基于并列结构的概念实例和属性的同步提取方法*

李文杰，穗志方
北京大学 计算语言学研究所，北京 100871
E-mail: lwj@pku.edu.cn; szf@pku.edu.cn

摘 要：在概念实例和属性的提取研究中，针对基于模式的方法召回率比较低的特点，本文提出了一种基于并列结构的概念实例和属性的同步提取方法。首先利用并列结构模式去网页集合中提取同类词语集合，然后再用基于种子的弱指导方法去学习实例和属性共现的上下文模式，最后再通过模式去提取候选实例或候选属性。在此过程中，每提取出一个候选，就将该候选所在的同类词语集合合并到候选集合中。实验结果表明，本文的方法在不降低准确率的基础上，能大大提高提取结果的召回率。

关键词：并列结构；搜索引擎；实例提取；属性提取；上下文模式

To Extract Concept Instances and Concept Attributes Based on Coordination

Li Wenjie, Sui Zhifang
Institute of Computational Linguistics, Peking University, Beijing 100871
E-mail: lwj@pku.edu.cn; szf@pku.edu.cn

Abstract: Most researches on concept instances and concept attributes extraction has focused on based-pattern approaches, motivated by the problem of these approaches always harvest a low recall rate, in this paper we present a method of extracting concept instances and concept attributes based on coordination. Because a part of candidate instances and attributes extracted by the coordination patterns can be putted into the similar-concept-phrases sets in advance, we can use these similar-concept-phrases sets to greatly expand the extraction results in the procedure of co-occurrence pattern-based extraction. Compared with the baseline without using the coordination patterns, experimental results show that the coverage of this method has greatly improved but without reducing the precision rate.

Keywords: coordination; search engine; instances extraction; attributes extraction; contextual pattern

1 引言

概念是反映客观事物及其特有属性的思维对象，它是知识表示的核心要素。对概念的提取研究分为：（1）概念实例提取，例如：提取“疾病”概念下的所有疾病名，包括：“感冒”、“心肌炎”等；（2）提取概念的属性名（例如：提取“疾病”概念的属性名“症状”、“并发症”、“用药”等）。

在概念实例和属性的自动提取方面，当前已有很多研究。

[1] Hearst 提出了一种利用手工指定的模式从非限定性文本中自动获取上下位关系的方法，这种方法可以获得很高的准确率，但是需要人工制订模式，这种方法提取出的结果往往都是有限的。[2][5][6]从 Web 文档中提取实例,其中[2]利用迭代的方式对种子实例集合进行扩展，[5][6]利用给定的概念和上下位模式通过在搜索引擎中构造查询请求来自动获取实例。[3]是利用搜索引擎的查询日志来进行实例提取,[4]提出一种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，[7]利用给定的概念和概念的实例集合从结构化的网页文本中提取概念属性，[8]利用手工指定的模式分别对网页文件和查询记录中提取属性的结果进行了比较。

以上研究大多关注的是单独的概念实例和属性提取，而[9]提出了一种利用非常少的种子属性从 Web 文档和搜索引擎查询日志中同时提取实例和属性的方法，[10]提出了一种基于 Web 弱指导

* 本文相关研究得到国家自然科学基金 60873156、61075067 以及国家社会科学基金 09BYY032 的支持。

的本体概念实例和属性的同步提取方法，利用给定的种子实例和属性集，在 Web 搜索引擎中查询，通过寻找实例和属性共现的上下文模式来提取新的概念实例和属性。

本文在[10]基础上提出了一种基于并列结构的概念实例和属性的同步提取方法。在同步提取之前，首先利用并列结构去获得一些同类词语集合，然后再用这些集合去扩充同步提取结果。实验结果表明这种使用并列结构的方法在不降低准确率的情况下，能大大提高提取结果的召回率。

本文的结构如下：第 2 部分提出了本方法的基本思想；第 3 部分介绍了本方法各部分的关键技术；第 4 部分介绍了实验设置及对实验结果的分析评价；最后对本文的工作进行了总结。

2 基于并列结构的概念实例和属性的同步提取基本思想

[10] 假设概念实例和属性往往出现在特定的上下文模式中，利用种子实例和属性构造形如“IH1AH2”（I 为种子实例，A 为种子属性，H1 和 H2 是上下文）的查询请求，在搜索引擎返回的结果中自动提取实例和属性共现的上下文模式。然后利用这些模式，再构造形如“*H1AH2”的查询请求去提取候选实例，构造形如“IH1*H2”的查询请求去提取候选属性。

这种基于模式的同步提取方法有一个很大的缺点就是召回率比较低，当提取出上下文模式后，只有当某个实例或属性能精确匹配该模式时才能被提取出。但是通过观察，我们发现语料中存在着很多这样的句子：“心律失常的并发症有冠心病、风心病、心肌病、高心病、肺心病等。”，即很多情况下多个概念实例和属性都是以某种并列结构的形式出现在语料中。如果我们提前能通过句子中的并列结构将这几种疾病归为一类，当我们利用模式提取“疾病”概念实例时，若发现了“冠心病”为一个实例，则只需将其他的“风心病、心肌病、高心病、肺心病”等都加入候选实例集合即可，这样可以提取出很多基于模式的方法提取不出的候选实例，属性的提取也同样如此。

因此本文提出了基于并列结构的概念实例和属性的提取方法，在进行提取时融入了并列结构这样一种特征，能大大提高系统的召回率。和前面的工作类似，本文也选取搜索引擎作为获得语料的工具。和单一文本相比，Web 的信息冗余性能获得更好的结果。

3 关键技术

3.1 整体结构

基于并列结构的概念实例和属性的同步提取方法，其输入是少量的种子概念实例和种子属性，以 Web 搜索引擎为语料获取工具。这种方法的整体结构如图 1 所示。

它主要包含四个模块：基于并列结构的同类词语提取、上下文模式的提取、候选实例的提取及候选属性的提取，下面对这几部分涉及到的关键技术分别予以介绍。

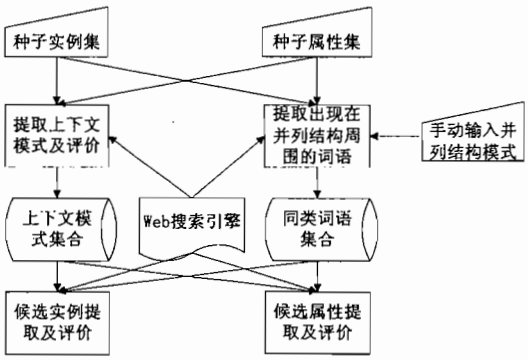

图 1 整体结构框架图

3.2 基于并列结构的同类词语的提取

之所以采用并列结构是由于具有并列关系的两个词语表现为相同概念下的特性，即如果并列结构中的一个词语是某个概念下的实例（或属性），那么跟它具有并列关系的其他词语也非常可能都是该概念下的其他实例（或属性）。

本文获取包含并列结构语料的方法是：利用给定的种子实例和种子属性，将种子实例两两组合和种子属性两两组合作为查询关键词，在搜索引擎中进行查询，将每个结果网页下载下来，对网页进行预处理，抽取出网页中的文本内容作为语料。

然后人工构造并列结构的规则，提取出语料中能匹配这类规则的所有词语。构造的规则为："$S \rightarrow$、词、词、"，即出现在三个"、"中间的两个词语就被认为是具有并列结构关系的词语对。我们只选取了这一种规则，是因为这种规则准确度很高。有一些其他的规则也可能包含并列结构信息，例如出现在"和"字周围的两个词，但是这些规则准确度不高，会提取出很多错误的词语对，这些噪音很大程度上会影响后面的提取结果。

接着将这些词语对按照贪心的思想进行简单的分块，分块遵循的原则是按并查集算法将具有并列结构关系的词语放在同一个集合中。

分块完成后会生成一个大的集合记作 Φ，Φ 中包含有很多个词语集合 Φi，每个集合中包含的是同一类的词语，并且对于 Φ 中的任意两个集合 Φi 和 Φj，它们之间没有交集，因为若 Φi 和 Φj 存在着某个共同的词语，则它们应该会被归并成同一个集合。

这样做的好处是，可以通过这种方法尽可能多地将表现相同概念的词放在同一个集合中。在后面的基于模式的提取过程中，若提取出的某个候选实例或属性出现在了某个集合 Φi 中，则可以认为 Φi 中的其他词语也都为候选概念实例或属性。

3.3 模式的提取

3.3.1 模式的提取

和文献[10]类似，本文提取模式的方法也是根据给定的种子实例和种子属性，对每个种子实例 ins_seed 和种子属性 attr_seed，构造这样的查询关键词"ins_seedC1attr_seedC2"（C1、C2 为通配符）在搜索引擎中进行查询。提取出每一对匹配到的字符串 C1 和 C2，若 C1 和 C2 的长度不超过阈值 L，就将{C1，C2}加入到候选模式集合 P 中。

3.3.2 模式的评价

一个候选模式在语料中出现的次数越多，则表明该模式是一个比较好的模式，应该具有相对较高的权重。因此对于模式集合 P 中的每一个候选模式 Pi，我们定义了如下的模式评价公式：

$$Conf(Pi) = \frac{freq(Pi)}{\sum_{Pj \in P} freq(Pj)} \tag{1}$$

其中 $Conf(Pi)$ 为模式 Pi 在语料中出现的频率，$freq(Pi)$ 为模式 Pi 在语料中出现的频次。

3.4 基于并列结构的概念实例提取

3.4.1 基于模式的概念实例提取及评价

- 提取

对于 3.3 提取出的模式集合 P 中的每一个模式 Pi={C1，C2}和每一个种子属性 attr_seed，我们构造这样的模式"*C1attr_seedC2"去搜索引擎中进行查询。然后搜索每一个以该模式开头的句子，将匹配"*"部分的字符串抽取出来，通过构造停用词表去掉字符串前后的无用信息，若最后剩下的字符串长度在 2~10 之间，则将其作为候选实例记作 ins。

- 评价

一个模式的置信度值越高，这个模式就越能反应概念实例和属性之间的关联程度，出现在该模式周围的候选实例和属性就越有可能是正确的概念实例和属性。对于由模式 Pi 和种子属性提取出的实例 ins，我们定义如下的公式来计算候选实例 ins 的置信度值。

$$Conf(ins)=\begin{cases} Conf(pi) & ins \notin Ins \\ Conf(ins) \oplus freq(pi) & ins \in Ins \end{cases} \tag{2}$$

初始时实例集合 Ins 为空。若提取出的候选实例不在实例集合 Ins 中，则令其置信度值直接等于模式的置信度值；若提取出的候选实例已经在实例集合 Ins 中，则将其以前的置信度值加上模式的置信度值作为其新的置信度值。

3.4.2 利用同类词语集合对原始候选进行扩充

提取出原始候选实例 ins 后，接着将 ins 在 3.2 提取出的集合 Φ 中进行查找，若发现 ins 在 Φ 中的某个词语集合 Φi 中，则 Φi 中的其他词语都可看作候选实例。

且 Φi 中某个词语的词频越大即其与其他词语出现过并列结构的次数越多，该词也越有可能是概念实例。因此对于通过候选实例 ins 在集合 Φi 中发现的每个候选实例 ins_par，我们定义如下的公式来计算其置信度值。

$$Conf(ins\_par)=\begin{cases} Conf(ins\_par) & ins\_par \in Ins \\ \dfrac{freq(ins\_par)}{freq(ins)} * Conf(Pi) * \lambda & ins\_par \notin Ins \end{cases} \tag{3}$$

若 ins_par 已经在实例集合 Ins 中，则不改变其置信度值；若 ins_par 不在实例集合 Ins 中，则通过下面的公式来计算其置信度值，其中 $freq(ins\_par)$ 为 ins_par 在集合 Φi 中的频次，$freq(ins)$ 为 ins 在集合 Φi 中的频次，$Conf(Pi)$ 为发现候选实例 ins 的模式 Pi 的置信度值，λ 为权重因子。

3.5 基于并列结构的属性提取

3.5.1 基于模式的属性提取及评价

- 提取

和实例提取的方法类似，对每一个上下文模式 Pi={C1，C2}和每一个种子实例 ins_seed，我们构造查询关键词“ins_seedC1*C2”去搜索引擎中获取属性提取的语料。然后构造同样的模式“ins_seedC1*C2”去语料中搜索，将匹配“*”部分的全部字符串抽取出来，若字符串的长度在 2~8 之间，则将其作为候选属性记作 attr。

- 评价

候选属性的评价我们也是采用和实例评价同样的方法，通过其和模式之间的关联程度来评价。对于每个通过模式 Pi 提取出的候选属性 attr，我们定义如下的置信度计算公式。

$$Conf(attr)=\begin{cases} Conf(pi) & attr \notin Attr \\ Conf(attr) \oplus freq(pi) & attr \in Attr \end{cases} \tag{4}$$

3.5.2 利用同类词语集合对原始候选进行扩充

同样的，将候选属性 attr 在 3.2 提取出的集合 Φ 中进行查找。对于通过候选属性 attr 在 Φ 中的某个集合 Φj 中发现的每个候选属性 attr_par，我们定义和上面类似的公式来计算其置信度值。

$$Conf(attr\_par)=\begin{cases} Conf(attr\_par) & attr\_par \in Attr \\ \dfrac{freq(attr\_par)}{freq(attr)} * Conf(Pi) * \lambda & attr\_par \notin Attr \end{cases} \tag{5}$$

4 实验

4.1 实验设置

本文以 Web 为语料，选取百度为获得语料的工具。除和文献[10]一样选取医学领域的“疾病”概念为实验对象外，我们还选取了“药物”、“汽车”、“国家”以及“宗教”这几个概念来评价本方法的提取效果。对于提取结果，通过人工去判别其准确性。因为 Web 上实例提取结果的召回率很难去计算，因此我们选取计算提取结果在所选的黄金标准中的覆盖率来代替召回率。

4.2 实验结果

4.2.1 “疾病”概念实验结果

医学领域以 MESH 的现代医学领域 Ontology 作为实例和属性提取的黄金标准。该 Ontology 中包含有 3904 个疾病实例。本文以文献[10]为 Baseline，使用和文献[10]相同的种子实例集{感冒、高血压、鼻炎、颈椎病、肾结石}和种子属性集{症状、治疗、病因}，选取了模式提取的前十个模式进行实验，评价时令权重因子 $\lambda = 1.0$ 。

表 1 给出了本文的方法和 Baseline 在黄金标准所有疾病实例上的覆盖率之间的对比，而图 2 则给出了本文提取结果前 2000 个的准确率。

表 1 实例提取在黄金标准上的覆盖率对比

	提取实例与黄金标准重合数	覆盖率 $\mathbb{R}$
Baseline	315	8.1%
本文的方法	746	19.1%

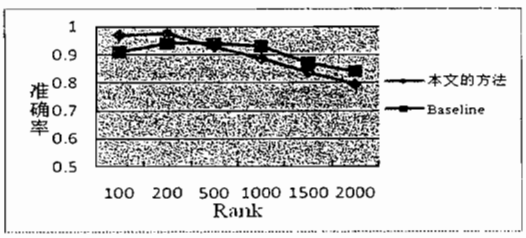

图 2 疾病概念下实例提取的准确率对比

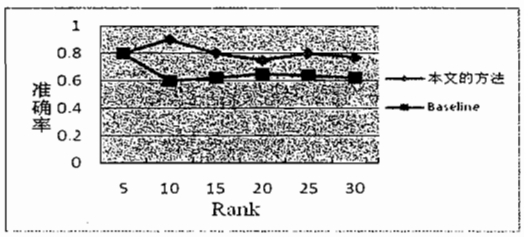

图 3 疾病概念下属性提取的准确率对比

由表 2 和图 2 我们发现，Baseline 在不使用并列结构的情况下只找出 315 个实例，而本文的方法在仍然保持很高的准确率上，比 Baseline 多找出 431 个实例，覆盖率 $\mathbb{R}$ 从 8.1%提高到了 19.1%。甚至在前两百个实例的准确率上，本方法的准确率还高于 Baseline。

通过对属性结果进行评价，我们发现本文的方法在属性提取的准确率上也有很大的提高，这是因为由并列结构提取出的同类词语具有很高的准确度。图 3 给出了属性提取的准确率对比。

4.2.2 其他概念实验结果

除了选取医学下的“疾病”概念为实验对象外，本文还选了“药物”、“汽车”、“国家”和“宗教”这几个概念下的实例和属性提取来比较使用并列结构和不使用并列结构的结果。在进行实验的时候，我们发现[10]的方法对种子的依赖很强，差的种子会得到很差的结果，并且当实例不是简单的出现在句子的开头时，通过模式的方法提取的候选结果会很差，而从并列结构获取的候选往往都比较准确，这个时候我们可以提高从并列结构获取候选的权重因子，来获得不错的准确率。

“药物”概念以上面的医学领域 Ontology 列出的 1576 种药物为黄金标准；“汽车”概念以汽车之家网站列出的 101 种汽车品牌为实例提取黄金标准；“国家”概念以当前联合国的 192 个会员国为实例提取黄金标准；“宗教”以中文维基百科提供的 154 种宗教为黄金标准。表 2 给出了几个概

念的实例提取结果对比。

在属性提取方面，对比前三十个候选属性的准确率，不使用并列结构与使用并列结构时："药物"概念下准确率从56.7%提高到86.7%，"汽车"概念下准确率从46.7%提高到60%，"国家"概念下准确率从50%提高到63.3%，"宗教"概念下准确率从56.7%提高到70%。表3给出了几个概念下的属性提取结果的前十个候选属性。

表2 实例提取结果对比

	不使用并列结构		使用并列结构	
	准确率（前100个）	覆盖率	准确率（前100个）	覆盖率
药物	40%	5.39%	90%	7.68%
汽车	70%	30.7%	80%	73.3%
国家	69%	60%	78%	75%
宗教	54%	13%	65%	20.8%

表3 属性提取结果

	置信度值排名前十的候选属性
药物	药理作用、疗效、不良反应、适应症、用法用量、副作用、效果、禁忌症、作用机制、成分
汽车	系列、油耗、底盘、爬坡能力 、标志、车型、加速时间、真实、价格、车名
国家	国歌、国徽、国旗、首都、面积、人口、货币、总统、语言、象征物
宗教	教义、教规、教理、经典、在华传教、本质、人才学、真相、区别、教法

5 结论

本文提出了一种基于并列结构的概念实例和属性的同步提取方法，这种方法在基于模式的方法中融入了并列结构这样一种特殊结构，首先通过并列结构提取出一些同类词语集合，然后再用基于种子的弱指导方法去提取候选实例和候选属性，每提取出一个候选，就将该候选所在的同类词语集合中的其他词语都加入到候选集合中。实验结果表明，和单纯的同步提取的弱指导方法比，本文的方法在不降低准确率的基础上，能大大提高提取结果的召回率。

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 [C]. Nantes, France, 1992: 539-545.

[2] R. C. Wang, and W. W. Cohen. Iterative Set Expansion of Named Entities using the Web. In Proceedings of ICDM 2008[C]. Pisa, Italy, 2008.

[3] M. Pasca. Weakly-supervised discovery of named entities using web search queries. In Proceedings of CIKM-07[C], pages 683-690, New York, NY, USA, 2007.

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

[5] R. C. Wang, and W. W. Cohen. Automatic Set Instance Extraction using the Web. In Proceedings of ACL-IJCNLP-09[C], Suntec City, Singapore, August 2009.

[6] Z. Kozareva, E. Riloff, and E. Hovy. Semantic class learning from the web with hyponym pattern linkage graphs. In Proceedings of ACL-08: HLT[C], pages 1048-1056, Columbus, Ohio, June, 2008.

[7] S. Ravi, and M. Pasca. Using structured text for large-scale attribute extraction. In Proceedings of the 17th CIKM (CIKM 2008) [C], Napa Valley, California, pages 1183-1192, 2008.

[8] M. Pasca, B. Van Durme, and N. Garera. The role of documents vs. queries in extracting class attributes from text. In Proceedings of the 16th CIKM (CIKM-07) [C], pages 485-494, Lisbon, Portugal, 2007.

[9] 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 [C]. Columbus, Ohio, USA, June 2008.

[10] 康为、穗志方. 基于Web弱指导的本体概念实例及属性的同步提取. 中文信息学报[J], 2010, 24(1), 54-59.