# 基于语义分析的软件需求提取技术研究*

陈 辞
（海军驻大连地区第一军事代表室 大连 116000）

**摘 要** 针对信息系统开发过程中需求来源种类多、需求描述不规范、人工软件需求分析效率低等问题，借助自然语言处理等技术，研究软件需求的表示和组织，构建基于语义分析的软件需求提取技术，实现对形式化、非形式化等多种软件需求文档的语义提取，辅助用户理解需求。为后面提取业务、信息活动、能力效果等顶层设计数据，以及软件需求形式化验证提供技术基础。

**关键词** 软件需求提取；语义分析；需求表示和组织
**中图分类号** TP311.52 **DOI**：10. 3969/j. issn. 1672-9730. 2020. 05. 028

# Research on Software Requirements Extraction Based on Semantic Analysis

CHEN Ci
（No.1 Navy Force Representative Bureau in Dalian, Dalian 116000）

**Abstract** In view of the problems of many types of requirement sources, irregular requirements descriptions, and low efficiency of manual software requirement analysis in the development of information systems, with the help of natural language processing and other technologies, the software requirement extraction is researched based on semantic analysis to achieve formal and informal semantic extraction of various software requirement documents to assist users in understanding requirements. This research, as a technical foundation for formal verification of software requirements, provides top-design data for subsequent extraction of business, information activities and capability effects.

**Key Words** software requirement extraction, semantic analysis, representation and architecture of requirement
**Class Number** TP311.52

## 1 引言

在信息系统研制过程中，软件需求横向可以划分为业务需求、用户需求和功能需求三个层次，纵向可以划分为功能性需求和非功能性需求两个类型。为此，人们提出了面向对象的方法、面向目标的方法和面向主体和意图的方法等不同的需求分析方法。

非形式化需求描述，多采用自然语言或类自然语言方法描述，容易建立和理解，但定义缺乏严格，存在二义性；半形式化描述虽具备了结构化特征，提高了需求描述的规范性，但形式定义仍不够严格，且不易推理和检验；形式化描述多建立在数学基础上，具有严格的形式定义，便于进行推理和检验，但多数不易理解且难扩展[1]。

而多源异构文本需求中蕴含着大量的语义要素，它们具有不同的层次和角度的特性[2]。因此，在对文本需求的实体、属性、关系和事件进行抽取和发现，需要构建源异构文本需求多层次多角度语义要素模型，以适应不同场景下需求文本中语义的不同粒度与视角[3]。

因此，本文提出了基于语义提取的软件需求分析技术，主要是在统一的软件需求表示和组织框架下，进行软件需求语义的提取。技术框架如图1所示。

* **收稿日期**：2019年12月3日，**修回日期**：2020年01月20日
**作者简介**：陈辞，男，硕士，高级工程师，研究方向：作战指挥系统，自然语言处理和需求工程。

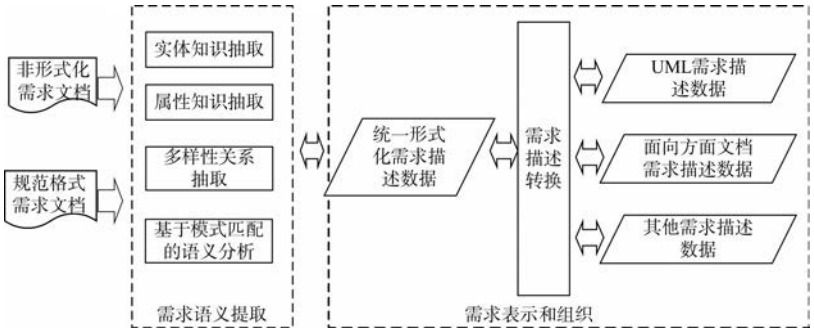


图1　基于语义分析的软件需求分析技术技术框架

## 2　软件需求表示和组织

针对软件需求描述中存在的多层次、多角度的特点，本文提出了一种多层次多角度的需求统一表示和组织框架。包括三方面内容：1）构建一个统一的、形式化的、规范化的功能需求描述模型；2）功能需求描述方式与非功能需求描述方式之间的转换方法；3）核心需求描述与非形式化、半形式化以及形式化的需求描述之间的转换机制[4]。

### 2.1　需求统一描述模型

同领域的信息系统具有极高的相似性，特别是在需求和功能上，共同点更为著性。这就决定了它们的解决方案也具有稳定性和内聚性。通过统一的本体描述方法，可以将同领域的特征模型和体系结构建模有机结合。合理定义基于本体语言需求描述的领域约束，以及可操作的验证规则，就可以借助本体与生俱来的严格推理能力，自动地建立和验证其领域内的需求描述模型。这种方法就兼顾了建模的高效性和验证准确性。

作为一种语义相关的知识概念模型，本体涵盖了信息系统领域中的对象、概念等实体及其之间的关系。本体有五个建模元素，分别是类、关系、函数、公理和实例，其概念如下[5]。

类：也可称为概念，它可以泛指任何事物，包括事件、功能、策略和推理过程等。

关系：用于刻画领域中概念间的交互作用，形式上通常被定义成 $n$ 维笛卡儿积的子集 $R:C_1\times C_2\times\ldots\times C_n$。基本关系包括：实例与概念间的关系、属性关系、部分与整体的关系，以及继承关系。

函数：本体概念范畴中的一类特殊关系。在函数中，前面的元素是可以唯一决定后面的元素的，形式化地可定义为 $Function:C_1\times C_2\times\ldots\times C_{n-1}\rightarrow C_n$。

公理：用于表示恒真断言。

实例：指代隶属于某一概念的对象，代表元素。

OWL的语义表达能力极为卓越，它在处理文档内容和文档其他丰富的相关信息上具有很大优势[4]。OWL本体中要素与本文要建立的需求模型要素是非常贴近。因此，采用OWL语言作为统一需求描述语言。

### 2.2　非功能需求建模

软件需求分析和规约中，通常会对功能和非功能需求分而治之。将非功能需求的描述语句结构化处理后，分别列在软件需求文档中的对应章节，是表示需求的通常做法。然而，该方法本质上对主要采用非形式化的自然语义来描述需求内容，避免不了不一致性、二义性等问题的产生[6]。

为了易于展现概念的分解、聚合，经常会用树结构和表结构等结构化的方式来表示非功能需求的层次关系。面向目标的分析和建模方法，也是一种能够深入分析、细致刻画功能需求的建模方法[7]。这种形式化本体建模方法主要采用非功能需求框架和软目标依存图来建立非功能需求表达模型，基于本体建模思想对非功能需求的概念层次、属性关系和公理化合理刻画，并采用OWL语言构建出本体模型[8]。其本体中蕴含非功能需求的概念层次关系、以及交互关系的语义表示和推理，以此来实现非功能需求自动化的分析。

### 2.3　描述间的相互转换

由于统一需求描述采用本体描述语言OWL，而UML需求描述可以用于作为输出给用户的需求形式，接下来，研究了OWL需求描述与UML需求描述之间的相互转换。

1）UML类图在OWL本体上的映射

UML类图，是通过类以及类之间的关系组成，类之间的关系主要有泛化、组合、聚合、关联和依赖等[5]。通常，UML类图可以由（概念集合，属性集合，行为集合，关系集合）等元素组成的四元组表示。它和OWL本体之间有很多概念上的相似或等价。例如，在UML类图中，关联可通过属性表示，

而在OWL中,关联也被定义为属性。UML类图在OWL本体上的映射规则包括有标识符规则、属性规则、取值范围规则、行为规则、关联规则、泛化规则、聚合规则、组合规则等。

2)OWL需求描述向UML需求描述的转换

基于软件开发领域的知识范畴,UML元模型具备紧密联系用户需求和软件设计的能力[5]。这实质上是对应用本体概念的实现,以及概念关联在面向对象软件设计方法上的映射[8]。为了实现从需求模型到面向对象模型的转变,需要将面向对象模型中对象、类和关联等概念从基于应用本体的需求模型中合理提取。OWL需求描述向UML需求描述的转换也有详细和严格的规则[5]。

3)统一需求描述与面向方面文档需求描述的相互转换

本体需求描述到面向方面文档需求描述的相互转换主要包括两个方面:第一个方面是从需求描述模型结构本身开始,通过对本体需求描述模型和面向方面文档需求描述模型进行深入比对,找出模型上联系和差异,并进行转换;第二个方面是从最初的建模条件出发,分析两个模型建模思想,从中找出两者的关联,并进行转换。

## 3 软件需求语义提取技术

在多层次多角度的需求统一表示和组织框架,通过定义句型模式规则集,来严格规范需求结构化描述。在对需求文本进行语义分析时,根据匹配到的句型进行语义分类。并基于前期识别出实体类,实体类的属性和操作,角色和用例[9]等建模元素,形成需求结构化描述。

### 3.1 实体知识抽取

面向多源异构的文本需求,本文提出了多层次多角度的实体知识抽取技术,主要包括1)基于结构化需求的实体抽取与初始需求实体库构建;2)基于半监督学习和模板匹配的非结构化需求中实体抽取与需求实体库补全;3)基于半监督学习的Bi-LSTM(双向长时短时记忆循环神经网络)与CRF(条件随机场)[10]结合的需求实体抽取模型三部分内容。多源异构文本需求多层次多角度实体抽取技术总体技术框架如图2所示。

1)基于结构化需求的实体抽取与初始需求实体库构建

信息系统的研制过程中,需求规格说明等软件开发文档是遵照标准或格式来编制,其部分需求描述具有结构性。可直接从这些结构化需求文本中获取实体,并构建初始需求实体库。

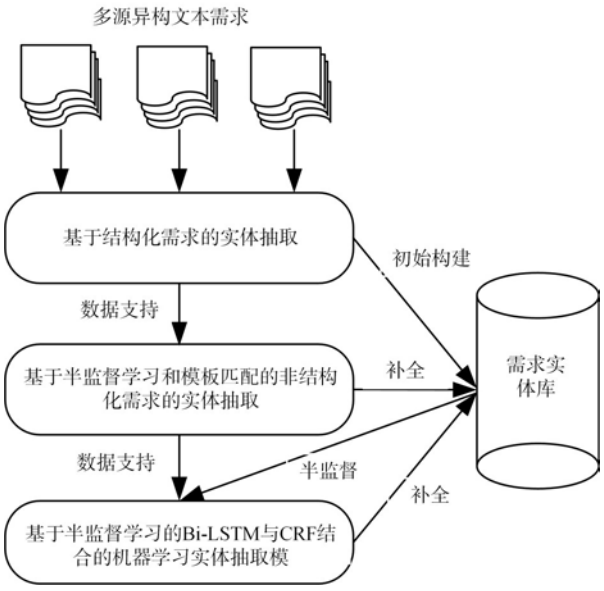


**图2 多源异构文本需求多层次多角度实体抽取技术总体架构**

2)基于半监督学习和模板匹配的非结构化需求文本中实体抽取与需求实体库补全研究

初始需求实体库虽然准确率较高,但是由于大量需求的描述不具有结构性,导致需求实体覆盖率低,不能自动提取。通过采用1)中的初始需求实体库为半监督种子,基于半监督学习和模板匹配的非结构化需求文本中需求实体的自动抽取,对需求实体库进行补全。

基于半监督学习和模板匹配的非结构化需求文本中实体抽取算法如下:

算法1:需求实体自动抽取与需求实体库补全算法
输入:初始需求实体库$E$,非结构化需求文本$C$
输出:需求实体
步骤1:遍历所有非结构化需求文本数据集$C$,获取符合顿等模式的所有需求描述语句集合$S$
步骤2:foreach $s_i \in S$ do
步骤3: 抽取$s_i$中符合顿等模式的同类词集合$T$
步骤4: foreach $t_i \in T$ do
步骤5: if $t_i$ not in $E$ do
步骤6: add $t_i$ to $E$
步骤7: end if
步骤8: end foreach
步骤9:end foreach

3)基于半监督学习的Bi-LSTM与CRF结合的需求文本中实体抽取模型研究

前面两种方法需要人工参与,其覆盖率提升缓慢。对此,采用需求实体库中的实体作为半监督学习的种子,对多源需求文本进行少量标注,构建

Bi-LSTM 与 CRF 方法结合的实体抽取机器学习模型，实现多源需求文本中的需求实体进行自动识别和抽取。

首先，利用词向量作为输入，使用 Word2vec 技术和 CBOW（Continue Bag of Words）方法，对需求文本数据集进行字向量学习，具体如下：

$$p(w_i|w_c)=\frac{\exp(v_{w_i}\cdot h)}{\sum_{\hat{i}\in\bar{V}}\exp(v_{w_i}\cdot h)} \tag{1}$$

其中，$w_i$ 是需要预测的字，$w_c$ 是字 $w_i$ 的该次训练中的上下文集合，$p(w_i|w_c)$ 是已知字 $w_i$ 的上下文 $w_c$ 的条件下获得字 $w_i$ 的概率，$h$ 的表示如下：

$$h=\frac{1}{n}\sum_{i=1}^{n}v_{w_{c_i}} \tag{2}$$

其中 $v_{w_i}$ 是字 $w_i$ 的待训练字向量，$i\in(1,\bar{V})$，$\bar{V}$ 是需求数据的字典库，$n$ 是该次训练中字 $w_i$ 的上下文词个数。

对此，整个需求实体库的字向量训练目标函数如下：

$$\frac{1}{V}\sum_{i\in V}\log p(w_i|w_c) \tag{3}$$

采用负采样技术对字向量进行训练。训练完成词向量后，利用需求实体库中的实体作为半监督学习的种子，使用远程监督方法，对需求文本进行自动标注。

基于 Bi-LSTM（双向长时短时记忆循环神经网络）与 CRF（条件随机场）方法结合的实体抽取机器学习模型整体框架如图 3 所示。

图 3 中模型的第一层是 look_up 层，并利用训练好的一个句子 $x$ 的字向量（$V_i=R^d$，$d$ 是字向量的维度）作为输入。

第二层是双向 LSTM 层，对 look_up 层输入的数据 $(V_1,\cdots,V_n)$ 进行自动特征提取。其中，正向 LSTM 输出 $(f_1,\cdots,f_n)$，反向 LSTM 输出 $(b_1,\cdots,b_n)$。对正反两两的 LSTM 输出单元进行按位置拼接形成 $C=(f_1\oplus b_1,\cdots,f_n\oplus b_n)$，$C\in R^{n\times 2d}$，并执行 Dropout 技术，并执行一个线性操作将 2d 维转为 K 维，记为：$P=(p_1,\cdots,p_n)\in R^{n\times K}$，$p_i\in R^K$。

第三层是 CRF 层，将 $p_i\in R^K$ 的每一维作为该字分到哪一个标签下的一个打分。CRF 层的参数是一个 $(K+2)\times(K+2)$ 的转移矩阵 $A$，$A_{ij}$ 表示从第 $i$ 个标签转移到第 $j$ 个标签的转移得分。那么一个句子的标签序列 $y=(y_1,\cdots y_n)$ 对于句子序列 $x=(x_1,\cdots x_n)$ 打分为

$$score(x,y)=\sum_{i=1}^{n}p_{i,y_i}+\sum_{i=1}^{n+1}A_{y_{i-1}y_i} \tag{4}$$

对得分使用 Softmax 得到归一化后的概率：

$$P(y|x)=\frac{e^{score(x,y)}}{\sum_{i=1}^{K}e^{score(x,y_k)}} \tag{5}$$

模型在训练时，目标函数使用最大化上述公式的对数似然函数：

$$O=\log(P(y|x)) \tag{6}$$

当模型训练完成后，采用动态规划的 Viterbi 算法来进行测试和预测：

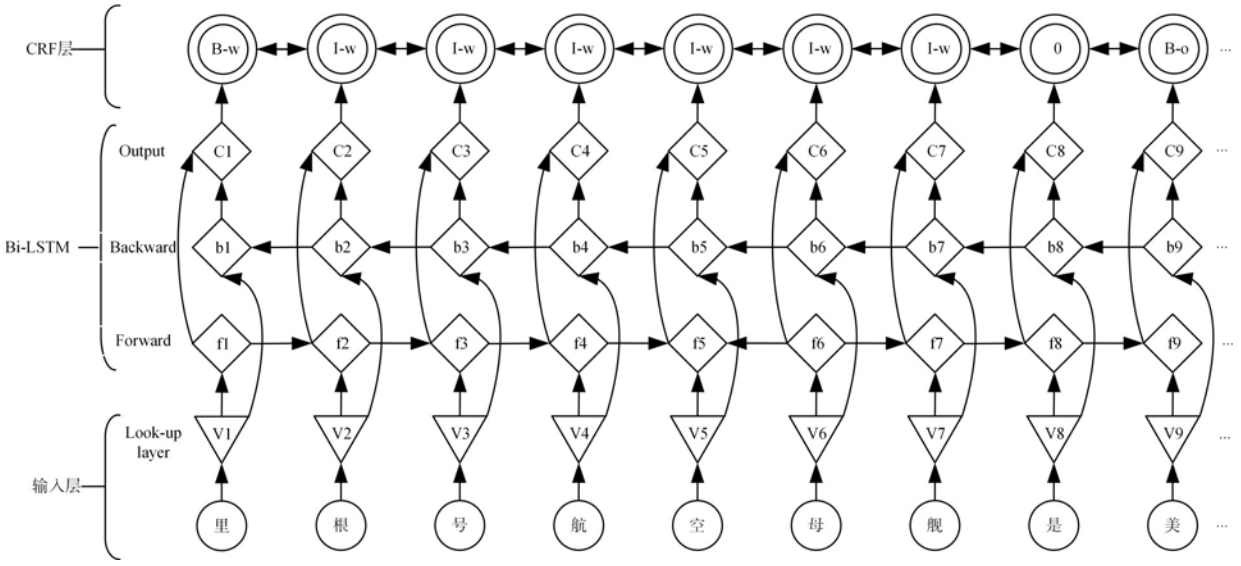


图 3　基于 Bi-LSTM 与 CRF 方法结合的需求实体抽取机器学习模型整体框架示意图

$$y^* = \arg\max_{y}(score(x, y')) \quad (7)$$

通过基于Bi-LSTM与CRF方法结合机器学习模型可以抽取大量的需求实体,进而可以对需求实体库进行实体的补全。

### 3.2 属性知识抽取

需求属性抽取的目标是从多层次多角度的数据源中采集需求实体的属性信息,实现对实体知识的整体勾画,为需求实体关系的抽取、需求知识的融合、表达提供基础和实现的载体[11]。而实体属性可以看作实体与属性值之间的一种名词性表述关系。实体属性的抽取可看作是一种特殊的关系抽取。

多源异构需求文本的属性抽取可以从两个方面来开展研究:1)直接属性抽取;2)基于关联语义链的语义属性抽取。直接属性抽取是利用数据挖掘等技术,从不同来源的结构化、半结构化和非结构化的需求文本中抽取实体属性知识。基于关联语义链的语义属性抽取则主要利用实体的关联语义链,抽取与实体高度关联的语义词语作为属性。

对于结构化、半结构化需求文本,利用结构信息抽取出实体对应的属性候选,然后利用关联规则挖掘Apriori算法,选取置信度高的候选作为实体对应的属性。对于非结构化需求文本,利用句法分析对需求文本进行处理,利用实体识别结果、句法分析结果和属性词典产生每个实体对应的属性候选,然后融合句法语义特征、候选属性本身的语义特征、候选属性与实体间的相对位置特征等多种特征来对每个候选属性进行打分,分值高的属性候选即为抽取结果。

### 3.3 多样性关系抽取

需求的关系包括:需求实体或属性的分类层次关系、部分-整体关系、相似关系、互斥-协同等抽象关系以及属性关系等。其中属性关系可以直接通过属性抽取的方法得到,而部分-整体关系和分类层次关系可以归纳为需求实体或属性的is-a关系。因此选取以下面三种具有代表性的需求实体或属性间的关系进行关系抽取的研究:is-a关系、相似关系、互斥-协同关系[12]。

采用基于实体注意力的深度神经网络实体关系抽取模型,来抽取需求实体的关联关系。并借助语义信息分析技术,预测需求实体间语义关系的类别,从而达到抽取多源异构文本中实体和关系的目的。具体来说,首先采用双向LSTM对需求实体所在的文本的上下文进行建模,随后利用实体注意力模型对辨别语义关系过程中起不同作用的语义特征分配不同权重,然后将不同需求实体相关特征的计算结果通过softmax归一化映射为每一类语义关系对应的概率,模型的参数通过梯度下降算法进行优化。

### 3.4 基于模式匹配的需求语义分析

针对需求描述的不确定性和语义表达多样性,通过定义需求句型模式规则集,来规范需求结构化描述。对需求文本进行语义分析时,根据匹配到的句型进行语义分类,并基于前期识别出实体类,实体类的属性和操作,角色和用例等建模元素,形成需求结构化描述[6]。

然后,利用模式匹配的方法,把从原文本中抽取到的句型模式和XML的结构相结合起来,自动化地生成需求条目。将本技术拆解为以下三个主要模块来实现:1)基于XML的句型模式规则集的定义;2)基于有监督机器学习方法的句型自动发现;3)基于模式匹配的语义分析。下面将详细解析每个模块使用的技术与方法。

1)基于XML的句型模式规则集的定义

XML是一种可拓展标记语言,它是一种具有结构性的标记语言,提供统一的方法来描述结构化数据,具有强大的描述能力与灵活的应用场景。

通常来说,需求主要描述了功能、性能、可靠性、出错处理、接口等方面的要素。而用户在对这些特定方面的描述常常会使用一些特定的语法结构,如用户在描述性能需求时通常会使用类似“XXX功能的响应时间应在XXX毫秒之内”的语句。可根据用户需求,定义一套句型模式规则集,用于后续的语义分析和需求结构化描述的生成,并采用XML来结构化地存储被提取的需求条目。

2)基于有监督机器学习方法的需求句型自动发现

对于给定的需求文本,通过文本句子中每个语素之间的联系,构建出语义树,以实现自动化地分析出句子的句型。在3.1中,已经将文本中的关键词提取出来,接下来就需要进一步分析每个句子,获取句中关键词间的关系。

一般可将关系抽取技术划分为无监督、半监督和有监督等三种学习方法[12]。由于软件需求文本中,原始数据往往是杂乱、弱逻辑化的文本,使用半监督和无监督的方法聚类的提取方法效果较差,而有监督的学习方法能抽取更有效的特征,从而获得更高的准确率和召回率。基于有监督机器学习方法的需求句型自动发现,主要步骤如图4所示。

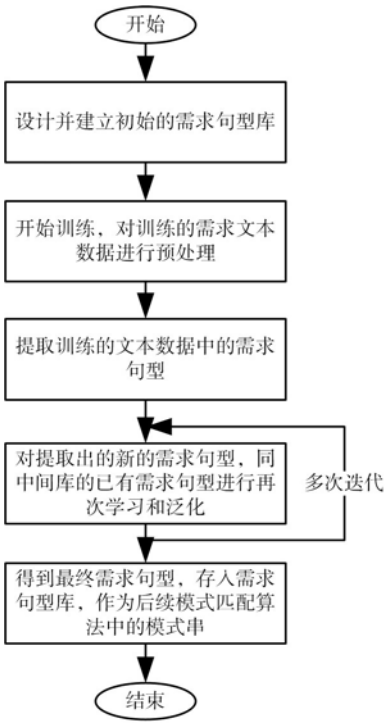


**图4　基于有监督机器学习方法的需求句型自动发现算法**

3)基于模式匹配的语义分析

模式匹配起初是一种用于进行字符串匹配的简单算法。给定两个字符串:其中,匹配子串为“ab”,被匹配的母串为“abbaabcdeeefffab”。从母串中找出所有与子串相同的部分,就是模式匹配的主要工作,其目的寻找母串和子串之间存在的映射关系。在逻辑层面,也可以将它理解为语义关系的获取。

前期通过基于有监督机器学习方法的需求句型自动发现方法,得到了需求句型模式串后,进一步采用基于规则的模式匹配方法,对软件需求文本进行语义分析。这些需求句型模式串被提取出,并以XML树形结构的来表示,并融入包括数据类型、成员名称以及数据结构等信息,用于对模式匹配全过程的协同控制。将所有可能的树形结构作为模式匹配中的文本串,进行多次对模式进行遍历匹配,匹配结果中效果最好的作为最终的提取结果,并同样以XML的形式结构化地存储到中间数据库中,形成基于语义分析的需求分析结果。

## 4　结语

针对信息系统开发过程中面临的需求变化频率高、人工提取需求效率较低等问题,本文设计多层次多角度需求统一表示和组织框架,实现需求的统一形式化描述以及与其他描述方法之间的转换,在此基础上通过实体知识抽取、属性知识抽取、多样性关系抽取和基于模式匹配的需求语义分析技术的实现,最终完成了基于多层次多角度语义分析的需求提取技术的构建,为信息服务系统的分析、设计、开发、测试和维护的全流程提供有力支撑。

## 参 考 文 献

[1] 李强. 基于本体的构件形式化描述与检索研究[D]. 昆明:昆明理工大学, 2017.

[2] Antoine Bordes N U, Alberto Garcia-Dur´an, Jasonweston, Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data[J]. Proc of NIPS. Cambridge, MA: MIT Press, 2013: 2787-2795.

[3] 漆桂林, 欧阳丹彤, 李涓子. 本体工程与知识图谱专题前言[J]. 软件学报, 2018 (10): 5-6.

[4] 江东宇, 康达周, 王顺. 基于本体的需求分析和软件体系结构设计研究[J]. 计算技术与自动化, 2017, 036 (02): 129-135.

[5] 王宇华, 印桂生. 基于本体的需求模型到UML模型转换方法[J]. 哈尔滨工程大学学报, 2013, 33(06): 735-740.

[6] 胡海波. 非功能需求交互的语义建模和自动化推理[D]. 重庆:重庆大学, 2012.

[7] 李代遗. 实例软件非功能需求知识构建与推荐研究[D]. 昆明:云南大学, 2019.

[8] 潘一之. 基于领域本体的需求模型到UML模型的转换方法研究[D]. 长沙:湖南大学, 2016.

[9] 郑梦悦, 秦春秀, 马续补. 面向中文科技文献非结构化摘要的知识元表示与抽取研究——基于知识元本体理论[J]. 情报理论与实践, 2020(02):157-163.

[10] 项威. 事件知识图谱构建技术与应用综述[J]. 计算机与现代化, 2020(1): 10.

[11] 苏佳林, 王元卓, 靳小龙, et al. 自适应属性选择的实体对齐方法[J]. 山东大学学报(工学版), 2020, 50 (01): 14-20.

[12] Yankai Lin Z L, Maosong Sun, Yang Liu, Xuan Zhu. Learning Entity and Relation Embeddings for Knowledge Graph Completion[J]. Proc of AAAI. Menlo Park, CA, 2015: 2181-2187.

[13] Suchanek F M, Abiteboul S, Senellart P. PARIS: Probabilistic alignment of relations, instances, and schema [J]. Proceedings of the VLDB Endowment, 2011, 5(3): 157-168.