### 3.2.1 Web信息采集方法分类

Web信息采集方法从不同的角度可以将分为不同种类[15]。根据网页信息源的不同可以分为自由文本采集、半结构化的文本采集和结构化的文本采集。根据自动化程度可以分为人工方式、半自动方式以及全自动化方式[16]。根据包装器原理的不同可以分为基于层次结构的信息抽取归纳方法和基于概念模型的多记录信息抽取方法。根据人工干预的程度分为基于知识工程的方法和自动训练法[17][18]。根据用户和抽取系统的交互将网页信息采集系统分为手工构造的信息抽取系统、基于监督的信息抽取系统、基于半监督的信息抽取系统、无监督的信息抽取系统[19]。但最常见的分类方式是根据包装器形成方式的不同将网页信息采集分为基于自然语言处理的网页信息抽取、基于本体（Ontology）的网页信息抽取、基于包装器归纳方式的信息抽取、基于HTML页面结构分析的网页信息抽取、基于Web查询的网页信息抽取和基于页面视觉特征的网页信息抽取[20][21][22][23][24][25]。表1分别对各种采集方法进行对比，并对其在开放获取资源采集中的应用进行分析。

表 1 各采集方法比较分析

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 方法名称 | 代表系统 | 包装器形成方式 | 适用领域 | 应用于开放获取资源采集 |
| 基于自然语言处理的网页信息抽取 | 基于WHISK[27]算法的信息抽取系统 | 依赖于自然语言处理技术，先对网页样本进行标注，建立语言字典。即将采集网页内容分割成多个句子，分析句子成分，使用自然语言处理技术来建立短语和短语之间的关系，进而获得抽取规则，所产生的抽取规则通常基于语法约束和语义约束[26] | 主要适用于含有大量文本且句子完整的网页 | 需要前期准备大量的能覆盖所有情况的训练样本，构建足够全面的语意字典，工作量大。 |
| 基于本体的网页信息抽取 | SNAX[28]系统 | 对特定主题或领域建立本体集，然后通过网络爬虫工具，对网页进行解析，完成对网页正文信息提取、分词等。将处理后的网页信息传递给计算模块，计算模块根据本体集来对网页信息进行加权计算。权值反映的是该网页与主题或领域之间的相关度，若与内容相关性高就保存，不相关就丢弃。 | 主要适用于特定主题或领域的网页 | 需先请领域专家对不同来源的开放获取资源构造完整的本体知识库，需要专家投入相当多的时间。 |
| 基于包装器归纳方式的信息抽取 | WIEN[29]系统 | 根据事先由用户标记的样本实例，应用机器学习方式归纳算法，生成基于定界符的抽取规则。 | 主要适用于具有相同结构和编码的网页。 | 对网页结构多变的网站采集效果不好。用户无法预知网页结构的变化，形成的抽取规则无法覆盖网站内所有资源。 |
| 基于HTML页面结构分析的网页信息抽取 | RoadRunner | 充分利用了HTML的半结构化特性，在信息抽取之前通过解析器将Web文档解析成结构树，通过自动或半自动的方式产生抽取规则，定位主题信息，将网页信息抽取转化成对解析树的操作[30]。 | 主要适用于由共同模板生成，有一定规律性的网页。 | 形成抽取规则时，为了归纳出通用的正则表达式，要使用了大量复杂的启发式搜索算法，实现起来有一定的难度。 |
| 基于Web查询的网页信息抽取 | Web-OQL[32] | 将Web文档解析成一棵抽象的HTML语法树，根据页面的结构和所需提取信息的唯一的标记，写出合适的类似于SQL语句的Web查询语言对语法树进行查询。 | 主要适用于网页源代码中抽取字段有清晰的标识的网页 | 需要人工对网页进行识别抽取字段的唯一标识，且对网页结构依赖性很强，对没有明确的唯一标识的网页抽取十分困难。 |
| 基于页面视觉特征的网页信息抽取 | 基于VIPS[33]算法的采集系统 | 利用Web信息的视觉特征，例如字体的大小和颜色、段落的长短、数据所处区域等信息，将一个网页分割成多个不同的视觉信息块，去除掉诸如导航条、广告条、版权信息等与主题无关的信息之后，再使用包装器对信息进行抽取。 | 主要适用于网页结构清晰的网页。 | 由于视觉特征具有复杂而又不确定的特性，往往需要人工不断地修改调整抽取规则，使用起来很难保证规则集的一致性[34]。 |

总结而言，基于不同方法的网页信息采集，最终都是使用构建的包装器对网页信息进行采集，区别在于形成包装器的方式不同。基于自然语言处理的信息抽取和基于本体(Ontology)的网页信息抽取事先需要领域专家的参与，编写语义字典或本体知识库，且专家编写的质量直接影响到采集效果。此类方法更加适用于对网页中相关主题信息的提取，但因受限制于专家的参与，通用性较差。基于Web查询的网页信息抽取和基于HTML页面结构分析的网页信息抽取需要用户对网页结构进行一定的分析，对单个或结构统一的网页元数据采集十分方便，但对于向开放获取资源这样结构多样的网页，就需要多次人工解析，工作量大。基于页面视觉特征的网页信息抽取适用于结构清晰的网页，而且往往还需结合其他方法对元数据进行采集，适用范围有限。基于包装器归纳方式的信息抽取通过样本页面归纳学习抽取规则，减少人工参与，而且适用于各种类型的网页。但是在样本网页的选择中有时无法覆盖到所有情况，如在开放获取资源的网站中，无法预知从哪一个网页开始结构发生了变化，训练所得到抽取规则无法抽取网页结构变化后的元数据信息。而各采集方法通过不同方式最终形成的抽取规则，采集到的元数据都有较高的准确率和召回率。这些方法在其适应的领域能保证采集数据的质量，而在对结构多样化的开放获取资源采集方面具有一定的不适应性。