到了67%，相比于[8]中使用了50,000,000个查询日志和100,000,000个Web文档，本文的方法使用的资源规模要小很多，但是却取得了相当的准确率。相比于概念实例，一个概念的属性的数目要远远小于概念实例的数目，因此属性提取的准确率较实例提取要低一些。图3展示了属性自动提取的准确率。

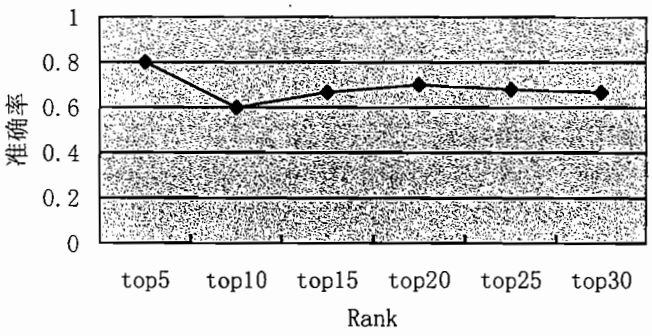

图3 属性自动提取的准确率

5. 结论

本文提出了一种基于Web弱指导的本体概念实例和属性的同步提取方法，利用小规模的种子实例和属性集，自动从Web上获取概念实例和属性共现的上下文模式，并利用种子实例和属性的关联性来评价这些模式。在根据上下文模式提取候选概念实例和属性后，本文利用概念实例和属性的关联性以及候选实例或候选属性与种子实例或属性在上下文模式分布上的相似度来评价准确度。以Web作为语料进行实例和属性提取，充分的利用了Web信息的冗余性，可以有效的克服单一文本中的数据稀疏问题。实验结果表明，本文方法提取出的概念实例和属性有不错的准确度，能够有效的辅助本体的自动构建。

参 考 文 献

[1] M. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the 14th International Conference on Computational Linguistics, pages 539-545, Nantes, France, 1992.

[2] M. Poesio, A. Almuhareb. Identifying Concept Attributes Using a Classifier. In Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition, pages 18-27, Ann Arbor, 2005.

[3] O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D.S. Weld, A. Yates. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. Artificial Intelligence, v.165 n.1, p.91-134, June 2005

[4] M.J. Cafarella, D. Downey, S. Soderland, O. Etzioni. KnowItNow: Fast, Scalable Information Extraction from the Web. In Proceedings of HLT/EMNLP, pages 563-570, Vancouver, October 2005.

[5] N. Yoshinaga, K. Torisawa. Open-Domain Attribute-Value Acquisition from Semi-Structured Texts. In Proceedings of the OntoLex 2007, Busan, South-Korea, November 11th, 2007.

[6] S. Ravi, M. Pasca. Using Structured Text for Large-Scale Attribute Extraction. In Proceedings of the 17th International Conference on Information and Knowledge Management(CIKM-08), pages 1183-1192, Napa Valley, California, USA, October 2008.

[7] G. Cui, Q. Lu, W. Li, Y. Chen. Automatic Acquisition of Attributes for Ontology Construction. In: ICCPOL2009, Vol.5459, Springer(2009), pages 248-259.

[8] M. Pasca, B.V. Durme. Weakly-Supervised Acquisition of Open-Domain Classes and Class Attributes from Web Documents and Query Logs. In Proceedings of the ACL-08: HLT, pages 19-27, Columbus, Ohio, USA, June 2008.

[9] F. Keller, M. Lapata, O. Ourioupina. Using The Web to Overcome Data Sparseness. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 230-237, Philadelphia, July 2002.

[10] P. Turney. Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL. In Proceedings of the 12th ECML-2001), pages 491–502, Freiburg, Germany, September, 2001.

[11]L. Lee. Measures of Distributional Similarity. In Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics (ACL-99), pages 25-32, College Park, Maryland, 1999.

通讯作者：穗志方 szf@pku.edu.cn