《智能信息处理》课程考试

基于本体的 Web 信息抽取的研究

刘行顺

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2020 年 12 月 18 日

基于本体的 Web 信息抽取的研究

刘行顺[1)]

（大连海事大学信息科学技术学院 辽宁大连）

摘 要 基于本体的 Ｗｅｂ信息抽取就是以所构建的本体为核心，利用本体中已定义的概念、分类层次、关系、函数、公理和实例及一些必需的外部资料对 Ｗｅｂ页面进行信息提取，得到结构化的知识并保存的过程，这一技术已经成为国内外研究的热点之一。本文介绍了本体与信息抽取的相关概念，并对应用到的相关技术进行了详细分析，最后本文对基本本体的信息抽取方法与其他方法之间进行了评析。

关键词 本体；信息抽取；web 页面；技术评析

Research on Web information extraction based on Ontology

Liu Xingshun[1)]

Abstract Web information extraction based on ontology is to build the ontology as the core, the use of ontology is defined in the concept, classification level, relations, functions, and the axiom for instance and some necessary external data information of Web pages are extracted, structured knowledge and save the process, this technology has become one of the hot spot of research at home and abroad. This paper introduces the concepts related to ontology and information extraction, and analyzes the relevant technologies in detail. Finally, this paper reviews and analyzes the information extraction methods of basic ontology and other methods.

Key words Ontology; Information extraction; The web page; Technical evaluation

引言

信息抽取是指将无结构或半结构化的文本转化为结构化的信息，并以一定的形式进行存储，供用户查询以及进一步分析利用的过程。Web 信息抽取则是从 Web 页面中抽取用户感兴趣的信息并过滤掉不相关的信息，将分散在半结构化 Web 页面中的信息提取出来，并以结构化、语义更为清晰的模式表示。

在信息时代的今天¨信息的增长速度已经让我们瞠目结舌¨随着 Web 的发展¨在 Web 上的各种各样的信息也以不同的形式分布¨我们该如何在这些繁多的、无结构的 Web 信息中找到我们真正需要的¨就成为现在急需要解决的问题。Web 信息抽取就是为了这个目的而存在的¨它可以把 Web 中的信息变成结构化的、 更有语义的模式结构。 在 Web 信息抽取技术中¨基于本体的 Web 信息抽取是其中一个比较重要的方向¨它的实现可以帮助用户更方便地在信息海洋中找到自己需要的信息¨减少应用程序的资源浪费。

1 相关概念

1.1 本体

本体（Ontology）的概念最初起源于哲学领域，20 世纪 70 年代末 John Mc Carthy 将这个哲学术语引入到计算机领域¨在人工智能界，最早给 Ontology 定义的是 Neches 等人。他们将 Ontology 定义为 “给出构成相关领域词汇的基本术语和关系¨以及利用这些术语和关系的构成规定这些词汇外延的规则定义”。1993 年 Gruber 给出了 Ontology 的一个最为流行的定义[1]。即“Ontology”是概念模型的明确规范说明”。Ontology 的目标是捕获相关领域的知识¨提供该领域知识的共同理解¨确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些词汇 （术语） 和词汇间相互关系的明确定义。

1.2 领域本体

领域本体 （Domain Ontology），是专业性的本体，描述的是特定领域中的概念和概念之间的关系，提供了某个专业学科领域中概念的词表以及概念间的关系，或在该领域里占主导地位的理论，能够独立地存在和被使用。

1.3 Web 信息抽取

Web 信息抽取（Web Information Extraction）是将 Web 作为信息源的一类信息抽取。简单地说，Web 信息抽取是指从 Web 页面中抽取用户感兴趣的信息而过滤掉不相关的信息，具体是指研究如何将分散在半结构化 Web 页面中的信息提取出来，并以结构化、语义更为清晰的模式表示，它为用户在 Web 中查询数据、应用程序直接利用 Web 数据提供了便利[2]。Web 信息抽取中输入信息抽取系统的是原始文本，输出的是固定格式的信息点。其主要功能就是把信息点从各种各样的文档中抽取出来，然后以统一的形式集成在一起。

2 基于本体的Web信息抽取相关技术

2.1 基于本体的 Web 信息抽取模型

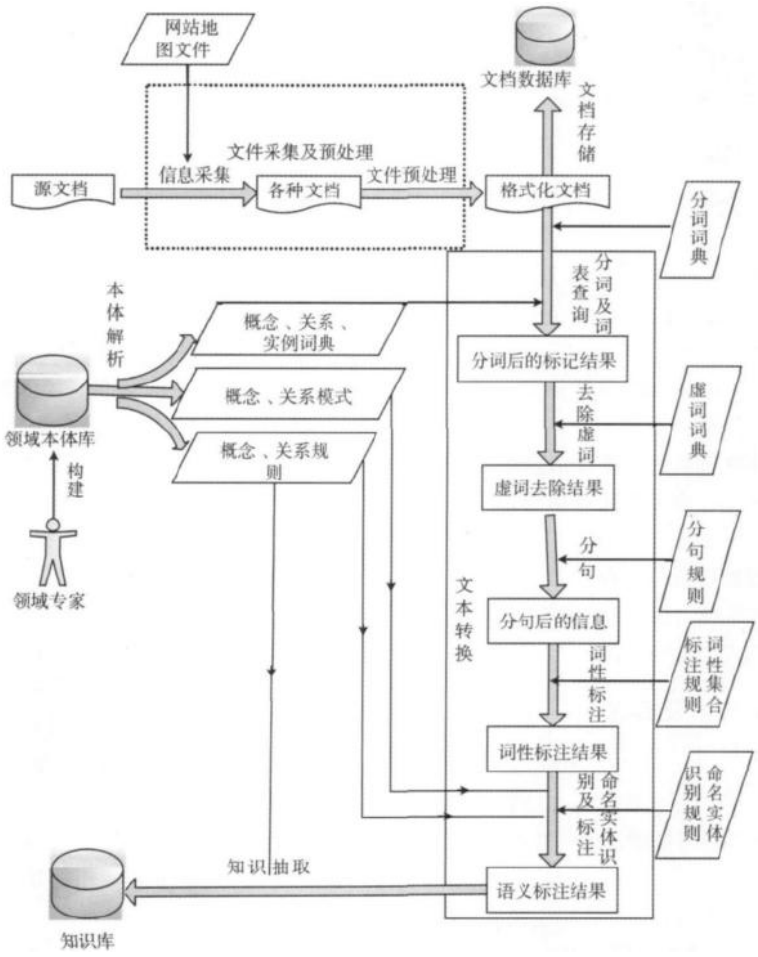

图 1 基于本体的 Web 信息抽取框架

基于本体 的 Web 信息抽取就是以所构建的本体为核心，利用本体中已定义的概念、分类层次、关系、函数、公理和实例及一些必需的外部资料对 Web 页面进行信息提取，得到结构化的知识并保存的过程。图 1 给出了基于本体的 Web 信息抽取框架。

2.1.1 系统构成

整个系统包括文件采集及预处理、文本转换、知识抽取 3 个部分。

在信息抽取过程中需要一些外部资料，包 括：①领域本体；②网站地图文件；③单词、短语、特殊字符词典和通用词典；④停用词、虚词、高频词词表；⑤分句规则；⑥词性集合及词性标注规则；⑦实体识别及标注规则；⑧知识抽取规则。

信息抽取用到的算法包括：① 爬网 算 法；②文件预处理算法；③文件存储算法；④ 分词和词表查询算法；⑤停用词、虚词、高频词去除算法；⑥分句算法；⑦词性标注算法；⑧命名实体识别及标注算法；⑨知识抽取算法[3]。

2.2 信息抽取的关键技术

2.2.1 存在客体的有效识别

存在的客体作为一个文本中的信息单位，是文本的基本构成元素，有效识别这些客观实体，是理解文本内容的基础。从更深层次的角度讲，这些客观实体在现实中的描述对象不只是属于具体或是抽象的范畴，还有可能包括时间等。而结合具体应用的情境，才能够确定命名实体的确切含义。客观实体识别的技术难点在于：

- ◆ 实体表达形式不统一，形式多样；
- ◆ 庞大的数量以至枚举的困难性，全部收录属于理想状态；
- ◆ 实体的外部延伸会随着领域和场景的变化而有所差别；
- ◆ 有些类型的实体名称不遵循严格的规律，变化频繁无常；
- ◆ 缩写形式普遍存在于各类文档；

2.2.2 语句解析

一般而言，计算机能够对于属于自然语言范畴的文本进行理解是有着一定的基础，该基础就是借助于文章语句解析来实现的，解析得到的结构化输入，能够对于文章的理解提供良好的技术支撑。通过近年来对于语句解析的研究方向来看，块分析技术成为了越来越多的系统所应用的技术支

持，这种趋势造成的因素也是多样化的[4]。

2.2.3 篇章分析与推理

信息抽取系统力求解决文本内及文本间的共指问题，希望能够在推理技术支撑下，达到正确合并描述属于同一事件或客体信息片段的目的。但从现阶段的研究状况来看，能够被借用的篇章分析理论和方法数量十分之少。已有的篇章分析理论的要求都比较苛刻，要借用大量的知识，需要更加规范的文本结构，以及在大规模语料上的初测性，在面向人、面向口语的进程中还需要技术更新。

2.2.4 知识捕获

信息抽取系统的处理本质上属于自然语言的范畴，其抽取信息的目标实现需要异常强大的知识库支持来确保信息的完备性。知识库的内容和结构随着信息抽取系统的变化而变化，尽管有其可变性，一般看来，都涵盖了一部词典、一个抽取模式库及一个概念层次模型。

2.3 基于本体的 Web 信息抽取

采用的抽取技术即为基于本体的 Web 信息抽取技术，此技术脱离了传统信息抽取对于网页结构过于依赖的尴尬境地，抽取的实现途径是借助于描述的数据项本身。理论上说，只要能够给出的应用领域的本体足够强大，足够丰富，系统就可以对目标领域中的给定网页进行完备率、准确率较高的信息抽取。一般来说，当使用这个方法来实现抽取目的时，第一步骤是由领域知识专家对领域本体进行设计构造，构建本体需要完善的信息当然都包括在内，准备工作也是必不可少，之后抽取规则的生成会根据本体概念及其层次关系的描述信息，放入规则库中。

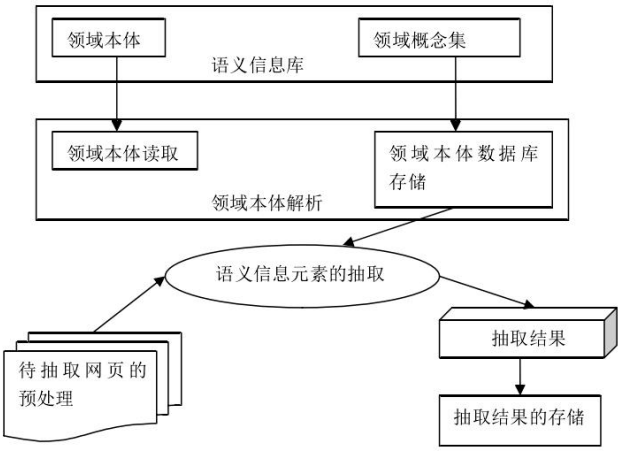

图 2 基于领域本体的 web 信息抽取框图

从图中都可以清楚地看到，该策略的核心就是本体的构建和信息的抽取。整个信息抽取的流程简单表述为：

① 语义信息库：根据本体构建规则，合理完善地建立任务所要抽取的该网页的信息所属的领域本体，作为信息抽取的基础。

② 领域 ontology 解析：本体生成后，对本体解析生成领域对象的一系列概念、关系以及层次关系，并把结果存放到数据库中；是生成抽取规则的依据。

③ 预处理：通过预处理将 HTML 网页解析成为任务需要的形式化文本，预处理阶段可能包括语义标注等环节，这是对任务所需数据源的合理化约束。

④ 语义信息抽取：结合任务构造的抽取规则，依据本体提供的语义关系，通过自身的抽取模型获取目标信息，存储抽取结果。

信息抽取的思路为：获取领域相关概念集，对目标领域设计本体模型，采用合适的本体构建工具构建领域本体；借助于最新的 OWL 语言进行本体描述，将领域本体内的类及它们之间的关系约束合理表示，之后运用基于本体产生的抽取规则，当然要用较好的语法来编写规则，完成对 web 信息的抽取[5]。

2.4 基于页面信息本体的 Web 信息抽取

上述策略是基于领域本体的 Web 信息抽取¨当然还可以结合 DOM 等技术进行进一步的改进¨随之系统框图也会有所改变¨我们提供的是最底层架构¨在此基础上可以为了提高查全率和查准率¨而结合其他的一些先进技术。

由于基于上述策略的方法实现都是以领域本体的建立为基础¨而将本体与其他方法结合的研究还较少；并且此类本体的构建中¨领域特性表现得极其明显；建立领域本体的过程需要领域专家参与¨过程复杂¨周期较长[6][7]因此¨简化本体建模过程¨缩短建模周期¨对基于本体的 Web 信息抽取技术有着重要的意义[8] 。

3 基于本体的Ｗｅｂ信息抽取方法的评析

3.1 基于本体的Ｗｅｂ信息抽取方法的类型

基于本体的 Web 信息抽取方法有多种。

(1) 根据其实现的自动化程度，可以把 Ｗｅｂ信息抽取方法分为人工、半自动化和全自动化 3 种。人工方式费时费力、代价昂贵，因此是不切实

际的；

(2) 根据学习的过程可以把 Web 信息抽取分为知识工程 （基于规则）方法和自动训练 （基于统计）的方法；

(3) 根据现有工具的抽取原理可以将基于本体的 Web 信息抽取划分为：基于 HTML 结构的信息抽取、包装器归纳方式的信息抽取、基于形式文法的半自动文本的信息抽取、基于事件 （框架）的信息抽取、基于主题 （场景）的信息抽取、基于自然语言处理方式的信息抽取、基于直接比对的信息抽取、基于本体相似的信息抽取等。这里主要对第三种方法进行详述。

3.2 基于ＨＴＭＬ结构的信息抽取

这类方法根据 Web 页面的结构定位信息，自动或半自动地从网页中寻找模式（模板）或规则，并利用这些模式或规则实现对网页内容的信息抽取。抽取模式或规则可以手工设置，也可以通过机器学习的方法得到。徐东兴首先选定抽取区域，然后基于领域本体解析结果产生抽取规则，对特定结构进行信息抽取。如文中对<TABLE>的抽取，使用了 XSLT 和 XPath 表示所产生的抽取规则，然后对源文档进行转换，得到抽取结果[9]。张鑫等根据数据容器的视觉特征准确划分数据区域，通过启发式学习从这些结构相似的数据区域树中得到信息项的抽取路径，然后通过抽取路径自动构建领域本体，最后通过对领域本体的解析得到信息项的抽取规则进行信息抽取。另外，也有研究者给出了利用本体基于 HTML 结构的信息抽取方法。

3.3 包装器归纳方式的信息抽取

这类方法首先由用户手工标注一组网页作为训练数据，然后利用机器学习方法从训练数据中学习得到抽取规则。此方法对有规范模式的 Web 文档是非常适用的。陆科进、李新颖首先经过分类找到所需数据在页面中的分布，然后利用本体概念信息使用所开发的针对特定结构的包装器对网页数据进行信息抽取[10]。周明健等根据本体定义，对用户感兴趣的信息区域使用应用归纳学习技术的包装器生成规则并进行信息抽取[11]。也有研究给出了包装器归纳方式的信息抽取。

3.4 基于本体相似的信息抽取

何召卫、陈俊亮利用本体相似和本体的推理能力，应用本体把信息抽取目标文档描述为特殊的本体格式，用受限本体相似度计算本体相似度，采用机器学习理论对本体进行分析和处理，进行信息抽取[12]。

4 总结

在基于本体的Ｗｅｂ信息抽取中，早期的信息抽取仅仅进行命名实体识别，抽取概念 （或实体）；现在的信息抽取常根据ＨＴＭＬ的结构信息，如表格，对半结构化的文档，利用文档对象模型树、包装器、框架、事件、语法分析树等抽取概念的属性信息，即抽取数据文档信息；以后的信息抽取，不仅需要抽取概念、概念属性，还需要抽取概念之间的关系，包括父子关系、包含关系、属性关系、实例关系，还包括一般定义的关系，这需要利用实体关系的识别、主题识别、自然语言处理、句法信息等通过规则的方法或基于统计的方法来实现。

参 考 文 献

[1] T·R·Gruber·Toward principles for the design of ontologies used for knowledge sharing·Presented at the Padua workshop on Formal Ontology¨March1993¨1 ater published in International Journal of Human－Computer Studies¨1995¨43（4－5）：907－928

[2] 李凌志¨张玉婷·基于本体的信息集成研究·情报杂志¨2008（1）：68－71

[3] 王志华,魏斌,李占波,赵伟.基于本体的Web信息抽取系统[J].计算机工程与设计2012,33(07):2634-2639.DOI:10.16208/j.issn1000-7024.2012.07.026.

[4] Grishman R, Information Extraction: Techniques and Challenges. In M-T. Pazienza, editor, Information Extraction: a Multidisciplinary Approach to an Emerging Information Technology, Springer, Berlin, 1997.

[5] Wen Zhang,Taketoshi Yoshida,Xijin Tang. Using ontology to improve precision of terminology extraction from documents[J]. Expert Systems with Applications,2009,7: 9333-9339.

[6] 刘耀¨穗志方·领域Ontology 概念描述体系构建方法探析·大学图书馆学报¨2006（5）：28－33

[7] 徐静¨孙坦¨黄飞燕·近两年国外本体应用研究进展·图书馆建设¨2008（8）：84－90

[8] 柳佳刚¨陈山¨贺令亚·基于本体和DOM相结合的 Web 信息抽取器·现代图书情报技术¨2009（5）：44－49

[9] 徐东兴 ．基于 Ｇａｔｅ框架的信息抽取系统的研究与实现 ．华东师范大学学位论文，２００７

[10] 陆科进，李新颖 ． 基于 Ｏｎｔｏｌｏｇｙ的文本信息抽取 ． 计算机应用研究，２００３ （７）：４６－４８．

[11] 周明健，高济，李飞 ． 基于本体论的 Ｗｅｂ信息抽取 ． 计算机辅助设计与图形学学报，２００４ （４）：５３５－５４１

[12] 何召卫，陈俊亮 ． 基于本体关系匹配的信息抽取 ． 计算机工程，２００７ （２１）：２０７－２０９