计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(11):243−249 [doi: 10.15888/j.cnki.csa.007697]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向电网安全监测的领域本体自动构建①

于碧辉[1], 孙　思[1,2], 李　岳[1,2]

[1](中国科学院 沈阳计算技术研究所, 沈阳 110004)
[2](中国科学院大学, 北京 100049)
通讯作者: 孙　思, E-mail: 1048829739@qq.com

摘　要: 针对电力监控系统面临的网络安全问题实际需求, 开展了本体自动构建技术研究, 以现有的领域本体自动化构建技术为基础, 从非结构化文本数据提取出电网安全监测领域本体 SafeAgent, 采用机器学习、自然语言处理、关联规则等方法抽取本体概念, 挖掘概念之间的关系, 完善了领域本体自动化构建方案. 经实验验证, 本文采用的方法能以较高准确率完成领域本体的自动化构建工作, 克服对人工以及专家知识的依赖.

关键词: 本体; 自动构建; 层次聚类; 关联规则; 电网安全

引用格式: 于碧辉,孙思,李岳.面向电网安全监测的领域本体自动构建.计算机系统应用,2020,29(11):243–249. http://www.c-s-a.org.cn/1003-3254/7697.html

Automatic Construction of Domain Ontology for Power-Grid Security Monitoring System

YU Bi-Hui[1], SUN Si[1,2], LI Yue[1,2]

[1](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110004, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Aiming at the actual needs of network security problems faced by the power monitoring systems, research on ontology construction technology was carried out. Based on existing domain ontology automation construction technologies, a power grid monitoring domain ontology named SafeAgent was proposed from unstructured text data. Using methods such as machine learning, natural language processing and association rules to realize extraction of ontology concepts. Futhermore, this study accomplished the mining of relationships between concepts, and perfecting of the domain ontology automation construction scheme. The experimental verification shows that the method proposed in this study can complete the automatic construction of domain ontology with higher accuracy while overcoming reliance on human and expert knowledge.

Key words: ontology; automatic construction; hierarchical clustering; association rules; power-grid security

本体原本是一个哲学概念, 随着人工智能领域的发展, 被赋予了新的定义, 领域内学者对此进行了深入的研究, 对本体的定义也在不断发展变化, 目前被广泛认可的是 1998 年 Studer 对本体的定义: “本体是共享概念模型的明确的形式化规范说明”. 本体主要依靠概念、概念之间的关系以及公理来发挥作用, 其中关系又包括层次关系以及非层次关系[1,2].

关于本体的研究与应用主要围绕 3 个方面: (1) 对本体概念以及分类等等理论上的研究. (2) 应用在信息系统中, 包括信息组织、信息检索以及异构系统互操作问题. (3) 应用在语义网中, 在知识层提供知识重用和共享的依据. 本体可以分为 3 个层次: 上位本体、领域本体和面向应用的本体. 上位本体是可跨领域复用的本体, 为不同本体之间的逻辑组织提供保证. 领域本

① 收稿时间: 2020-04-03; 修改时间: 2020-05-18, 2020-05-27; 采用时间: 2020-06-01; csa 在线出版时间: 2020-10-29

体针对某一个特定的学科、专业或领域, 表述适用于这一范围内广泛使用的概念和关系. 面向应用的本体是为了特定应用构建的本体知识库.

如今, 本体构建主要有 3 种方法, 由领域专家和本体专家参与的手动构建方法; 使用机器学习、深度学习或者自然语言处理的自动构建方法; 融合了上述两种方法的半自动本体构建方法. 然而, 手动构建本体方法中本体概念的抽取以及概念之间的关系均通过人工来定义, 依赖于本体专家的意见耗费大量人力, 时间, 而且依赖于人的主观性, 具有高度局限性. 因此, 手工构建方法逐渐被半自动化、自动化构建方法取代, 自动构建本体方法可以方便的和其他机器学习、自然语言处理领域相结合, 可以使用不同的数据源来进行构建, 文本数据具有数据来源广泛、便于获取等特点[3]. 鉴于此, 本文采用电力安全相关文本作为数据源进行领域本体的自动构建并对构建出来的本体进行评估.

1 相关研究

文献 [4] 采用形式概念分析 FCA 来进行本体构建,基于概念格的相关理论, 但是构造过程中计算代价大, 适用于小规模本体的构建研究. 文献 [5] 以叙词表为依据,针对叙词表等级结构及其包含的概念间关系开展基于叙词表的本体构建方法研究, 但是仅适合应用于医学领域. 文献 [6] 提出基于模板识别的 SSE_CMM 领域本体自动构建技术. 文献 [7] 基于维基百科等开放知识库进行本体构建,但由于这些开放知识库的异构性, 关于此类本体构建方法还处于初级阶段. 在概念抽取方面, 文献 [8] 采用 TF-IDF 公式进行相关性的判断, 得到术语在领域的相关程度, 筛选出相关性高的作为领域内概念. 文献 [9] 采用 LDA (Latent Dirichlet Allocation) 主题模型将语料中最核心的概念提取出来.

2 领域本体自动构建

依据电力行业相关规定, 结合电力监控系统的实际需求, 本文采用了电力监控系统网络安全管理平台基础支撑功能规范以及中国知网中电力监控系统网络安全相关论文作为数据集. 通过以下步骤对输入的文本数据进行处理, 从而实现领域本体文件的自动构建:

第 1 步. 文本数据预处理, 该过程将成段的文本进行分词并去除停用词;

第 2 步. 本体概念抽取, 该过程将中文词汇转换为本体的基本元素——本体概念;

第 3 步. 本体概念间关系抽取, 该过程抽取并建立本体概念间的相互关系以完成本体网络的构建.

图 1 为本文所建立的领域本体自动构建流程图, 图解本文自动构建领域本体的整体过程.

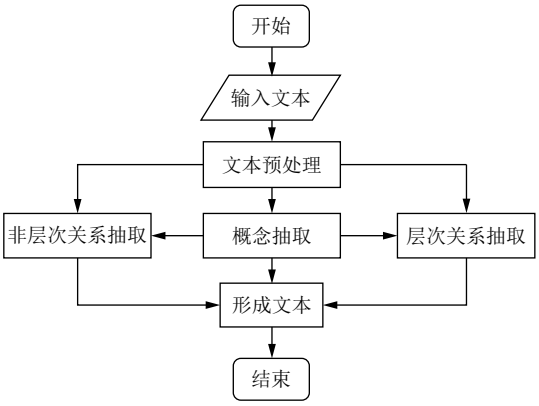

图 1 领域本体自动构建流程图

下文将对各步骤中所涉及的设计细节进行详尽的介绍.

2.1 本体概念抽取

主要有 3 种常用的概念抽取方法: 基于规则的方法、基于统计的方法和规则与统计的混合方法[10]. 本文采用基于统计的方法, 因为该方法易于扩展、不受具体领域语言学限制, 易于实现.

首先, 对文本数据进行预处理, 采用开源的 Jieba 中文分词工具对文本进行分词, 本文使用 Jieba 分词时加载自行定义的电网安全监控词典来提高分词的效果. 自定义的电网安全监控词典由搜狗细胞词库中电力词汇表、计算机词汇表以及网络工程词汇表等组成.

使用百度停用词表、哈工大停用词表、四川大学停用词表等中文停用词表组成的停用词表进行过滤.

目前, 已有多种成熟的统计学方法可实现从文本数据中抽取本体概念. TF-IDF (Term Frequency-Inverse Document Frequency) 是一种基于统计的方法, 衡量一个词语在文档中的重要程度, 词语的重要性与出现在文档中的次数成正比, 与出现在语料库里的其他文档中的频率成反比. TextRank 算法是一种用于文本的基于图的排序算法, 它的思想来源于 PageRank 算法, 把文本分为若干部分, 建立图模型, 使用投票机制对文本中的重要词汇进行排序. 不同于 TF-IDF、LDA 等方法, 该算法是一种无监督的学习算法, 不强烈依赖语料

库, 不需要对多篇文档进行学习训练, 能够有效地处理本文所使用的文本资源. 因此, 本文采用 TextRank 算法实现本体概念的抽取. 在该算法中, 单词的 TextRank 权重计算公式如下:

$$WS(V_i)=(1-d)+d\times\sum_{V_j\in In(V_i)}\frac{w_{ji}}{\sum_{V_k\in Out(V_j)}w_{jk}}WS(V_j) \quad (1)$$

其中, d 是一个人为设置的可调整参数, 经过实验调整, 本文将上式中的 d 设置为 0.85. $In(V_i)$为每个单词 i 在单句内成线性关系排列的单词的集合, 单词 i 的权重 $WS(V_i)$ 取决于在 i 之前的各点 j 组成的 (j, i) 边的权重, 以及 j 点到其他各边的权重之和. 形如 w_{ji} 的权重值由计算两个不同的文本单元同时出现在同一个文本窗口中的比率而得, 该权重的取值通常为 2. 初始化时, 每个单词的权重统一初始为 1, 经过多次计算后所有权重整体达到一致性, 分别以单个文档、单句为单位进行权重排名, 取权重排名最高的单词为关键词.

现有的 TextRank 算法主要基于统计学获取权重排名, 在部分情况下, 对文本资源中出现频次低却包含领域内关键概念的词汇抽取效果较差. 实验表明, 在应用中时常造成关键概念的遗漏, 从而导致抽取准确度存在较大的提升空间. 针对上述问题, 改进 TextRank 算法将原本单一的权重排名队列扩大为 3 个队列组成的多重权重排名队列. 通过计算权重得到原始队列后, 基于电网安全监控词典以及上下文语义关系, 统计各词语与领域内的关键词汇的关联度. 直接在词典中出现的关键词关联度置 1, 与词典中关键词产生语义关系的依照关系强弱置为 [0.2,0.9) 区间内的值. 从队尾反向搜索, 设定关联度阈值, 将高关联的词汇认定为领域关键词升至上位队列. 从队首正向搜索, 将低关联词汇认定为高频次的非关键词汇降入下位队列. 通过添加上述过程, 能够有效地提升概念抽取的准确率, 并在关系抽取过程之前过滤非关键词, 从而提高了算法整体的运行效率.

2.2 本体概念间关系抽取

本体概念间关系主要划分为两种: 层次关系与非层次关系. 层次关系主要是概念之间的父子关系[11]; 非层次关系是指除了层次关系之外的关系, 包括整体与部分之间的关系、属性关系等. 所以先进行层次关系的抽取, 再此基础上对非层次关系进行抽取. 例如 USB 是设备的子类, USB 和设备之间具有层次关系; 设备的名称和设备的编号是设备的数据属性, 设备与设备名称和设备编号具有非层次关系.

2.2.1 层次关系抽取

本体概念之间关系的抽取方法主要包括: 基于模板的方法、基于关联规则的方法、基于词典的方法以及层次聚类的方法. 聚类方法的思想在于根据事物的属性最小化类内距离, 最大化类间距离, 将一组具有异同特征的对象数据集依据特征的相似性分类为相似的对象类, 同一分类下的对象具有相对的高度相似性, 这一思想适用于本文中对本体概念进行层次关系的抽取过程. 层次聚类根据不同的聚类策略又分为自顶向下的分裂方式和自底向上的凝聚方式, 凝聚方式将每个概念作为一个簇, 计算概念之间的相似度不断的进行合并, 将簇不断扩大直到所有概念都合并为一个簇; 而分裂的方式和他相反, 初始情况将所有概念作为一簇, 依据相似度将概念细分, 不停迭代直到概念各成一簇为止. 两种不同方式如下图所示. 本文采用凝聚的层次聚类方法将 1.1 小节得到的领域内概念向量化, 根据向量之间的相似度对概念进行聚类, 抽取它们之间的层次关系, 方法的核心思路如图 2 所示意.

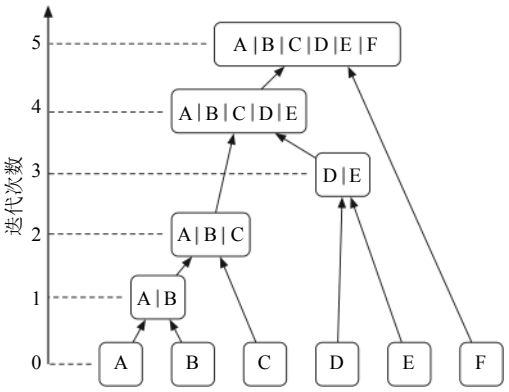

图 2 基于凝聚的自底向上层次聚类方法示意图

使用空间向量模型, 定义概念-文档矩阵, 领域本体概念用 W 表示, 特征项用 t 表示, 其中 t 使用 *tf-idf* 权值表示. 公式如下:

$$\overline{W_i}=(t_{i1},t_{i2},t_{i3},\cdots,t_{in}) \quad (2)$$

其中, $\overline{W_i}$表示第 i 个概念 t_{ij}:

$$t_{ij}=tf_{ij}\times\log\left(\frac{n}{n_i}\right) \quad (3)$$

tf_{ij} 表示抽取出的概念出现在文档集中的频率, n 表示数据集中文档总数, n_i 表示出现概念 i 的文档数. 按照

上述公式构建向量空间模型, 建立相似度矩阵步骤如下:

Step 1. 计算抽取出来的每个本体概念向量 $\overline{W_i}$, 构成本体概念向量空间.

Step 2. 计算两两概念之间的相似度:

$$sim(\overline{W_p},\overline{W_q})=\cos<\overline{W_p},\overline{W_q}>=\frac{\sum_{i=1}^{n}t_{pi}t_{qi}}{\sqrt{\sum_{i=1}^{n}t_{pi}^2\sum_{i=1}^{n}t_{qi}^2}} \tag{4}$$

Step 3. 构建概念相似度矩阵 S_{ij}, 其定义如下:

$$S_{ij}=\begin{array}{c}\\ \overline{W_1}\\ \overline{W_2}\\ \overline{W_3}\\ \cdots\\ \overline{W_n}\end{array}\begin{array}{c}\begin{array}{ccccc}\overline{W_1} & \overline{W_2} & \overline{W_3} & \cdots & \overline{W_n}\end{array}\\ \begin{pmatrix}1 & sim(\overline{W_1},\overline{W_2}) & sim(\overline{W_1},\overline{W_3}) & \cdots & sim(\overline{W_1},\overline{W_n})\\ & 1 & sim(\overline{W_2},\overline{W_3}) & \cdots & sim(\overline{W_2},\overline{W_n})\\ & & 1 & \cdots & sim(\overline{W_3},\overline{W_n})\\ & & & \cdots & \cdots\\ & & & & 1\end{pmatrix}\end{array} \tag{5}$$

簇间平均距离的定义如下:

$$d_{\text{avg}}(X,Y)=\frac{\sum_{x\in X,y\in Y}sim(x,y)}{|X||Y|} \tag{6}$$

其中, X, Y 表示两个簇, $|X|$和$|Y|$表示两个簇内元素的个数.

概念层次关系抽取步骤如下:

Step 1. 将抽取出的每个概念单独作为一簇.

Step 2. 计算两个簇之间的相似度即 $d_{\text{avg}}(X, Y)$.

Step 3. 取相似度最大的两簇进行合并, 若所有对象合并成一簇则跳转到 Step 4, 否则跳转至 Step 2.

Step 4. 结束.

在初始阶段, 将每个领域本体概念作为一簇, 根据相似度矩阵, 逐一将相似度大于规定阈值 threshold 的两簇合成一簇, 直到簇内平均距离小于给定阈值为止.

聚类的方法可以将本体概念分为多个簇, 但是簇内父概念和子概念的划分需要进一步定义, 使用簇内平均相似度来划分. 计算簇内概念两两之间的相似度, 若某一个概念的簇内平均相似度越大, 则说明此概念与其他概念联系广泛, 更有可能为簇内的父概念. 簇内平均相似度定义如下:

$$sim_{\text{avg}}(\overline{W_i})=\frac{\sum_{j=1}^{n}sim(\overline{W_i},\overline{W_j})}{|n|} \tag{7}$$

其中, $sim(\overline{W_i},\overline{W_j})$表示两个概念之间的相似度, n 为簇内概念的数量.

通过上述方法抽取的部分层次关系如图 3 所示.

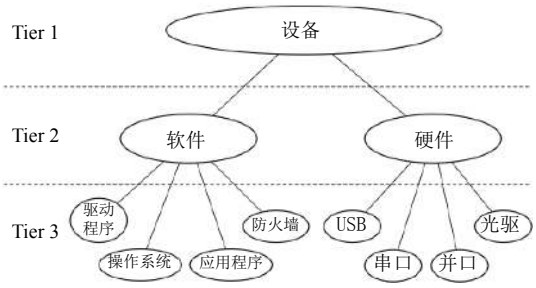

图 3 部分层次关系示意图

2.2.2 本体概念非层次关系抽取

本体概念之间的非层次关系主要包括: 部分与整体之间的关系、概念与属性之间的关系如对象属性和数据属性等. 本文基于统计学方法进行构建, 具有可移植性强, 对语言依赖性低等优点. 采用关联规则方法, 该方法可发现事物之间的相互依存性和关联性. 普通的关联规则方法只能得出概念之间确实存在非层次关系, 但无法得出具体的关系名称, 而概念之间的非层次关系可以用 (主语, 谓语, 宾语) 三元组表示, 所以用动词可以作为概念之间的非层次关系改进关联规则. 概念 W_i 和 W_j 之间具有关系 R_t 的关联强度可以用支持度和置信度来衡量. 支持度 *Support* 表示两个概念出现在同一个句子里的概率, 置信度 *Confidence* 表示在 W_i 出现的情况下 W_j 出现的概率, 定义如下:

$$Support(W_i\rightarrow W_j)=p(W_i\cup W_j) \tag{8}$$

$$Confidence(W_i\rightarrow W_j)=\frac{Support(W_i\rightarrow W_j)}{Support(W_i)} \tag{9}$$

使用以下改进的关联规则进行非层次关系抽取的步骤如下:

Step 1. 从抽取到的本体概念中选取概念 W_i 和 W_j.

Step 2. 根据上述公式计算$Support\,(W_i\rightarrow W_j)$和$Confidence(W_i\rightarrow W_j)$.

Step 3. 给定支持度和置信度阈值 *min_Support* 和 *min_Confidence*,如果 $Support(W_i\rightarrow W_j)$>*min_Support* 且 $Confidence(W_i\rightarrow W_j)$>$min\_Confidence$ 则概念 W_i 和 W_j 具有非层次关系, 进行 Step 4, 否则转到 Step 1.

Step 4. 统计出现在 W_i 和 W_j 中的所有动词及其共现频率. 如果概念与该动词的共现频率大于给定阈值, 则把该动词定义为概念之间的非层次关系.

Step 5. 验证所有动词之后结束.

以上方法抽取的部分非层次关系如表 1 所示.

表 1 部分非层次关系

对象属性	定义域	值域	数据属性	定义域	值域
madeBy	设备	生产厂商	IP_address	主机	string
hasGrade	告警	告警级别	file_path	文件	string
hasStatus	登录	登录状态	typeOfGrade	告警级别	string
isStatusOf	登录状态	登录	timeOfTrigger	安全事件	string
triggerBy	告警	安全事件	typeOfEquip	设备	string

2.3 依据概念关系构建本体

通过上述两种本体概念间关系的抽取, 完成概念间的分类关系、分层关系, 以及跨层次的归属关系等关系的罗列, 归纳得到本体构建所需的连接方式. 根据领域概念和概念间的关系, 可在 Protégé工具软件中构建树状的领域本体. Protégé是由斯坦福大学开发的本体开发工具, 该软件提供图形化界面可用于模拟概念类之间的关系以及属性. 本文对于层次关系的抽取结果可以在 Classes 选项卡定义,并且可以生成树状关系图, 如图 4 所示. 非层次关系抽取的结果主要包括对象属性和数据属性[12], 可以在 Protégé中的 object properties 选项卡和 data properties 选项卡中完成定义.

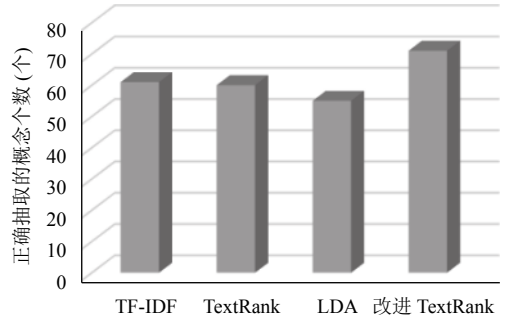

图 4 本体概念抽取对比实验结果

3 实验验证与分析

3.1 实验环境

基于上文中提出的自动化构建方法, 本文针对性地搭建了完整的实验环境以验证本文所提出方法的有效性. 数据方面, 采取了电力监控系统网络安全监测装置技术规范以及电力监控相关论文作为实验的文本数据源, 与之相配套的开发环境及使用到的工具软件列举如表 2 所示.

3.2 实验结果

本体的评价一般可以从两个角度来进行: 从应用的角度和从本体自身的角度. 基于应用的角度是比较是否使用本体对应用效果的影响, 依赖于具体的应用, 不够直观,所以本文采用基于本体自身的评价. 使用搜狗细胞词库中电力行业与计算机行业专业词汇表作为数据源, 手工构建本体作为参照本体, 其中包括 87 个概念类, 64 条数据属性以及 49 条对象属性. 为了提高实验评价的客观性在手工构建本体时使用《知网》(HowNet) 词汇相似度计算工具进行概念以及概念之间关系的建立, 并且增加适当的人工修正, 提高评价的可信度.

表 2 实验环境

工具	名称	版本号
编程语言	Python	v3.6.5
分词工具	Jieba	v0.39
相似度工具	Word2Vec	–
本体工具	Protégé	v5.5.0

(1) 本体概念抽取实验

在实验的本体概念抽取环节中, 本文基于相同的文本数据源开展了多种本体概念抽取方法的对比实验, 包括现有的 TF-IDF 算法、TextRank 算法、LDA 主题模型, 与本文提出的 TextRank 改进算法进行横向对比, 实验结果如图 4 所示.

通过实验对比可以看出,本文所提出的 TextRank 改进算法能够在相同的文本数据源中正确地抽取到更多的概念, 本体概念的抽取能力有显著的提升. LDA 主题模型在短文本数据上进行概念抽取的效果不佳, 而 TF-IDF 算法以及一般的 TextRank 算法本质上是依据词频, 当领域核心概念出现频次较低时, 容易产生遗漏, 效果一般.

(2) 概念间层次关系抽取实验

在层次关系抽取过程中, 采用了准确率 *Precision*、召回率 *Recall* 以及 *F*1 值等 3 种衡量指标来多角度地衡量关系抽取结果. 准确率为正确抽取出的关系与实际抽取出的关系总数的比值, 召回率为正确抽取出的关系与数据集中抽取出的所有关系总数的比值, *F*1 值为准确率与召回率的调和平均值. 上述 3 个衡量指标具体的计算方式如下:

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

$$Recall = \frac{TP}{TP+FN} \tag{11}$$

$$F1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \tag{12}$$

在层次关系抽取过程中, 选取不同的相似度阈值 threshold 对上述衡量指标的影响如表 3 所示.

表 3 Threshold 对实验结果的影响

相似度阈值	准确率	召回率	*F*1
0.90	0.68	0.59	0.63
0.80	0.74	0.69	0.71
0.70	0.67	0.60	0.63
0.60	0.58	0.53	0.55

将本文使用的层次聚类算法与文献 [1] 中使用的形式概念分析法 FCA, 以及一种基于 Beta 分布的聚类算法 BRT (Bayesian Rose Tree) 进行对比, 如图 5 所示.

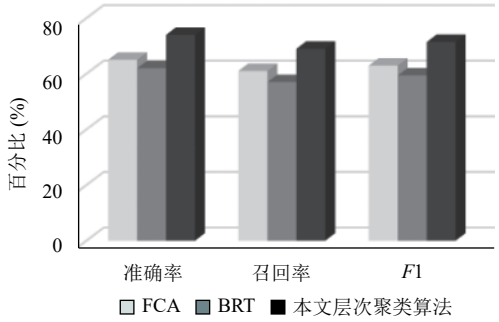

图 5 层次关系抽取对比实验结果

使用形式概念分析法得到的准确率、召回率和 *F* 值分别是 0.65、0.61 和 0.63; BRT 聚类算法的准确率、召回率和 *F* 值分别是 0.62、0.57、0.59; 本文采用的层次聚类算法的准确率、召回率和 *F* 值分别是 0.74、0.69 和 0.71. 可以看出本文使用的方法具有较好的抽取效果. 原因如下: 当句子中的概念存在并列关系时,层次聚类方法可以将这些概念归并到一个簇中, 有效防止簇内概念被分开, 而 BRT 算法需要计算概念之间的合并概率, 容易产生误差.

(3) 概念间非层次关系抽取实验

在非层次关系抽取中,使用式 (8) 和式 (9) 计算概念之间支持度和置信度,当支持度和置信度的阈值 *min_Support* 和 *min_Confidence* 取不同值时,对非层次关系结果的影响如表 4 所示,根据结果进行阈值选取.

使用词典的方法进行非关系抽取得到的非层次关系种类少, 而传统的关联规则方法无法得到非层次关系的名称, 所以上述方法无法进行实验对比. 本文采用基于模板的方法, 定义主语、谓语、宾语形式作为非层次关系的获取模板, 与本文提出的改进关联规则方法进行实验对比, 结果如图 6 所示.

表 4 不同支持度与置信度阈值情况下的准确度 (%)

支持度	置信度			
	0.01	0.05	0.10	0.15
0.01	57	59	62	53
0.05	63	66	57	49
0.10	53	57	50	44

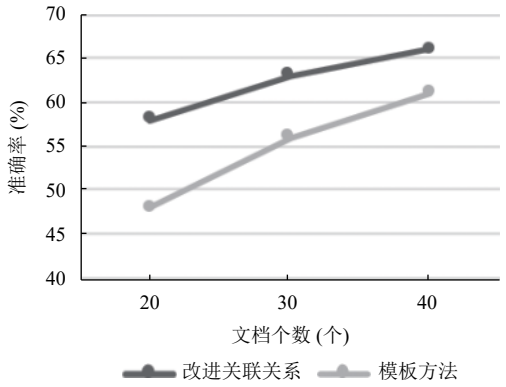

图 6 非层次关系抽取对比实验结果

可以看出随着使用的文档数量的增加, 两种方法的准确率均有所提高. 基于模板的方法由于筛选条件简单, 抽取到的非层次关系数量较多, 但是准确率较低; 本文提出的改进关联规则方法准确率较高.

通过统计结果可以看出, 本文所提出的领域本体构建方法准确率、召回率达到实际应用中的可用性要求, 能够为本体的自动化构建提供可靠的本体概念数据. 自动化抽取得到本体概念后, 依次进行了本体概念间层次关系、非层次关系的抽取. 最终, 依据概念、概念间的关系, 在 Protégé中构建了 SafeAgent 本体. 构建的本体 (部分) 如图 7 所示.

本文基于上述自动构建的电力监控安全本体开展进一步的实际应用, 开发一套电网网络安全智能监控系统软件. 该系统以 SafeAgent 本体作为后台的逻辑内核, 对电网监测设备采集的监测数据进行实时语义标注, 后续处理中依据数据语义特征实施不同操作. 在实际运行过程中, 对比于早期由开发人员手动构建的领域本体, 采用本文提出的方法进行自动化构建的本体具有可观的准确率、可靠性, 可以实现对人工构建本体的初步替代应用. 实验证明, 在确保替代不影响系统整体性能的前提下, 自动化构建本体方法可以切实有效地节省开发过程中的人力、物力, 并且在大规模、多领域的语义网建设中保持高度的可扩展特性.

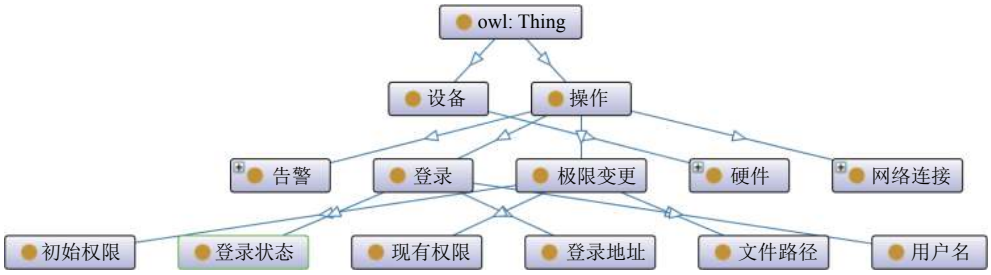

图 7 电力监控安全本体 (部分)

4 结束语

本文针对电力监控系统网络安全方面的实际需求, 开展了领域本体的自动化构建研究, 在现有的本体自动化构建方法基础之上, 针对文本数据到领域本体概念的转化、本体概念间层次关系的抽取、非层次关系的抽取等多个必要步骤进行了改进, 并初步实现了该领域本体的自动化构建过程. 经实验验证, 本文能够以较高的效率、准确率完成领域本体的自动化构建, 避免了耗费大量人力、物力的领域本体的人工构建过程, 从而实现对电力监控系统的网络安全行为进行快速的语义标注, 为未来的电力监控系统中的物联网设备标准化、智能化奠定了基础.

参考文献

1 Maedche A, Staab S. Ontology learning for the semantic web. IEEE Intelligent Systems, 2001, 16(2): 72–79. [doi: 10.1109/5254.920602]

2 廖莉莉, 沈国华, 黄志球, 等. 本体评估方法研究综述. 计算机应用研究, 2015, 32(3): 647–651. [doi: 10.3969/j.issn.1001-3695.2015.03.002]

3 任飞亮, 沈继坤, 孙宾宾, 等. 从文本中构建领域本体技术综述. 计算机学报, 2019, 42(3): 654–676.

4 Chen RC, Bau CT, Yeh CJ. Merging domain ontologies based on the wordnet system and fuzzy formal concept analysis techniques. Applied Soft Computing, 2011, 11(2): 1908–1923. [doi: 10.1016/j.asoc.2010.06.007]

5 李晓瑛, 李军莲, 冀玉静, 等. 基于叙词表及其语义关系的本体构建研究. 情报科学, 2018, 36(11): 83–87.

6 刘磊. 基于模板的 SSE-CMM 领域本体自动构建研究 [硕士学位论文]. 衡阳: 南华大学, 2011.

7 杜晶. 本体知识库的完全化过程研究 [硕士学位论文]. 北京: 北京交通大学, 2010.

8 王向前, 桂冬冬, 李慧宗. 面向文本的本体自动构建研究综述. 图书馆理论与实践, 2019, (4): 45–50.

9 郑姝雅, 黄奇, 张戈, 等. 面向用户生成内容的本体构建方法. 情报科学, 2019, 37(11): 43–47.

10 Shih CW, Chen MY, Chu HC, *et al*. Enhancement of domain ontology construction using a crystallizing approach. Expert Systems with Applications, 2011, 38(6): 7544–7557. [doi: 10.1016/j.eswa.2010.12.112]

11 Fang Q, Xu CS, Sang JT, *et al*. Folksonomy-based visual ontology construction and its applications. IEEE Transactions on Multimedia, 2016, 18(4): 702–713. [doi: 10.1109/TMM.2016.2527602]

12 Li JX, Ke G. Object-oriented method of constructing electric power system domain ontology based on Protégé. Advanced Materials Research, 2014, 997: 827–830. [doi: 10.4028/www.scientific.net/AMR.997.827]