情报学报 ISSN 1000-0135
第34卷第6期616-627，2015年6月
JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 34 No. 6, 616-627 June 2015

doi:10.3772/j.issn.1000-0135.2015.006.007

基于形式概念分析的学科术语层次关系构建研究[1)]

王 昊 朱 惠 邓三鸿

（1. 南京大学信息管理学院，南京 210023；
2. 南京大学 江苏省数据工程与知识服务重点实验室，南京 210023）

摘要 本体是领域知识的有效组织和描述，本体学习则是实现本体自动构建的方法体系和技术集合。本文以本体学习理论为指导，提出了一种以文档-术语空间为核心、形式概念分析（FCA）为手段的中文领域本体层次结构自动构建的有效方法，并以“白血病”领域为例，对面向学科资源的医学专业术语层次关联的抽取进行了详细论证，具体包括专业术语的抽取和筛选，术语文档关联的修正等数据清洗过程；文档术语矩阵的建立，领域概念格的自动生成，以及概念格中术语属性的层次关联建立等FCA过程；术语层次关联的自动OWL描述和存储，和领域本体的概念检索和可视化展示过程等。

关键词 学科术语 层次关系 本体学习 形式概念分析 概念格 OWL

Study on Construction of Hierarchy Relationship of Subject Terms Based on Formal Concept Analysis

WangHao, Zhu Hui and Deng Sanhong

(1. *School of Information Management of Nanjing University, Nanjing* 210023;
2. *Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing* 210023)

Abstract Ontology is the effective organization and description for domain knowledge and Ontology Learning (OL) is the methodology and technology to construct Ontology automatically. With the OL theory as a guide, this paper proposes an effective method, which is with documents-terms space as a core and with Formal Concept Analysis (FCA) as a means, to construct hierarchy structure of Chinese Domain Ontology automatically. Taking "leukemia" field for an example, it in detail demonstrates the extracting process on hierarchy relationship of Medical professional terms oriented disciplines resource, which specifically contains 3 processes. First is the data clearing process as initialization including extracting and filtering of professional terms, and amendment of association of terms from documents. Second is the FCA process including building of documents-terms matrix, automatic generation of domain concept lattice and construction of hierarchy relationship of properties from terms in concept lattice. Third is the terms ontology description process including automatic OWL description and storage of hierarchy associations of terms, concept searching and visually displaying of domain ontology.

Keywords subject terms, hierarchy relationship, ontology learning, formal concept analysis (FCA), concept lattice, OWL

收稿日期：2014年12月14日。

作者简介：**王昊**，男，1981年生，情报学博士，南京大学信息管理学院副教授，主要研究方向：从事智能信息处理与检索、知识本体构建及应用、科学评价和引文分析等，E-mail：ywhaowang@nju.edu.cn；**朱惠**，女，1978年生，南京大学信息管理学院讲师，主要研究方向：从事知识本体构建及应用，E-mail：zhuhui@nju.edu.cn；**邓三鸿**，男，1975年生，南京大学信息管理学院副教授，主要研究方向：从事知识图谱构建、学科评价等，E-mail：sanhong@nju.edu.cn。

1） 本文受国家社科重大招标项目“面向学科领域的网络信息资源深度聚合与服务研究”（12&ZD221）、江苏省自然科学基金项目“面向专利预警的中文本体学习研究”（BK20130587）等的资助

1 引 言

本体是对领域知识的有效组织和描述，也是WWW向语义网转化的核心要素[1,2]。因此，构建本体，即本体工程(Ontology Engineering)，对于知识的传承和扩散具有重要的意义。然而，本体工程是一项极其繁杂的工作，需要耗费大量的时间和财力，这与语义网建立所需要的“知识本体快速开发”形成了尖锐的矛盾。研究人员逐渐感觉到了半自动或自动生成领域本体的必要性，于是本体学习技术由此产生。

本体学习(Ontology Learning，OL)是指利用语言分析、机器学习和数学统计算法等技术，通过计算机自动或半自动地从已有的数据资源中发现潜在的概念、概念间的关系和公理等本体元素的方法体系和具体过程[3]。其实质是信息抽取在知识层面上的进一步延伸。结合信息抽取的5层次理论[4]，OL根据抽取任务的不同，可以分解为术语(Terms)、同义词(Synonyms)、概念(Concept)、概念层次(Concepts' Hierarchies)、语义关系(Semantic Relations)以及公理/规则(Axioms/Rules)等6个层次。在OL的层次体系中，自下而上其知识的复杂度逐渐上升，知识抽取特别是从非结构化文本中抽取的难度也逐层提高。对于英语等字符语言，本体学习的研究和应用已经覆盖全部层次，但在公理和约束规则的抽取上仍然涉及较少[5,6]；而中文等象形语言，由于其语法特征的复杂性和规则的多样性，目前研究主要聚焦在第4层次，很少涉及非层次语义关系特别是公理的抽取研究，而且层次关系的抽取也多停留于实验论证阶段，基本上还没有可进行实际应用的工具或算法出现。

由于学科的特殊性，医学通常是知识库研究和应用的重点领域。早在本体概念被引入知识工程领域之前，人工设置规则以演绎推理方式实现诊疗专家系统[7]就已经是研究热点，人们试图通过自动化方式实现从症状、疾病到诊治、处方的完整过程，以避免人为操作可能产生的差错和专家离世导致的经验流失。随着本体概念的兴起，构建医学本体[8]作为专家系统的知识基础成为了研究人员的共识，MeSH[9]就是其中的典型代表。本文试图将在英语文本上具有较广泛应用的层次关系抽取方法引入到中文医学文本中，从实用角度出发，在没有任何外部知识库的支持下实现中文文本到医学学科术语层次结构的衍化，同时针对中文文本的特征，探索改进层次关系抽取方法的策略，从而形成一整套针对学科资源抽取领域术语分类体系的正确而有效的解决方案。本文研究一方面是对中文术语层次关系自动构建方法的验证和应用，另一方面则是从专业学术著作中挖掘可能的术语分类体系作为MeSH的合理补充。

2 近期相关研究

概念层次关系是最重要的本体元素，有研究甚至直接将本体描述为具有包含关系的概念之间的一种层次结构[10,11]。这样的阐述虽然并不非常完整，但是在实际应用中，特别是学科领域，概念分类体系的揭示也基本能够满足需求。概念层次关系包含两种类型，对称关系和非对称关系。对于前者，由于查询扩展、词表构建等的需要，研究相对比较成熟[12,13]；后者则是目前本体学习研究的重点，不同领域的学者针对来源数据的特征及知识层次的具体应用提出了各种构建方法。

依据语言学规律构建术语层次是早期常用的一种方法。模式匹配，或称为Hearst模式(Hearst-Pattern)[14~16]，是其中的典型代表。所谓模式就是能够指示上位/下位关系的关键短语，例如“such as”，“and other”，“especially”等。对含有诸如此类短语的句子进行语法分解可以识别相关术语的层次关系。通过术语筛选、术语相似度和术语分散度等3个依赖维度从对称关系衍化生成非对称关系也是术语层次构建的一种思路[17]。这种方法认为术语的相关文档频率(Relevant Document Frequency，RDF)值越大，说明其比同义术语所处的层次越高。借助词汇层次(Lexical Hierarchies)和包含层次(Subsumption Hierarchies)的建立算法也可生成术语层次，前者从词法组成角度出发，根据短语或合成词中术语的结合规则来构建术语层次[18]，而后者则利用“如何包含术语B的文档集合是包含术语A的文档集合的一个子集，那么术语B是术语A的下位类”的理论假设[19]，基于相关文档集合之间的包含关系来建立术语层次，实际上是形式概念分析在信息检索领域的一种应用描述。

事实上，上述术语层次构建方法是以优化信息检索为目的从词语角度提出的，OL的概念在此时还没有被完整地阐述。随着快速构建本体以推广语义网应用需求的提出，Maedche等[20]对OL的过程和

方法进行了详细描述，至此术语层次结构被提升至概念层次，涌现出了各种系统性和理论性更强的实现算法，笔者将其概括为4个方向：①最为典型的就是Maedche等在本体学习方法论中提出的层次聚类分析（Hierarchical Clustering Analysis）方法，即将术语通过聚类形成若干个具有层次关系的主题簇，然后用一个简洁的说明（即概念）对每个主题进行概括[21~23]；②潜在语义索引（Latent Semantic Indexing，LSI）方法则是通过奇异值分解（Singular Value Decomposition，SVD）获取原始矩阵的所有特征（奇异值），进而通过主成分分析（Principal Component Analysis，PCA）选择主要特征以获得近视矩阵，根据近视矩阵的有效性评价要推导向量之间的上位/下位关系[24~26]；③形式概念分析（Formal Concept Analysis，FCA）认为下层概念的属性集合要包含上层概念的属性集合，由此根据概念特征的关系来推导概念之间的上下位关联[27,28]，这是近年来最常用的本体层次关系识别技术，与模糊数学结合而形成的模糊FCA（Fuzzy FCA，FFCA）方法更是将上下位关系拓展成为一种概率结构，致使模糊本体（Fuzzy Ontology）的产生[29]；④基于知识库的方法，即利用现有的知识库（如Wikipedia、WordNet等）资源，对需要分层的术语进行定义，然后通过解析定义之间的语义关系来建立术语之间的层次结构[30~32]，其实质是通过将一个“短”术语扩展为一个“长”而“准”的定义来增加术语的属性描述，从而更好的揭示术语的特征，增大上下位术语被准确识别的概率，该方法的识别效果较好，但是对知识库的领域依赖性较强。上述各种本体概念/术语层次构建方法存在一定的相似性，即均需要建立“术语×属性”矩阵，然后借助属性之间的关系计算来判断术语间关联。

3 采用方法

本文试图将FCA的理念应用于医学学科领域，借助期刊论文文档建立中文医学专业术语之间的层次关系，并将其作为本体的组成部分进行OWL描述和可视化展示。

3.1 基于FCA的学科术语层次关系构建模型

笔者对中文医学专业术语层次关系的构建过程和实现方法进行了系统建模，如图1所示。模型的基本目标是在没有其他知识库的支持下，从文本文档中衍生出学科专业术语的分类体系；以学科期刊论文文档作为数据输入，术语层次关系的OWL文件及可视化视图作为输出。模型的基本思路被划分为3个部分：①学科数据清洗过程。首先从学科论文文档中抽取关键词作为专业术语的候选集合，并根据文档集合中关键词的出现频率筛选术语，以明确专业术语集合，然后对术语在文档中的出现概率进行修正，补充术语文档关联，建立相对完整的<术语，属性，权重>3元组。②形式概念分析过程。将3元组转化为文档×术语矩阵，形成学科领域的形式化背景，然后利用现有的概念格生成算法，将形式

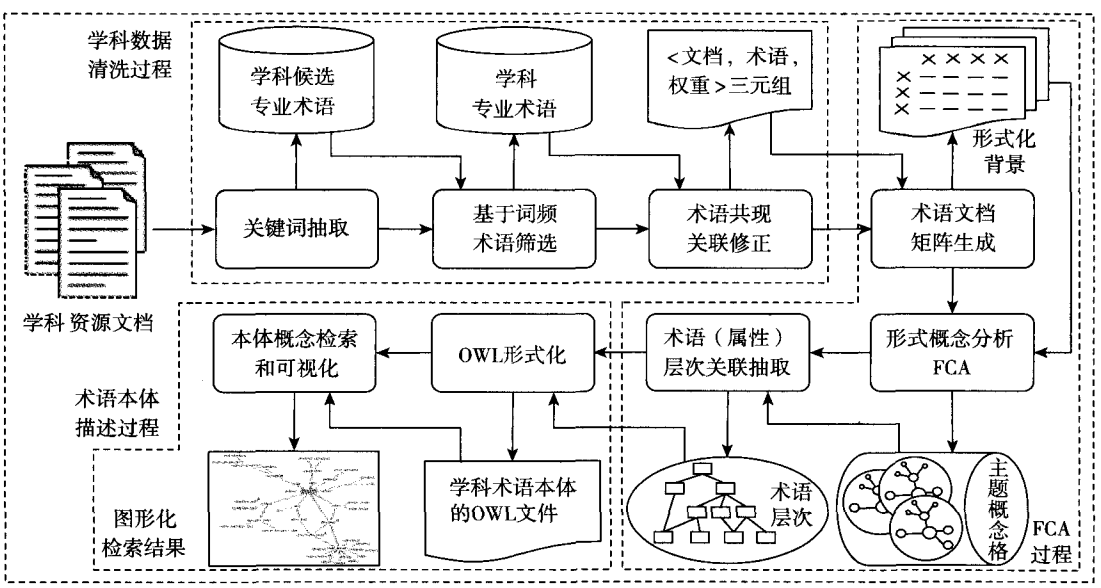

图1 基于FCA的学科专业术语层次体系构建模型

化背景转化为领域主题概念格，并根据主题概念间的直接和间接继承关系抽取主题属性（即术语）之间的层次关联。③术语本体的描述过程。将术语之间的层次关联用国际标准化语言 OWL 进行形式化描述，以文本形式存储以备进一步使用，包括以图形可视化方式展示术语概念的层次体系等。

3.2 数据预处理方法

本文从 2009～2012 年万方数据收录的医学专业期刊论文中检索出题名中含有“白血病”的 6194 条题录数据（不含摘要和全文），作为“白血病”领域的文档集合，并直接以论文中提供的关键词作为候选专业术语。然而，这些关键词主要由作者给出而没有经过规范化处理，而作者由于用词习惯、文化程度、专业素养等各方面均存在差异，因此给出的关键词特别是中文关键词存在较大的随意性、不一致性以及误差性。为此，有必要对关键词进行预处理，具体包括两个工作，一是对关键词转化而来的候选术语做进一步筛选，以获得具有较高领域认可度的专业术语集合；二是对术语与文档之间的共现关联进行修正，以弥补术语来源单一造成的术语文档间语义关联的缺失。

领域术语一般来说是专业文档内比较重要的词汇。英文术语抽取的传统做法是：对文档分词，去除停用词后计算术语对文档的 TF-IDF 值作为术语在文档中的权重，继而根据权重筛选出满足阈值的术语作为专业术语或候选。然而在中文文档集中，上述方法却基本上无法操作，一是中文专业术语通常是篇幅较长的短语，无法通过简单分词操作获取；另一方面在篇幅较短的题名和关键词文本中，TF-IDF 的作用无法发挥。为此，笔者采用术语在整个文档集合中的出现频率 N_k 作为筛选条件，即若：

$$Nk > C \quad (1)$$

则认为该术语被领域普遍认可，可作为该领域的专业术语。其中 C 为词频阈值，可根据筛选出来的术语集合的文档覆盖情况进行确定，以保证术语的领域认可度。

从关键词集合中筛选出专业术语，那么术语与文档之间的关联在来源数据中即已存在。但是由于论文中存在的关键词数量较少，即术语与文档关系不充分，直接以这些关联来构建形式背景将无法充分表达术语之间的包含关联。因此，笔者对术语与文档之间的关联进行了修正，基本思想是：对所有专业术语进行检测，若术语在论文的题名或关键词文本中存在，说明该术语与论文之间存在共现关联，若这种关联在原有数据集合中不存在，则添加。例如，某文档题名为“伊马替尼对慢性粒细胞白血病患者生育及生殖的影响”，原文中含有 4 个关键词“伊马替尼”、“慢性粒细胞白血病”、“生育”和“生殖”。根据上述方法，对术语“白血病”进行检测时发现该字符串在上文的题名和关键词文本中均有出现，且该术语原与文献没有关联，那么添加“白血病”术语与该论文的关联。通过如此计算，可对术语之间的共现关联进行补充修正，在一定程度上强化了术语之间的语义关系。

3.3 基于 FCA 的术语层次体系构建方法

FCA 是 Wille 在 1982 年提出的一种数学理论[33]，是一种用于数据分析、知识表示、信息管理以及本体生成的重要方法[34,35]。其用对象和属性间的二元关系来表达领域中的形式化背景，并从中抽取包括内涵和外延在内的概念层次结构，即概念格（concept lattice）[36]。

定义 1：形式化背景（formal context）是一个三元组 $F=(O, M, R)$，包含 3 个集合，其中 O 是对象的集合，M 是属性的集合，R 是 O 和 M 之间的一个二元关系集合，即 $R \subseteq O \times M$。oRm 表示 $o \in O$ 与 $m \in M$ 之间存在关系 R，读作“对象 o 具有属性 m”。

形式化背景实际上就是对象 × 属性矩阵。于是，在信息检索中被广泛应用的文档 × 术语矩阵可以映射为形式化背景 $F=(D, T, I)$[37]，其中 D 表示文档集合，T 表示术语集合，I 则是文档与术语之间的共现关联，可以用术语在文档中是否存在或存在频次描述。表 1 为形式化背景的示例，列出了“白血病”领域部分文档与术语之间的假设关联，表中√表示文档与术语间存在关联，构成了集合 I，且 $D=\{D1, D2, D3, D4, D5, D6, D7, D8\}$，$T=\{$白血病，髓系白血病，化疗，细胞，治疗，淋巴细胞，干细胞移植$\}$。

定义 2：在一个形式化背景 $F=(O, M, R)$ 中，可以定义两个映射 f 和 g：$\forall O_x \subseteq O: f(O_x)=\{m \in M \mid \forall o \in O_x, oRm\}$：对象集合 O_x 中所有对象的共同属性集合；$\forall M_y \subseteq M: g(M_y)=\{o \in O \mid \forall m \in M_y, oRm\}$：具有相同属性集合 M_y 的所有对象集合。如果 $f(O_x)=M_y$ 且 $g(M_y)=O_x$，则称 $c=(O_x, M_y)$ 为概念，其中 O_x, M_y 分别称作概念 c 的外延（extent）和内涵（intent）。F 中所有概念 c 的集合用 C 表示。

表1 “白血病”领域文档与术语的形式化背景 $F=(D, T, I)$ 示例

D \ T	白血病	髓系白血病	化疗	细胞	治疗	淋巴细胞	骨髓移植
D1	√	√	-	-	√	-	-
D2	√	√	√	-	√	-	-
D3	√	-	√	-	√	-	-
D4	√	-	-	-	√	-	√
D5	√	-	-	√	-	√	-
D6	√	√	-	√	-	-	-
D7	√	-	-	√	√	-	√
D8	√	-	-	√	√	√	-

在 $F=(D, T, I)$ 中，设 $X \subseteq D$，$Y \subseteq T$，根据定义2可得：

$\sigma(X)=\{t \in T \mid \forall d \in X:(d,t) \in I\}$，文档集合 X 包含的公共术语集合

$\tau(Y)=\{d \in D \mid \forall t \in Y:(d,t) \in I\}$，术语集合 Y 所在的公共文档集合

若 $X=\tau(Y)$ 且 $Y=\sigma(X)$，那么 $c=(X, Y)$ 被称为主题概念。例如表1中 $c_1=(\{D1, D2, D6\}, \{$髓系白血病，白血病$\})$ 可称为主题概念，其外延为 $\{D1, D2, D6\}$，内涵为{髓系白血病，白血病}，这个主题描述的是“髓系白血病”和“白血病”的相关内容，文档 $D1$、$D2$ 和 $D6$ 均是对这个主题的研究。即若某个术语集合中的每个术语均出现在了文档集合中的每个文档中，那么这个公共的术语集合和文档集合一起形成了一个主题概念，文档集合被称为这个主题的外延，而所有术语一起形成了其内涵。

定义3：如果 $c_1(O_1, M_1)$，$c_2(O_2, M_2)$ 都是形式化背景 F 中的概念，并且 $M_2 \subseteq M_1$，那么 c_1 被称作 c_2 的子概念（sub-concept），c_2 则是 c_1 的超概念（super-concept），记为 $c_1 \leqslant c_2$。“$\leqslant$”称为序，反映了概念间的层次关系。由序所描述的 F 的所有概念及其层次关系记作 $C(F, \leqslant)$[38]，称为概念格（concept lattice）[39]。

定义3表明了上层概念的属性集合应该包含于下层概念的属性集合，即特征越多，概念级别反而越低；映射到文档术语环境中，可认为文档包含的术语越多，说明其阐述的主题（术语的交叉部分）就越专业，应该处于主题概念的下层，反之亦然。例如在表1中，令 $c_1=(\{D2, D3\}, \{$化疗，治疗，白血病$\})$，$c_2=(\{D1, D2, D3, D4, D7, D8\}, \{$治疗，白血病$\})$，不难发现 $M_2\{$治疗，白血病$\} \subseteq M_1\{$化疗，治疗，白血病$\}$，因此 c_2 是 c_1 的父概念；c_2 主题为白血病的治疗，而 c_1 专门探讨白血病治疗中的化疗，可见就主题而言前者研究范围更大，是后者的上位概念，显然探讨上位概念的文档（外延）相对更多。

根据FCA的上述定义，笔者计算出了表1所示形式化背景的概念格 $C(F, \leqslant)$，用Hasse图表示，如图2所示。①图中圆形结点表示主题概念 c，圆形大小表示主题外延的个数；在层次结构中，上层为父概念，下层表示子概念；最上层概念的属性集合为所有对象均具有，因此其对应的外延最多，相反最下层概念含所有属性，其对应的外延也最少；自上而下，概念层次降低，属性集合逐渐增大，而对应的外延数量将会越来越少。在本例中，“白血病”出现在了所有文档中，为所有对象均具有的属性，而具有所有属性的文档对象为空。②每个概念由两部分构成，上半部分代表属性（内涵）M，下半部分代表对象（外延）O。为了简化概念格，每个结点仅显示了相对其父结点新增的属性和相对其子结点新增的对象。因此，若属性半圆呈蓝色表示有新增属性分布于该结点上，对象半圆呈黑色表示有新增对象分布在该结点上，而每个概念结点的属性集合和对象集合分别为以该结点为根节点的上子树上所有属性的总和（继承其父类的所有属性）和下子树上所有对象的总和（涵盖了其所有子类的外延）。例如图中间最右侧“$D5$，淋巴细胞”结点，其属性集合为{淋巴细胞，细胞，白血病}，对象集合为$\{D5, D8\}$，形成一个完整的主题概念 $c\{\{D5, D8\}, \{$淋巴细胞，细胞，白血病$\}\}$。③重点考察图中上半部分，发现作为概念属性的术语之间似乎存在一定关系。概念格

描述的是概念之间上下位关联，即下层概念通过新增属性从上层概念中派生出来，根据 Hasse 图的示意，新增属性所在的对象等于以其为根节点的下子树的外延总和，映射到本例中就是，新增术语所在的文档为以其为根节点的下子树的所有文档总和，包括了下位概念新增属性的所在文档集合。例如，图中“*D*4，骨髓移植”结点，新增“骨髓移植”属性，该属性出现在 *D*4 和 *D*7 中，该结点的父节点，新增了“治疗”属性，其出现在文档 *D*4 和 *D*7 以及 *D*3/*D*2/*D*1/*D*8 中，即父节点新增属性所在文档集合必然包括了子结点新增属性所在文档集合。那么，根据当且仅当包含术语 *A* 的文档集合是包含术语 *B* 文档集合的超集时，术语 *A* 包含术语 *B*[40]，即术语“治疗”是“骨髓移植”的上位术语。如果一个术语出现的文档比较多或其分散度较大，那么其泛化程度比较高。同理可得，“白血病”是“细胞”、“治疗”和“髓系白血病”的上位术语，“细胞”是“淋巴细胞”的上位术语，“治疗”是“化疗”和“骨髓移植”的上位术语。于是，用于构建概念之间层次关系的 FCA 应用于文档术语环境中，不仅可以生成主题概念之间的层次结构，同时根据“概念格中子概念新增的属性是父概念新增属性的下位类，且新增的属性均为首次出现”的结论，可以构建术语之间的层次关联。

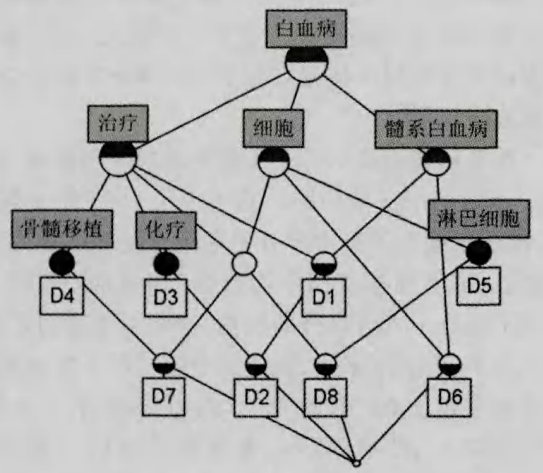

图 2　表 1 所示形式化背景的概念格

综上所述，基于 FCA 实现术语层次关系构建可分为 3 个步骤：①建立领域形式化背景，即构建文档 × 术语矩阵；②将形式化背景转化为概念格，即生成主题概念之间的层次结构；③根据概念格中术语属性首次出现的概念之间的上下位（包括传递）关系推导术语属性之间的层次语义关联。

3.4　术语层次体系的 OWL 描述及可视化展示方法

OWL[41]（Ontology Web Languages）是由国际标准化组织 W3C 发布的本体的标准描述语言。它基于描述逻辑，可用以明确表达领域概念（类）的含义以及概念间的语义关系，从而实现计算机对其所标注信息资源的“理解”。本文构建的领域术语分类体系实际上是学科领域本体的一个子集，可以采用 OWL 进行形式化描述，为领域知识的进一步应用奠定基础。本文将每个术语视为一个类，而通过 FCA 获得的同义术语，即具有完全相同文档集合的术语，则将其用分隔符相连作为一个特殊术语；术语之间的层次关联则被视为类间的上下位关系。OWL 中用于描述类上下位关系的标签是 owl:Class 和 rdfs:subClassOf。前者用于定义一个类，其语法如公式（2）所示，其中 rdf:ID 指明了该类的标识，通常为类名，“内容”则是类的组成，包括其他子标签，例如 rdfs:label 为类提供一个自然语言名称，类似于注释；后者用于指明当前类的父类，有两种使用方式：①如公式（3）所示，指明一个已定义好的父类，rdf:resource = "#父类名称" 用于引用已经定义的类，②如公式（4）所示，若父类之前没有定义，那么在指明父类的同时定义父类。图 3 所示的 OWL 编码显示了“白血病”为“细胞”的父类。仅描述类上下位关系的 OWL 编码较为简单，在类数量较大时，可通过程序自动生成类定义及其层次结构的 OWL 编码。

```
<owl:Class rdf:ID = "名称" > 内容 </owl:Class >                    (2)
```

```
<owl:Class rdf:ID = "子类名称" >
    <rdfs:subClassOfrdf:resource = "#父类名称"/
> …                                                               (3)
</owl:Class >
```

```
<owl:Class rdf:ID = "子类名称" >
    <rdfs:subClassOf > <owl:Class rdf:ID = "父类
名称" > 内容 </owl:Class >                                        (4)
    </rdfs:subClassOf > …
</owl:Class >
```

Protégé[42]是一款基于插件的本体编辑和展示工具，能够对 OWL 文件进行读写，实现 OWL 文件到可视化图形的转换。目前已经出现了多种用于 Protégé 的本体可视化插件，其中 Ontograf[43] 的可视化功能非常齐全，能够以多种布局方式展示本体关系并进行有效过滤，支持本体概念检索，能够对概念进行快速定位和局部展示。

```
<owl:Class rdf:ID="白血病"></owl:Class>
<owl:Class rdf:ID="细胞">
    <rdfs:subClassOf rdf:resource="#白血病"/>
</owl:Class>
```

```
<owl:Class rdf:ID=" 细胞">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="白血病"></owl:Class>
  <rdfs:subClassOf>
</owl:Class>
```

图3 “白血病”和“细胞”术语上下位关系的OWL编码

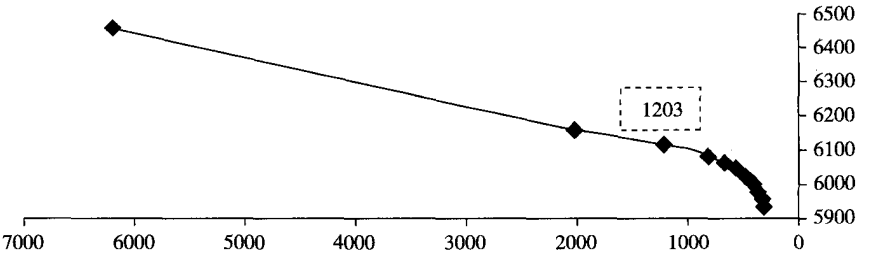

图4 文档量和术语量随词频阈值 C 的变化趋势

4 结果分析

根据上述方法,可对本文的数据集进行处理,以获得“白血病”领域专业术语的分类体系,并对其进行形式化描述和可视化展示。

4.1 学科数据清洗结果分析

在6194篇关于“白血病”的专业论文中,共可抽取6454个关键词作为候选专业术语。令词频阈值 C 从1变化到10,可得论文数量随术语数量变化的趋势如图4所示。图中横轴为术语量,纵轴为论文量。不难发现,随着 C 值的增长,术语量减少,其所能覆盖的文档量也逐渐减少;而且文档量减少的速度随术语量的减少不断增大。可见当 C 值越大时,每过滤掉一个术语都可能造成相当数量文档的减少,可见此时术语的文档覆盖面较大,其领域的认可度较高。笔者选择 $C=2$,使得术语量介于1000~2000,以此作为“白血病”领域的专业术语集合;由此获得的上下位术语之间至少存在3个文档以上的重合,以保证术语之间的层次关系具有一定可信度。

经过词频筛选,共可获得1203个专业术语,分布于6116篇论文文档中,术语与文档之间共存在19 943个关联。为了强化术语与文档之间的关系,笔者根据前文介绍的思想对术语在文档中的出现情况逐一进行了修正,修正后共可获得70 436个共现关联,以<文档,术语,权重>3元组形式存储,这些语义关联将是构建术语间层次关系的主要依据。

4.2 基于FCA的术语层次体系生成结果分析

将文档术语关联3元组转化为文档×术语矩阵,形成了医学领域的形式化背景 $FM=\{D, T, I\}$,D 中共有6116个对象,T 中有1203个属性,I 中有70 436个关联。对FM进行FCA,共可生成主题概念81 323个,并根据属性首次出现的概念之间的上下位及其传递关系可推导出术语属性间的直接上下位关系共1688对。由于概念数量巨大,生成的概念格Hasse图非常复杂,本文只能从横向和纵向两个方面对局部术语间层次关联进行可视化分析,而所有术语之间的层次体系在保存为OWL文件后以类检索方式展示。

图5从横向显示了在文档中出现频率最高的前25个术语的层次结构图。共生成了2595个主题概念,图中仅截取了其中部分主题概念及其层次关系。很明显,由于选择文档中均包含“白血病”术语,因此该术语处于术语集合的顶端,表明了领域的方向,其下有14个下位术语,包括图中第二层上所有术语以及第三层上的“护理”和“白血病患者”;此外,①“细胞”>“粒细胞”>“粒细胞白血病”(图6),②“细胞”>“淋巴细胞”>“淋巴细胞白血病”,③“细胞”>“淋巴细胞”>“急性淋巴细胞”>“急性淋巴细胞白血病”,④“细胞”>“白血病细胞”,⑤“急性”>“急性白血病”,⑥“急性”>“急性髓系白血病”,⑦“髓系”>“髓系白血病”>“急性髓系白血病”等均形成了上下位关联。

图7从纵向显示了以“白血病”的下位术语“髓系”为根节点的医学主题概念格,共涉及687个文

档,8 个术语,生成了 10 个主题概念。由于生成的概念较少,图中清晰展示了术语之间的层次结构以及概念的外延数量。该子树涉及的 8 个术语一共形成了 5 个分支 7 对上下位关联,①“髓系”>“髓系抗原”,②“髓系”>“髓系白血病”>“慢性髓系白血病”,③“髓系”>“髓系白血病”>“急性髓系白血病”>“EVAGREEN 染料”,④“髓系”>“髓系白血病”>“急性髓系白血病”>“毛细管电泳”,⑤“髓系”>“髓系白血病”>“急性髓系白血病”>“老年急性髓系白血病”。每个结点下方显示的为该主题概念包含的文档外延数量及其比例,可见在“髓系”研究领域内,“急性髓系白血病”是当前的研究热点,近 82% 的文献均涉及了该主题的研究;而在该方向上,“老年急性髓系白血病”是研究的重点内容。“慢性髓系白血病”和“急性髓系白血病”结点附近的方框显示了主题概念的基本内容,列出了对应主题的内涵、外延总量以及结点自身所包含的外延数量,其中“慢性髓系白血病”结点有 56 篇文档涉及,只有 1 篇是其子概念的外延,而且该文档同时介绍了“慢性髓系白血病”和“急性髓系白血病”;右侧“急性髓系白血病”结点的主题概念有 63 篇文档是来自其子类的,其中最多的为“老年急性髓系白血病”,而“EVAGREEN 染料”和“毛细管电泳”极有可能是与“急性髓系白血病”相关的材料、方法或技术。类似的,也可以分别列出以“白血病”所有一级下位术语为根节点的子概念格进行分析。

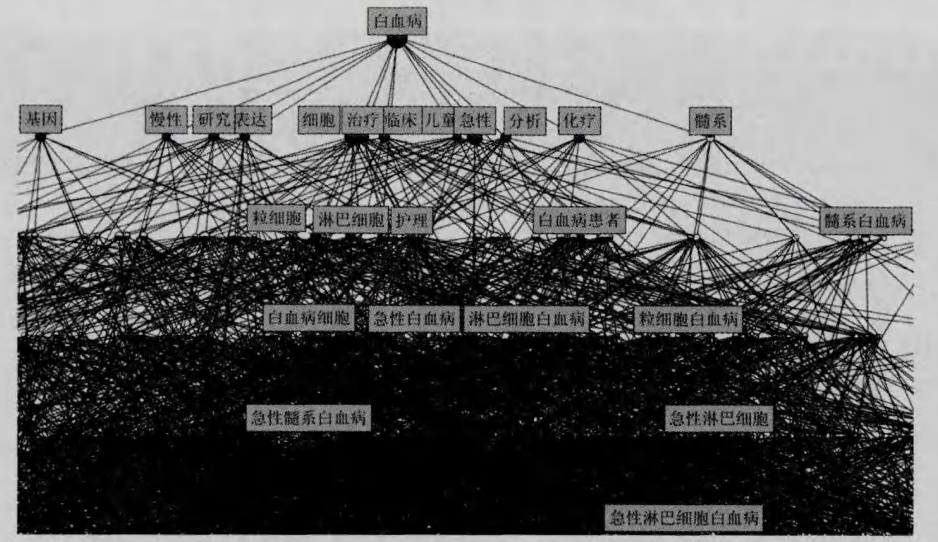

图 5　文档中出现频率最高的 25 个术语的局部层次结构图

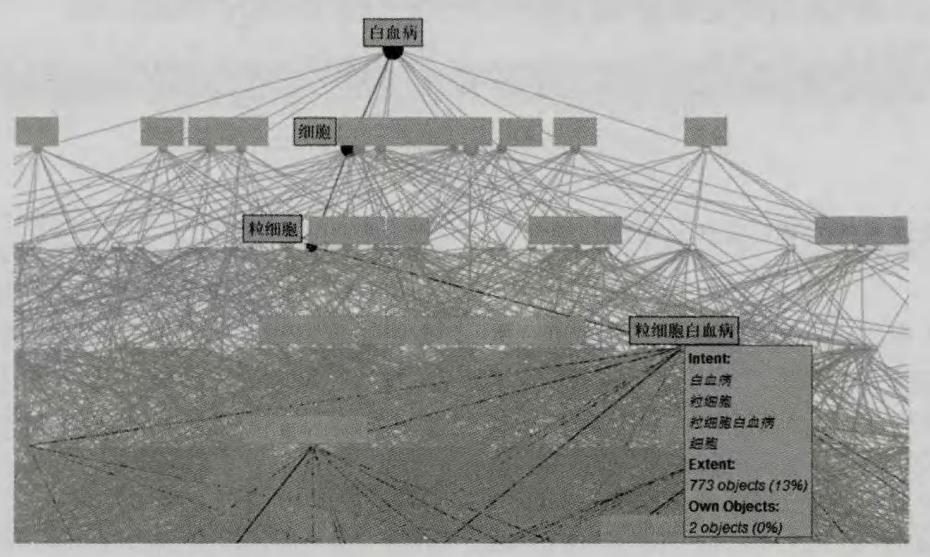

图 6“白血病”>“细胞”>“粒细胞”>“粒细胞白血病”间层次关系的高亮显示

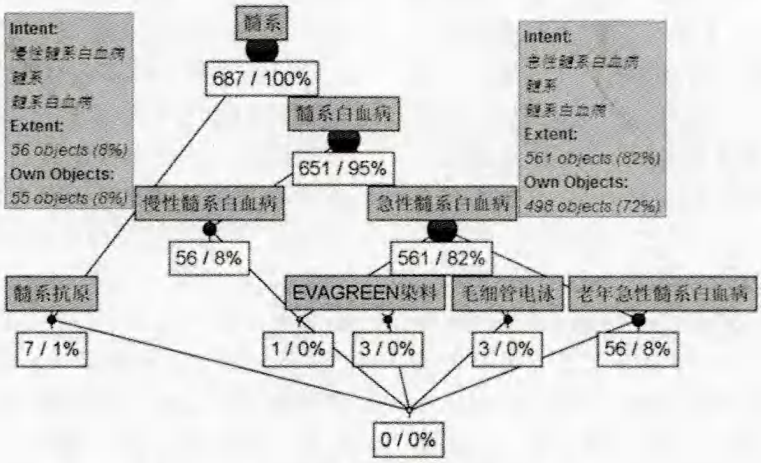

图7 以“髓系”术语为根节点的医学主题概念格

图8 医学“白血病”领域专业术语本体的OWL文件

4.3 术语层次体系的OWL描述及可视化展示结果分析

根据OWL中类和子类定义的基本语法和采用标签,笔者对本文生成的“白血病”医学领域1203个术语间的1688对上下位关联进行了自动OWL编码(图8),生成了仅含层次关系的医学“白血病”领域本体,可用于诸如语义检索、蕴含推理等实际应用。笔者用Protégé对其进行读取,并采用OntoGraf插件对其进行了类检索和可视化展示。图9展示了“造血干细胞”类的层次结构。图中左侧以树形结构显示了“白血病”领域本体的所有类及其相互包含关系,可实现本体概念的顺序浏览;右侧上方有search框,可实现本体类的随机定位;在此输入“造血干细胞”,即可在右侧下方以树状图方式显示出了与该类相关的所有概念,例如其上位类有“造血干”和“干细胞”,下位类包括“造血干细胞移植”等,连线中箭头指向了子类方向,通过绘图区上方的工具栏可对图形进行调整,例如将树形图转变为星形图或spring图。图10采用spring图对“干细胞”类的层次结构进行了展示。

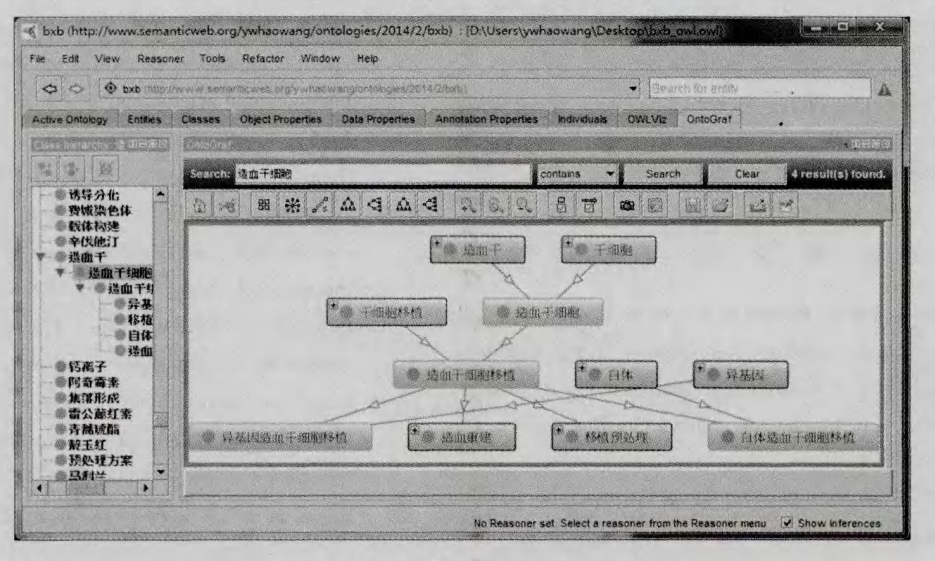

图 9 “造血干细胞”类层次结构的树形图展示

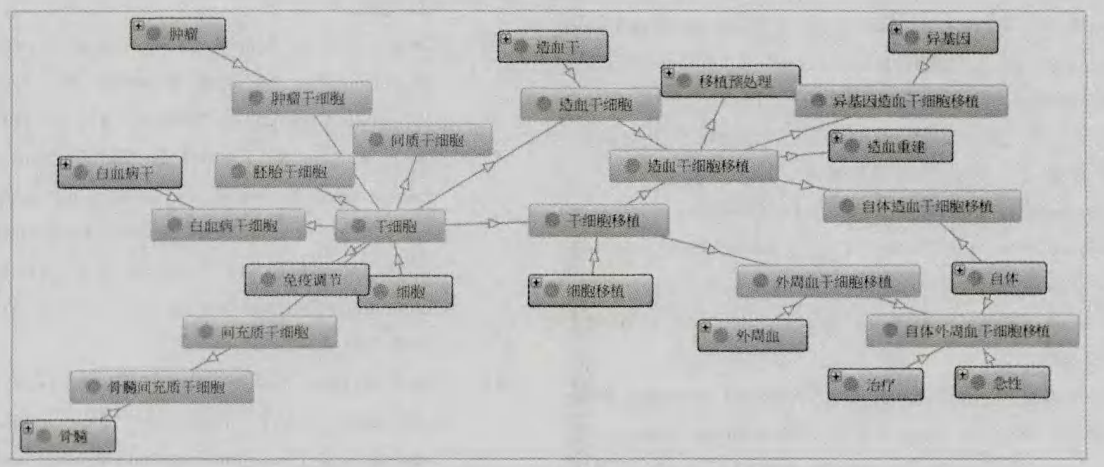

图 10 “干细胞”类层次结构的 spring 图展示

5 结 语

本文提出了一种以文档 - 术语空间为核心、FCA 为手段的中文领域本体层次结构生成的有效方法，并以“白血病”领域为例，对基于 FCA 的医学专业术语层次关联的自动构建进行了详细的论证。在抽取并筛选领域专业术语的基础上，对术语与文档关联进行了修正，进而建立文档 - 术语形式化背景；然后采用 FCA 理论将形式化背景转化为主题领域概念格，进而利用主题概念之间的上下位关系以及文档集合包含对术语层次关联的推理规则，自动生成作为概念属性的术语之间的上下位关联；这种上下位关联可以采用 OWL 进行形式化描述，最终形成领域术语本体，为领域知识的进一步应用如可视化展示等奠定了结构基础。综上所述，本文的创新之处可以归纳为两个方面：一是总结出了一整套利用相关软件及程序自动构建学科层次知识结构的完整方案，为学科知识本体的快速开发提供了一种可参考的操作模式；二是提出了将 FCA 算法作用于文档 × 术语空间，辅以文档包含理论，以构建术语层次关系的方法，并进行了实验论证。

本文也存在不足之处。第一，本文自动构建术语层次关联的理论基础是，术语所在文档集合的包含关系决定了术语之间的层次关系。由于该思想源自对英文术语的描述，直接将其应用于中文文档 - 术语空间所取得的术语层次关系效果并不理想，原因是中文术语的粒度（一般为短语）远远大于英文术语（一般为单词），而导致大量的上下位关系没有完全体现出来。第二，文中没有对构建的术语层次关系进行系统测评，一方面是因为可对照的知识语料较为缺乏，文中引言部分即提到本文构建的术语

分类体系可作为 MeSh 的补充;另一方面则是目前多采用领域专家主观评价,可操作性较差。因此对中文文档-术语空间的修正以及学科知识本体测评的有效开展将是领域本体自动构建今后进一步的研究方向。

参 考 文 献

[1] Nanda J, Simpson T W, Kumara S R T, et al. A methodology for product family ontology development using formal concept analysis and Web ontology language[J]. Journal of Computing and Information Science in Engineering, 2006, 6(2): 103-113.

[2] De Maio C, Fenza G, Lola V, et al. Hierarchical web resources retrieval by exploiting Fuzzy Formal Concept Analysis[J]. Information Processing & Management, 2012, 48(3): 399-418.

[3] Yu M M, Wang J L, Zhao X D. A PAM-based ontology concept and hierarchy learning method[J]. Journal of Information Science, 2014, 40(1): 15-24.

[4] 王昊, 邓三鸿. HMM 和 CRFs 在信息抽取应用中的比较研究[J]. 现代图书情报技术, 2007(12): 57-63.

[5] Terrientes L, Moreno A, Sánchez D. Discovery of relation axioms from the Web [C]// Knowledge Science, Engineering and Management, Lecture Notes in Computer Science, Belfast, Northern Ireland, UK, 2010, 6291: 222-233.

[6] Shamsfard M, Barforoush A A. Learning ontologies from natural language texts [J]. International Journal of Human-Computer Studies, 2004, 60(1): 17-63.

[7] Metaxiotis K S, Samouilidis J E. Expert systems in medicine: academic exercise or practical tool? [J]. Journal of Medical Engineering & Technology, 2000, 24(2): 68-72.

[8] Hwang S H, Kim H G, Kim M K. A data-driven approach to constructing an ontological concept hierarchy based on the formal concept analysis[C]// Computational Science and its Applications (ICCSA 2006), Glasgow, UK, 2006, 3983: 937-946.

[9] MeSH. [EB/OL]. [2014-03-10]. http://www.ncbi.nlm.nih.gov/mesh.

[10] Gruber T. A translation approach to portable ontology specifications[J]. Knowledge Acquisition, 1993, 5: 199-220.

[11] Rios-Alvarado A B, Lopez-Arevalo I, Sosa-Sosa V J. Learning concept hierarchies from textual resources for ontologies construction [J]. Expert Systems with Applications, 2013, 40(15): 5907-5915.

[12] Bast H, Majumdar D. Why spectral retrieval works[C]// Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2005,: 11-18.

[13] Henriksson A, Moen H, Skeppstedt M, et al. Synonym extraction of medical terms from clinical text using combinations of word space models [C]//5th International Symposium on Semantic Mining in Biomedicine (SMBM), Sep. 3-4, Zurich, 2012: 10-17.

[14] Hearst M A. Automatic acquisition of hyponyms from large text corpora [C]// Proceedings of the 14th Conference on Computational Linguistics, Morristown, NJ, USA, 1992: 539-545.

[15] Hearst M A. Automated discovery of WordNet relations [A] // Fellbaum C. WordNet: An Electronic Lexical Database, Cambridge, MA: MIT Press, May 1998: 131-153.

[16] Choi I, Rho S, Kim M. Semi-automatic construction of domain ontology for agent reasoning[J]. Personal and Ubiquitous Computing, 2013, 17(8): 1721-1729.

[17] Nanas N, Uren V, Roeck A D. Building and applying a concept hierarchy representation of a user profile[C]// Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03), New York, NY, 2003: 198-204.

[18] Nevill-Manning C G, Witten I H, Paynter G W. Lexically-generated subject hierarchies for browsing large collections [J]. International Journal on Digital Libraries, 1999, 2(2-3): 111-123.

[19] Sanderson M, Croft B. Deriving concept hierarchies from text[C]// Proceedings of the 22nd annual international ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 1999: 206-213.

[20] Maedche A, Staab S. Ontology Learning for the Semantic Web[J]. IEEE Intelligent Systems, 2001, 16(2): 72-79.

[21] Chuang S L, Chien L F. A practical web-based approach to generating topichierarchy for text segments [C]// Proceedings of the 13th ACM Conference on Information and Knowledge Management (CIKM'04), New York, NY, USA, 2004: 127-136.

[22] Bloehdorn S, Cimiano P, Hotho A. Learning ontologies to improve text clustering and classification[C]// From Data and Information Analysis to Knowledge Engineering, Springer Berlin Heidelberg, 2006: 334-341.

[23] Sumg S, Chung S, McLeod D. Efficient concept clustering

for ontology learning using an event life cycle on the web [C]// Proceedings of the 2008 ACM Symposium on Applied Computing (SAC'08), Fortaleza, Ceara, Brazil, March 16-20, 2008: 2310-2314.

[24] Dupret G, Piwowarski B. Principal components for automatic term hierarchybuilding [C]// Proceedings of the 13th International Symposium on String Processing and Information Retrieval (SPIRE 2006), LNCS, 2006, 4209: 37-48.

[25] Bast H, Dupret G, Majumdar D, et al. Discovering a term taxonomy from term similarities using principal component analysis[C]. Semantics, Web and Mining, LNAI, 2006, 4289: 103-120.

[26] Rizoiu M A, Velcin J. Topic Extraction for Ontology Learning [C]// Wong W, Liu W, Bennamoun M. Ontology learning and knowledge discovery using the web: Challenges and recent advances, Hershey, PA: IGI Global, 2011 : 38-61.

[27] Hwang S H, Kim H G, Yang H S. A FCA-Based ontology construction for the design of class hierarchy [C]//Proceedings of the 2005 International Conference on Computational Science and Its Applications, ICCSA'05, Singapore, May 9-12, 2005, 3482: 827-835.

[28] Fowler M. The taxonomy of a Japanese stroll garden: An ontological investigation using formal concept analysis [J]. Axiomathes, 2013, 23(1): 43-59.

[29] Chen R C, Bau C T, Yeh C J. Merging domain ontologies based on the WordNet system and Fuzzy Formal Concept Analysis techniques [J]. Applied Soft Computing, 2011, 11(2): 1908-1923.

[30] Khalida Ben Sidi Ahmed, AdilToumouh, MimounMalki. Effective ontology learning: concepts' hierarchy building using plain text Wikipedia [C]// CEUR Workshop Proceedings, ICWIT, 2012, 867: 170-178.

[31] Ponzetto S P, Strube M. Deriving a large scale taxonomy from Wikipedia[C]// Proceedings of the 22nd National Conference on Artificial Intelligence, AAAI '07, Vancouver, BC, Canada, 2007, 2: 1440-1445.

[32] Sangno Lee, Soon-Young Huh, Ronald D. McNiel. Automatic generation of concept hierarchies using WordNet[J]. Expert Systems with Applications, 2008, 35(3): 1132-1144.

[33] Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts [C]// Proceedings of the NATO Advanced Study Institute, Banff, Canada, 1982: 445-470.

[34] Priss U. Formal concept analysis in information science [J]. Annual Review of Information Science and Technology, 2006, 40(1): 521-543.

[35] Kuznetsov S O, Poelmans J. Knowledge representation and processing with formal concept analysis[J]. Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery, 2013, 3(3): 200-215.

[36] Xu W, Li W J, Wu M L. Deriving event relevance from the ontology constructed with formal concept analysis [C]//Computational Linguistics and Intelligent Text Processing, Mexico City, Mexico, 2006, 3878: 480-489.

[37] Poelmans J, Elzinga P, Viaene S, et al. Text mining scientific papers: a survey on FCA-based information retrieval research[C]// Proceedings of 12th Industrial Conference, ICDM 2012, Berlin, Germany, July 13-20, 2012, 7377: 273-287

[38] Formica A. Ontology-based concept similarity in Formal Concept Analysis[J]. Information Sciences, 2006, 176 (18): 2624-2641.

[39] Quan T T, Hui S C, Fong A C M, et al. Automatic fuzzy ontology generation for Semantic Web[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(6): 842-856.

[40] Weng Sung-Shun, Tsai Hsine-Jen, Liu Shang-Chia, et al. Ontology Construction for Information Classification [J]. Expert Systems with Applications, 2006, 31(1): 1-12.

[41] OWL[EB/OL]. [2014-03-10]. http://www.w3.org/TR/owl-features/.

[42] Protégé[EB/OL]. [2014-03-10]. http://Protege.stanford.edu.

[43] Falconer S. OntoGraf. [EB/OL]. [2014-03-10]. http://protegewiki.stanford.edu/wiki/Onto- Graf.

(责任编辑 赵 康)