本体构建工具设计与实现*

李晓瑛 李军莲 李丹亚

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 从基于已有资源的本体重用与本体转化的本体构建需求出发，结合国内外技术进展和本体构建工具研究现状，以 WebProtégé 类库为基础进行本体构建工具代码开发与功能实现。在实现过程中注重工具自动处理与人工审核相结合的交互模式，收集具体的样例数据验证所开发的本体构建工具可行性和易用性。

〔关键词〕 本体；本体构建；知识组织；WebProtégé

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2017.08.013

Design and Realization of Tools for Ontology Construction *LI Xiao－ying, LI Jun－lian, LI Dan－ya, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 The paper begins with the ontology construction requirements based on ontology reuse and transformation of existing resources, combines the current situations of study on technical progress and tools for ontology construction at home and abroad, and conducts code development of tools for ontology construction and function realization based on WebProtege class library. In the process of realization, pay attention to the interactive mode of combination of automatic tool processing and manual review, and collect specific sample data to verify the feasibility and usability of applying the developed tools for ontology construction.

〔**Keywords**〕 Ontology; Ontology construction; Knowledge organization; WebProtégé

1 引言

本体（Ontology）是共享概念模型的明确的形式化规范说明[1]，其诞生的意义在于能够在知识与语言的层面以及语义和知识的角度上对信息系统进行深刻的描述和阐述。在网络信息化时代，本体已成为语义 Web 发展的基础与核心，在很多应用中起到重要的知识组织、知识分类、知识管理和知识推理等积极作用[2]，并且逐渐呈现出学科领域更广泛、内容结构更专业、目标针对性更强等特点[3]。在人工智能领域，本体的诞生和发展一直受到重现，在知识工程、信息检索、数字化图书馆、信息异构化处理及语义 Web 等领域得到了广泛的应用，已成为众多领域研究对象和应用的基础和前提。因此，其设计与构建是否完备关系重大。就本体的构建过程而言，一般涉及两个场景：从零构建和基于已有外部资源的本体复用与本体转化。相比而言，鉴于目前已存在许多具有较高影响力的领域知识组织成果与资源（如叙词表、主题词表、本体等），而抛开现有成果的本体从零构建不仅费时费力，而且未必会取得更显著的效果，因此基于已有资源的本体重用与本体转化是当前本体构建的主流方法，

〔收稿日期〕 2017－01－11

〔作者简介〕 李晓瑛，副研究员，博士，发表论文 10 余篇。

〔基金项目〕 中央级公益性科研院所基本科研业务费“生物医学术语服务系统建设关键问题研究”（项目编号：15R0109）。

受到越来越多的关注与青睐。本文即是针对这一本体构建场景中涉及的本体映射、本体裁切、本体合并、本体语义丰富及本体可视化 5 个环节，在文献调研及需求分析基础上，开展本体构建工具设计与实现研究，并且有机地嵌入到具有较高影响力的本体构建平台中，以期更好地服务于各种类型的本体构建实践。

2 国内外本体构建工具研究现状

2.1 各种工具优缺点

随着本体在人工智能、知识工程、语义 Web、数据库设计、电子商务、信息检索和抽取领域的广泛应用，诸多本体工具不断涌现。据统计，截至 2004 年 7 月 14 日，有记录可查的本体工具就达到 96 种[4]。严格说来，这 96 种工具并非都适合用来构建本体，它们各有所长，也各有缺陷；在诸多参差不齐的本体工具中，选择一个合适的工具来构建本体会事半功倍。在国内，武汉大学信息管理学院徐国虎等依据易获取性、是否自带有帮助文档或示范本体、是否支持 Unicode 字符集、工具名称在文献或网页中的使用率、工具版本更新时间及周期、是否支持国际本体标准格式共 6 个主要判断标准，从这 96 种工具中遴选出 8 种比较成熟且具有较高影响力的本体构建工具：Ontolingua Server、Ontosaurus、WebOnto、Protégé[5]、OntoEdit、WebODE、OILED 及 DUET，并且运用一种科学的评价体系对这 8 种本体构建工具的优缺点进行了全面的分析与比较[6]。同期，中国标准化研究院李景依据本体构建工具是否可免费使用、是否具有英文版本、是否支持 Unicode 字符集、最新更新时间、输入输出是否支持国际本体标准格式、是否提供可视化视图以及工具名称是否经常出现在文献或网页中这 7 条标准，并充分考虑其适用性，详细介绍了 5 种主要的本体工具：Ontolingua Server、Ontosaurus、WebOnto、Protégé 以及 OntoEdit[7-8]。另外，在上述两项研究之后才逐渐产生的 NeON 本体工具集[9]，不仅提供了开源的本体工程环境，而且支持本体工程生命周期管理，也是一种比较适用的本体构建工具。

2.2 Protégé 优势

相比而言，上述多种本体构建工具中，目前使用最广泛、最受关注的是由美国斯坦福大学生物医学信息研究中心开发的 Protégé，其显著优势主要体现在如下几个方面[10]：（1）作为一种开放资源，提供在线及本地两种使用模式，且支持用户免费获取其本地版工具。（2）提供平台手工编辑本体及基于 Java 编程语言自动生成本体（即 WebProtégé[11]）两种不同的本体工程环境，操作界面及 Java 源码风格简洁友好，易学易用。（3）支持本体网络语言（Ontology Web Language，OWL）、资源描述框架（Resource Description Framework，RDF）、可扩展标注语言（eXtensible Markup Language，XML）等多种国际标准本体文件格式存储，且提供在不同格式之间相互转化功能。（4）本地版工具集成了 OWLViz、OntoGraf 等多种可视化插件，便于用户直观浏览本体内容结构，且支持可视化图形结构的多方位调整功能。（5）不断更新，功能日趋完善，已受到全球 29.1 万用户的信赖。

2.3 Protégé 仍需完善之处

然而，就基于已有外部资源的本体重用与本体转化等本体构建活动与过程而言，Protégé 的已有功能仍需继续完善，具体为：（1）同时仅能打开一个本体，不支持多个本体之间的映射、裁切、合并等本体复用活动。（2）无法兼容外部非本体资源，由此限制了自动扩充本体概念及关系的语义丰富活动。（3）本地版工具虽然兼容了 OWLViz、OntoGraf 等多种可视化插件，但仅有 OntoGraf 表现出较好的兼容性；而 OntoGraf 插件仅提供对可视化结果的 Graph 图形保存功能，且暂不支持与 BMP、PNG、JPEG 等常见图形格式进行交互。因此，本研究将以 WebProtégé 本体构建工具为基础，面向基于已有外部资源的本体重用与本体转化等本体构建活动与过程，研究本体构建工具的本体映射、本体裁切、本体合并、本体语义丰富及本体可视化功能；此外继承 WebProtégé 界面风格及功能展现方式，以 Tab 模块封装所实现的本体构建工具，无缝地嵌入到

WebProtégé 平台中，从而与现有的本体构建功能形成一个有机整体。

3 本体构建工具需求分析与功能设计

3.1 概述

在充分考虑基于已有外部资源的本体重用与本体转化的本体构建活动实际需求及深入调研 WebProtégé、NeON 等具有较高影响力的本体工具已有功能之后，本研究从实用性、可靠性、安全性、可操作、可扩展、可维护等基本设计原则出发，通过 Protégé 本体调用和访问的开放接口，提供基于已有外部资源的本体重用与本体转化等本体构建活动与过程所需的本体裁切、本体映射、本体合并、本体语义丰富及本体可视化工具的基本功能，见图 1，并且支持各本体工具与 WebProtégé 平台之间的交互使用。

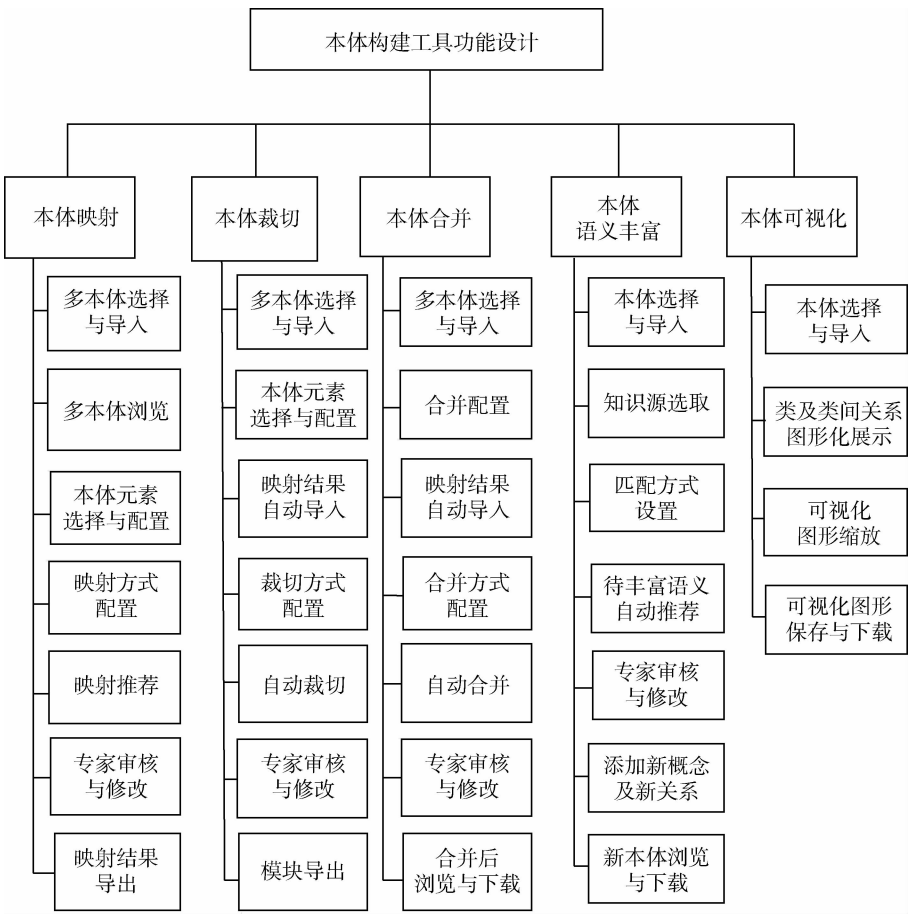

图 1 本体构建工具基本功能设计

3.2 本体映射工具设计

本体映射工具需要计算两个或多个本体之间元素的语义匹配关系，发现类和属性的对齐情况，可以实现本体对齐（Alignments），映射结果可进一步用于本体裁切或本体合并中。本体映射工具需满足的基本功能包括：（1）多本体选择与导入，支持对多种格式的待映射本体进行选择和导入；支持用户选择两个本体，彼此之间互为平行关系，不分主次。（2）多本体浏览，支持对选定的多本体的同时浏览，便于用户操作和配置相应的映射关系。（3）本体元素选择与设置，支持对本体中待匹配元素的选择与设置，包括本体类及属性；本体元素匹配支持精确和模糊两种方式。（4）映射方式配置，支持用户选择显示类映射或属性映射结果。（5）映射自动推荐，通过设计算法自动推荐满足不同配置的映射

结果，从而辅助用户初步建立映射关系，尽可能减少大量的人工操作。（6）专家审核与修改，对于工具自动推荐的映射结果，支持领域专家的审核和修改操作。（7）映射结果导出，支持导出经专家审核后的结果，便于本体裁切、本体合并等工具复用。

3.3 本体裁切工具设计

本体裁切工具的设计目标是支持将重用本体中与当前建设本体相关的部分进行拆分，用以支持本体重用活动。为实现这个目标，本体裁切工具需满足的基本功能包括：（1）多本体选择与导入，支持对多个本体进行选择和导入，并设置待裁切本体和参考本体。（2）本体元素选择与配置，通过提供功能窗口界面，支持用户主导的配置任务，选择参考本体与待裁切本体建设领域相关的类、实例和属性等本体元素，以及配置精确和模糊两种不同类型的匹配方式；工具将根据用户对本体元素的选择和配置，自动推荐相应配置下的概念和概念元素。（3）映射结果自动导入，支持复用本体映射结果，从而免去再次选择和配置本体元素环境。（4）裁切方式配置，支持用户选择裁切匹配项或未匹配项，并支持用户同时选择精确与模糊两种不同类型的匹配方式。（5）自动裁切，按照用户配置，自动完成逻辑上的本体元素裁切。（6）专家审核与修改，支持将用户手工配置、工具自动裁切后的本体和概念元素，经过系统的一致性和完整性检查后，提交到领域专家进行审核，支持修改。（7）模块导出，执行物理裁切，保留结构完整的独立本体模块；支持对裁切后的本体进行一次性模块化抽取与导出，保存为 OWL、RDF、XML 等国际标准本体文件格式，以支持持久性的复用。

3.4 本体合并工具设计

本体合并工具的设计目标为遵循在建本体的体系规范，将选择的重用本体与现有本体进行整合。基本功能包括：（1）多本体选择与导入，支持对多个本体进行选择和导入，设置待合并本体（当前在建）和参考本体。（2）本体元素选择与配置，与本体映射工具类似，支持用户选择参考本体与待合并本体建设领域相关的类、实例和属性等本体元素，以及配置精确和模糊两种不同类型的匹配方式；工具将根据用户对本体元素的选择和配置，自动推荐相应配置下的概念和概念元素。（3）映射结果自动导入，支持复用本体映射结果，从而免去再次选择和配置本体元素环境。（4）合并方式配置，支持用户选择合并精确匹配、模糊匹配、映射以及未匹配结果，也支持用户同时选择精确匹配与模糊方式。（5）自动合并，按照用户配置，自动完成逻辑上的本体合并操作。（6）专家审核与修改，支持将用户手工配置、工具自动合并后的本体和概念元素，经过系统的一致性和完整性检查检查后，提交到领域专家进行审核，支持修改。（7）合并后浏览与下载，执行物理合并，支持对合并后本体的浏览功能，用户亦可选择按 OWL、RDF、XML 等国际标准本体文件格式下载新建本体。

3.5 本体语义丰富工具设计

本体语义丰富是对现有本体添加新的概念（类）及语义关系的过程，来源包括外部本体、知识库、文献等。其主要目标是对本体概念进行丰富，在基础本体之上细化词间关系。由于在丰富过程中添加了新的概念和语义关系，所以本体结构在一定程度上也发生了改变。本体语义丰富工具的主要功能包括：（1）本体选择与导入，支持对待丰富本体的选择和导入操作。（2）知识源选取，即配置来源语义知识库，如外部知识组织体系、已有本体、预先从文献中发现的语义关系 3 元组等。（3）匹配方式设置，配置当前在建本体的类名称与知识库中概念的匹配方式，支持精确匹配与模糊匹配。（4）待丰富语义自动推荐，工具根据用户配置的知识源与匹配方式，自动推荐可用于本体语义丰富的语义知识，以及相应的来源信息。（5）专家审核与修改，支持将用户手工配置、工具自动推荐的语义知识，经过系统的一致性和完整性检查后，提交到领域专家进行审核，支持修改与删除。（6）添加新概念与新关系，即工具将专家审核后的语义关系自动添加到本体中；若该语义关系对应的概念不在当前在建本体的类名称中，应自动将其作为新本体类添加到本体

中。(7) 新本体浏览与下载，支持用户对完成语义丰富后的本体进行浏览，用户也可选择按 OWL、RDF、XML 等国际标准本体文件格式下载本体。

3.6 本体可视化工具设计

本体可视化工具主要用于帮助用户浏览本体元素，以图形结构展示本体元素之间的语义关系。基本功能包括：(1) 本体选择与导入，支持本体的选择和导入操作，便于进一步开展本体可视化。(2) 类及类间关系图形化展示，工具支持以多种图形展示本体类及类间关系，包括树形、左树形、力导向形等图形化格式。(3) 可视化图形缩放支持对本体可视化图形的放大、缩小功能。(4) 可视化图形保存与下载，支持用户将可视化展示的本体类及类间关系，以 BMP 等图形格式保存并下载。

4 本体构建工具实现与验证

4.1 概述

依据上述本体构建工具的功能设计，本研究在 WebProtégé 2. 5. 0 版本上完成了代码开发与功能实现。WebProtégé 是 Protégé 的网页在线版，支持基于 Java 的源码开发，最终应用程序可通过 Web 浏览器独立运行，且使用者无须安装任何插件。此外，结合 WebProtégé 的功能模块架构，本研究所开发的 5 个本体构建工具都被分别封装成独立的选项卡 Tab，并且无缝地嵌入到 WebProtégé 中，从而与现有的本体构建功能形成一个有机整体，便于使用者在实际本体构建过程中自由选择与切换。因此，本研究所设计并实现的本体构建工具，亦可看作是对 WebProtégé 当前版本功能的一次系统改进与完善。本研究开发本体构建工具的初衷，即是支持基于已有外部资源的本体重用与本体转化的本体构建场景中，所涉及的本体映射、本体裁切、本体合并、本体语义丰富及本体可视化等一系列活动与过程。因此，在完成基于 WebProtégé 源码架构的工具开发与功能实现后，本研究将收集一些具有代表性的测试数据进行验证。具体而言，以基于已有食品本体（简称“Food”）及其他知识组织系统，构建动物生活必需品本体（简称“Animal”）为例，展示本体构建工具的功能应用与建设成果，以期为其他本体构建活动与实践提供一些有价值的借鉴与参考。

4.2 本体映射工具实现与验证

本研究所构建的本体映射工具，其主要功能是通过精确匹配、模糊匹配两种不同的字符串匹配方式，计算当前本体与参考本体中类名之间的重叠及对齐情况，为进一步的本体复用活动提供数据基础。通过选择精确匹配，发现 Animal 本体与 Food 本体之间共有 2 个相同类，在两个本体中分别用数字 1、2 表示，详细的统计结果也显示于右上方。

4.3 本体裁切工具实现与验证

本体裁切工具的核心功能是自动批量地裁切满足一定条件的本体类、属性及实例，裁切后的“小”本体可进一步用于本体合并活动中。本体裁切工具将自动继承本体映射结果，并在满足条件的裁切项前添加箭头图标，同时将其颜色变灰，以示区别；而当使用者完成审核并点击“物理裁切”后，工具将实际执行裁切操作，并支持对裁切后本体的导出功能；另外，使用者可在右上方浏览相应的统计结果。

4.4 本体合并工具实现与验证

基于本体合并工具，可将两个本体合并为一个内容及结构更加完备的新本体，从而实现基于已有本体资源的本体构建；其中，不同本体类名称直接添加，相同本体类名称将实现属性及实例的合并；如若使用者勾选“合并子类”选项，工具将自动合并本体类的下位类及其属性、实例。工具首先加载本体映射结果并用数字标识出；合并后的本体类及其下位类以蓝色高亮显示，相应的统计结果显示在右上方。

4.5 本体语义丰富工具实现与验证

在完成基于已有本体资源的新本体构建后，往往需要借助其他外部资源实现对本体语义关系的丰富。本研究中，所支持的外部资源包括已有本体、可公开获取的主题词表、从文献中发现的关系三元

组。基于联合国粮农组织编制的 AGROVOC 农业叙词表[12]，对经过本体合并后的 Animal 本体，完成语义关系丰富的结果（以蓝色高亮显示）。此外，使用者可双击新添加的本体类，查看可供进一步丰富本体的语义关系并在选择确认后添加到本体中。

4.6 本体可视化工具实现与验证

本体可视化工具支持以图形格式浏览本体类、等级结构及其他语义关系。目前，在 Web 前端实现可视化的技术主要包括 JavaScript、Flash、Silverlight、JavaApplet 等，其中 JavaScript、Flash 相对而言绘图速度较快，是 Web 可视化的首选。由于使用 JavaScript 不需要插件（但 Flash 需要），且用户体验较好，因此本研究选择 JavaScript。而在相关的 JavaScript 开源类库中，D3（Data - driven document）不仅对当前绝大部分浏览器的兼容性较好，且已成功应用于知识组织系统的可视化动态交互展示[13]，为本研究开发本体可视化工具奠定了一定基础。通过深入调研与对比分析，本体可视化工具实现 3 种图形展示功能：树形（自顶向下展示等级层次），左树形（自左向右展示等级层次，见图 2），力导向形，见图 3；相对而言，树形与左树形适合浏览本体的等级结构，而力导向形借助于颜色与节点大小，更有助于突出本体的不同分支（颜色各异）及同一分支内部的同位兄弟关系（节点尺寸相同且聚集展示）。

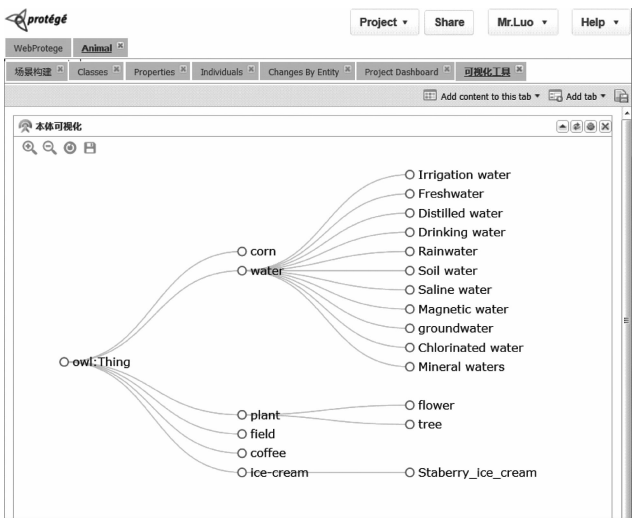

图 2 本体可视化工具功能示例（左树形）

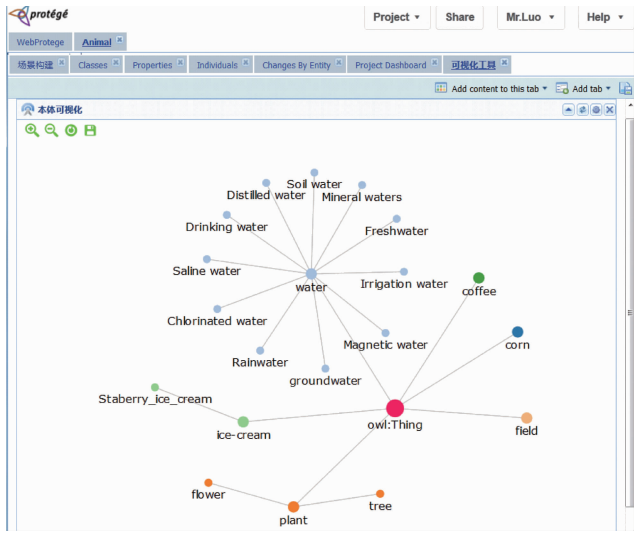

图 3 本体可视化工具功能示例（力导向形）

5 结语

本研究从基于已有资源的本体重用与本体转化的本体构建需求出发，以 WebProtégé 为基础，实现了本体映射、本体裁切、本体合并、本体语义丰富及本体可视化工具的功能设计及程序开发；并且选取样例本体作为测试，在展示本体构建工具建设成果的同时，完成了对其相应功能的验证，从而为各领域本体的构建提供了一定的工具支持与实践参考。此外，本研究沿用 WebProtégé 的功能模块架构，将所开发的 5 个本体构建工具无缝地嵌入到 WebProtégé 中，从而与其现有功能形成有机整体。因此，本研究也可看作是对 WebProtégé 当前版本功能的一次系统改进与提升，今后将在工具使用细节及功能完善上持续更新和实现。

参考文献

1 Neches R, Fikes R, Finin T, et al. Enabling Technology for Knowledge Sharing [J]. AI Magazine, 1991, 12 (3): 36 - 56.

2 Studer R, Benjamins VR, Fensel D. Knowledge Engineering: principles and methods [J]. Data & Knowledge Engineering, 1998, 25 (1): 161 - 197.

3 余倩. 近年来领域本体的应用新进展 [J]. 图书馆建设, 2008, (8): 95 - 99.

（下转第 68 页）

体不适症状的患者占比 14.0%，伴有精神不振、烦躁、压抑、焦虑等不良情绪症状的患者占比 36.1%，伴有记忆力下降的患者占比 3.0%（部分患者同时伴有多种症状）。

3.2.4 失眠患者病因分析 不同年龄层次的患者，导致失眠的主要因素有所不同，各个年龄段的失眠病因，见表 4。

表 4 失眠患者的主要病因

年龄段	病因
<25 岁	心理因素、压力、气血不足、肾虚、脾气虚
25 ~ 44 岁	心理因素、压力、气血不足、心肾不交、脑供血不足
45 ~ 59 岁	心理因素、气血不足、压力、肾虚、脑供血不足
>60 岁	脑供血不足、气血不足、压力、肾虚、焦虑

值得注意的是，在任何一个年龄层次，大部分患者都伴有神经衰弱和植物神经紊乱，这两种神经内科疾病可导致失眠，而失眠会加重神经衰弱和神经功能紊乱，从而导致恶性循环。

4 结语

通过统计结果可以得知，目前失眠患者中青年人占绝大多数，通过病因分析可以得知，年龄在 45 岁以下的患者主要是由心理因素和压力引起失眠，心理因素以焦虑和紧张为主，可以推断这些不良的情绪和压力多与生活、工作和身体健康状态有关，因此对于青年人，学会释放压力、减少心理负担、控制负面情绪尤为重要，这也从侧面反映出社会竞争越来越激烈。对于 60 岁以上的患者，引起失眠的原因主要为生理原因，如气血不足、脑供血不足等，这类患者应多从养生保健方面进行调理，从而达到改善睡眠的效果。在失眠患者中，女性所占比例高于男性，这有可能是由于生理因素导致，如女性进入 40 岁后，雌激素和孕激素分泌减少，从而导致失眠；另外女性的性格相比男性更脆弱和敏感，容易受到家庭等因素的干扰而导致失眠。

参考文献

1 于娟，刘强. 主题网络爬虫研究综述 [J]. 计算机工程与科学，2015，37（2）：231 - 237.

2 杨靖韬，陈会果. 对网络爬虫技术的研究 [J]. 科技创业月刊，2010，（10）：170 - 171.

3 周宏宇，张政. 中文分词技术综述 [J]. 安阳师范学院学报，2010，（2）：54 - 56.

4 甘秋云. 基于最短路径的二元语法中文词语粗分模型的研究 [J]. 现代计算机，2013，（25）：7 - 10.

5 Onan A，Koruko，Lu S，et al. Ensemble of Keyword Extraction Methods and Classifiers in Text Classification [M]. Pergamon Press，2016.

6 张莉婧，李业丽，曾庆涛，等. 基于改进 TextRank 的关键词抽取算法 [J]. 北京印刷学院学报，2016，24（4）：51 - 55.

（上接第 58 页）

4 Michael Denny. Ontology Tools Survey [EB/OL]. [2016 - 07 - 14]. http://www.xml.com/pub/a/2004/07/14/onto.html.

5 Stanford University. Protégé [EB/OL]. [2016 - 07 - 01]. http://protege.stanford.edu/Stanford University.

6 徐国虎，许芳. 本体构建工具的分析与比较 [J]. 图书情报工作，2006，50（1）：44 - 48.

7 李景. 主要本体构建工具比较研究（上）[J]. 情报理论与实践，2006，29（1）：109 - 111.

8 李景. 主要本体构建工具比较研究（下）[J]. 情报理论与实践，2006，29（2）：222 - 226.

9 European Commission's Sixth Framework Programme. NeON Toolkit [EB/OL]. [2016 - 07 - 01]. http://www.neon - toolkit.org/wiki/Main_Page.html.

10 李晓瑛，李丹亚，夏光辉，等. 肿瘤本体构建研究 [J]. 数字图书馆论坛，2015，（8），37 - 42.

11 Stanford University. WebProtégé [EB/OL]. [2016 - 03 - 10]. https://github.com/protegeproject/webprotege.

12 Food and Agriculture Organization AGROVOC [EB/OL]. [2016 - 04 - 10]. http://aims.fao.org/vest - registry/vocabularies/agrovoc - multilingual - agricultural - thesaurus.

13 张运良，张兆锋，张晓丹，等. 使用 D3.js 的知识组织系统 Web 动态交互可视化功能实现 [J]. 现代图书情报技术，2013，（7/8）：127 - 131.