情报学报 ISSN 1000-0135
第33卷第6期594-604,2014年6月
JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 33 No. 6, 594-604 June 2014

doi:10.3772/j.issn.1000-0135.2014.06.004

中文医学专业术语的层次结构生成研究[1)]

王 昊[1,2] 苏新宁[1] 朱 惠[1]

(1. 南京大学信息管理学院,南京 210093;
2. 威斯康辛-密尔沃基大学信息研究学院,威斯康辛州 53211)

摘要 本文基于术语共现理论,利用形式概念分析中概念格的自动生成来推理作为属性的领域专业术语的层次结构并进行可视化展示,进而提出了一整套用于实现领域本体概念层次关系构建的解决方案,具体包括文档/词汇与术语语义关联的识别、领域形式化背景的建立、基于形式概念分析的主题概念的生成、基于主题概念格的术语层次关系抽取、术语层次体系的OWL描述和图形展示等。笔者以"白血病"领域为例,详细论证了无知识库支持环境下中文文本到医学学科术语层次结构的衍化过程,并对以文档术语矩阵(DTM)和词汇术语矩阵(WTM)为形式化背景生成的术语层次体系进行了比较分析。

关键词 医学专业术语 层次结构 本体学习 文档术语矩阵 词汇术语矩阵 形式概念分析 OWL

Study on Hierarchy Structure Generation of Chinese Medical Terminology

Wang Hao[1,2], Su Xinning[1] and Zhu Hui[1]

(1. *School of Information Management of Nanjing University, Nanjing*, 210093;
2. *School of Information Studies of UWM, WI, USA*, 53211)

Abstract Based on the theory of Terms Co-occurrence, this paper uses automatic generation of Concept Lattice in Formal Concept Analysis (FCA) to reason the hierarchy structure of domain terminology as attributes and to achieve the visualization display, and then proposes a set of solutions which could implement the construction of Concept Hierarchies in Domain Ontology, including the identification of the semantic relations between documents/words and terms, building of domain formal contexts, generation of topic concepts based FCA, extraction of terminology hierarchical relations based topic concept lattice, and OWL description and graphical display of terminology hierarchy. With "leukemia" domain as an example, the author demonstrates in detail the derivation process from Chinese text to Medical Terminology Hierarchy Structure without the support of knowledge base, and makes a comparative analysis on the terminology hierarchies which are generated based on formal contexts respectively from the Documents-Terms Matrix (DTM) and Words-Terms Matrix (WTM).

Keywords medical terminology, hierarchy structure, ontology learning, documents-terms matrix (DTM), words-terms matrix (WTM), formal concept analysis (FCA), OWL

收稿日期:2014年4月24日

作者简介:**王昊**,男,1981年生,南京大学信息管理学院副教授,主要研究方向:知识本体构建及应用、数据挖掘技术应用,科学评价和引文分析等研究,Email:ywhaowang@nju.edu.cn。**苏新宁**,男,1955年生,教授,南京大学信息管理学院博士生导师,长江学者,主要研究方向:智能信息处理与检索、科学评价和引文分析等研究。**朱惠**,女,1978年生,南京大学信息管理学院情报学博士生,主要研究方向:知识本体构建及应用研究。

1) 本文受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)、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)资助。

1 引 言

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[1]。OL在"快速开发知识本体以适用于语义网"[2,3]的背景下产生并得到了迅速发展,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],本体学习根据抽取任务的不同,可以分解为如图1所示的自底向上复杂度逐层上升的6个层次。术语(Terms)是指由自然语言描述的在学科中有意义的词语或词组;同义词(Synonyms)是指具有相同语义的术语集合;概念(Concept)是对术语的抽象描述,由内涵(Intension)、外延(Extension)和词汇描述(Lexical realizations)等3部分构成;概念层次(Concepts' Hierarchies)描述了概念之间的非对称层次关系,表现为术语之间的包含关系;语义关系(Semantic Relations)则是指除层次关系以外的所有语义关系的总称,一般通过对象属性的方式加以描述;本体中最复杂的是规则(Rules)元素,公理(Axioms)是指永真规则,目前具有普适性的规则抽取方法还没有出现。

从非结构化文本中抽取本体元素是当前OL的研究热点。对于字符语言,面向文本资源的OL研究和应用已经覆盖所有层次,但在公理和约束规则的抽取上仍涉及较少[5,6];而对于象形语言,由于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[7],而且面向文本资源层次关系的抽取也多停留于实验论证阶段,基本上还没有可实际应用的工具或方法出现。本文试图将形式概念分析方法以及术语共现理论引入到医学本体的术语层次关系构建中,利用概念格的生成结果来自动推导术语属性之间的包含关系,进而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,并通过对形式化背景的改造来优化术语层次体系,从而形成一整套针对学科资源抽取术语层次关系的行之有效的解决方案。

2 近期相关研究

术语层次结构建立是本体学习中概念层次关系识别的一个具体应用,它着眼于领域中现实存在的术语,根据领域文本的描述,由机器自动建立术语之间的上下位关系。①Hearst模式(Hearst-Pattern)[8,9]是依据语言学规律抽取术语层次的一种典型方法,其借助于具有上/下位关系指示功能的模式短语,例如"such as","and other","especially"等来判断术语的层次关系。然而具有指示作用的模式无法穷举,特别是在语言规律极其复杂的中文环境下,术语层次关系的召回率非常低。②术语文档空间是信息检索领域用于描述文本的一种简洁而有效的统计手段,对其进行数据分析是目前获取术语关联的主要方法,具体包括术语共现理论[10],即"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"的应用;聚类分析则是依据术语的文档属性描述将术语聚集成若干个具有分层结构的主题簇[11,12];潜在语义分析(Latent Semantic Analysis, LSA),则是对术语文档矩阵进行奇异值分解(Singular Value Decomposition, SVD)以获取其所有特征,进而选择主要特征计算近视矩阵,并根据近视矩阵的有效性评价推导术语向量之间的上/下位关系[13,14]。③基于知识库的方

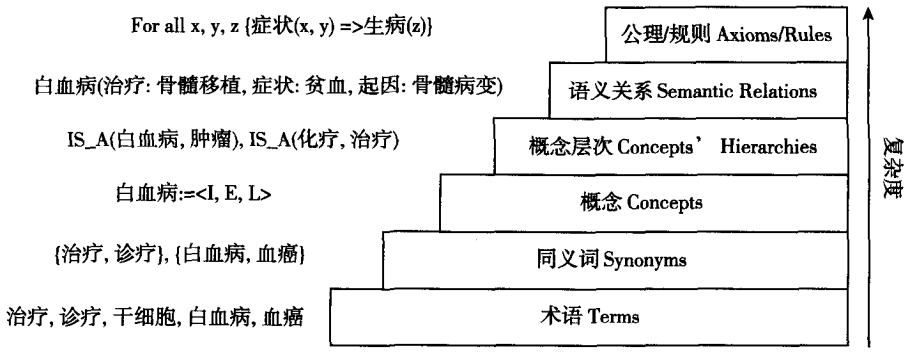

图1 本体学习的层次体系

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[15,16]。其实质是将一个"短"术语扩展为一个"长"而"准"的文字描述来更好的揭示术语的语义特征。但是,中文知识库资源的缺乏使得该方法具有较强的领域依赖性。

形式概念分析(Formal Concept Analysis, FCA)是 Wille 在 1982 年提出的一种数学理论,后来逐渐演化成一种用于数据分析、知识表示、信息管理的重要方法[17~19]。FCA 用对象和属性间的二元关系来表达领域中的形式化背景,从中派生出包括内涵、外延和泛化/例化关系等在内的概念格。由于概念格与本体理论中概念层次的定义不谋而合,因此利用概念格生成算法来自动构建本体概念层次结构成为了一种典型的有效方法[20~22]。于是,研究人员试图将 FCA 与术语文档空间相结合来构建学科本体(仅包括层次结构)[23~25]。然而,FCA 基于属性集合的交叉来生成概念及其层次结构的操作模式与领域术语的原始存在以及术语的属性描述相对匮乏形成了固有矛盾[26],即:①FCA 作用于术语文档空间生成的概念是抽象主题,而非术语实体;②以文档作为术语的描述属性,具有属性越多对象越泛化的特点,与 FCA 的描述正好相反。

3 基于 FCA 的术语层次结构生成方法

本节重点探讨基于 FCA 的术语层次结构的构建方法和过程,并分析形式化背景的变化对术语层次结构的影响方式和效果。同时,基于期刊论文文本生成的医学专业术语层次体系可作为 MeSH 本体的有效补充[27],以应用于其他医学知识发现研究[28]。

3.1 文档-术语空间在 FCA 中的应用

定义 1:形式化背景(Formal Context, FC)是一个三元组 $F=(O, M, R)$,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"。

FC 实际上就是对象×属性矩阵。那么文档×术语矩阵(Documents-Terms Matrix, DTM)也可以映射为形式化背景 $F_{DTM}=(D, T, I)$[29],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,用术语在文档中是否存在来表示。在 F_{DTM} 中,术语被认为是文档对象的属性。表 1 为 F_{DTM} 的一个示例,列出了"白血病"领域部分文档与术语之间的假设关联,表中"√"表示术语在文档中出现,$D=\{D1, D2, D3, D4, D5, D6, D7, D8\}$,$T=\{$白血病,粒细胞白血病,干细胞,细胞,治疗,粒细胞,干细胞移植,基因疗法$\}$。

定义 2:在 $F=(O, M, R)$ 中,可定义两个映射 f 和 g:$\forall O_x \subseteq O: f(O_x)=\{m \in M \mid \forall o \in O_x, oRm\}$,$\forall M_y \subseteq M: g(M_y)=\{o \in O \mid \forall m \in M_y, oRm\}$。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)。

表 1 "白血病"领域文档与术语的形式化背景示例

D \ T	白血病	粒细胞白血病	干细胞	细胞	治疗	粒细胞	干细胞移植	基因疗法
D1	√	√	-	√	-	√	-	-
D2	√	-	√	√	√	-	√	-
D3	√	-	√	√	√	-	√	-
D4	√	-	-	-	√	-	-	√
D5	√	-	-	√	-	√	-	-
D6	√	-	√	√	-	-	-	-
D7	√	√	-	√	√	√	-	√
D8	√	√	-	√	√	√	-	-

那么在 $F_{DTM}=(D, T, I)$ 中，设 $X\subseteq D, Y\subseteq T$，根据定义 2 可得：$\sigma(X)=\{t\in T\mid \forall d\in X:(d,t)\in I\}$，$\tau(Y)=\{d\in D\mid \forall t\in Y:(d,t)\in I\}$。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 $c_1=(\{D4,D7\},\{$白血病，治疗，基因疗法$\})$ 即称为一个主题概念，其外延为 $\{D4, D7\}$，内涵为{白血病，治疗，基因疗法}，该主题描述的是“白血病”的一种“治疗”方法“基因疗法”，而文档 D4 和 D7 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，文档集合被称为该主题的外延，而所有术语一起形成了其内涵。在这里，笔者发现一个现象，对象的属性一般都是从不同角度对对象进行描述，如汽车，可以从排量、颜色、形状等方面进行描述，这些属性一般来说不同类且相互之间没有交叉，然而在 F_{DTM} 这一特殊的形式化背景中，作为文档对象的属性却是同一类对象“术语”，它们之间不可避免的存在一定内容交叉。例如，“基因疗法”似乎是“治疗”的一个子集，而“治疗”又是“白血病”的一个描述方面。可见，在 F_{DTM} 的主题概念中似乎隐藏着术语属性间的某种关联。

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 $c_1\leqslant c_2$。F 中所有概念及其层次关系被记作 $C(F)$，称为概念格（concept lattice）[30]。

定义 3 表明，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 $c_1=(\{D1,D5,D7,D8\},\{$白血病，细胞，粒细胞$\})$，$c_2=(\{D1,D2,D3,D5,D6,D7,D8\},\{$白血病，细胞$\})$，不难发现 $M_2\{$白血病，细胞$\}\subseteq M_1\{$白血病，细胞，粒细胞$\}$，因此 c_1 是 c_2 的子概念；c_2 主题为白血病相关细胞的内容，而 c_1 则专门探讨白血病相关细胞中的粒细胞，可见就主题而言前者研究范围更大，是后者的上位概念，而探讨上位概念的文档也相对更多。

根据 FCA 的上述定义，笔者基于概念格生成算法计算出了表 1 所示形式化背景的概念格，用图 2 所示的 Hasse 图展示。①图中每个圆形节点表示一个主题概念 c，圆形大小表示主题外延的个数；处于上层为父概念，下层为子概念，自上而下，概念层次降低，而属性集合逐渐增大，对应的外延数量将会越来越少；最上层概念包括了所有外延，其内涵为所有外延的共有属性，而最下层概念包含了所有属性。在本例中，“白血病”出现在所有文档中，其应该在最上层概念的属性集合中，而含有所有术语的文档不存在，即最下层概念外延为空。②每个概念由两部分构成，上半部分代表属性 M，下半部分代表对象 O。为了简化概念格，图中每个节点仅显示出了相对其父节点新增的属性和相对其子节点新增的对象。因此，若属性半圆呈蓝色表示有新增属性分布于该节点上，对象半圆呈黑色表示有新增对象分布在该节点上，而每个概念节点的属性集合和对象集合分别为以该节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）。例如图中最右侧“D6，干细胞”节点，其属性集合应为{白血病，细胞，干细胞}，对象集合为 $\{D6,D2,D3\}$，形成一个完整的主题概念 $c\{\{D2,D3,D6\},\{$白血病，细胞，干细胞$\}\}$。

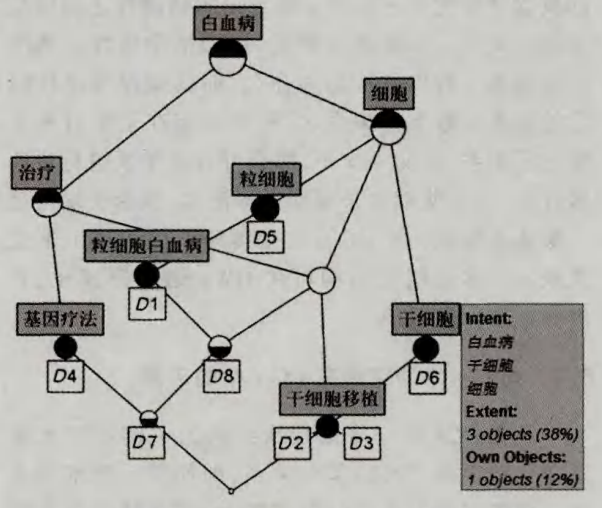

图 2　根据表 1 所示形式化背景示例所生成的概念格

笔者重点考察图中属性标签，发现这些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方式从上层概念中派生出来，因此根据 Hasse 图的示意，新增术语所在主题概念的文档集合为以其为根节点的下子树的所有文档总和，包括了其下位概念新增术语的所在文档集合。例如，图中“D4，基因疗法”节点，新增了“基因疗法”属性，该属性应该出现在 D4 和 D7 中；而该节点的父节点，新增了“治疗”属性，其出现在文档 D4 和 D7 以及另一分支的 D2 和 D3 中，即父

节点新增术语所在文档集合必然包括了子节点新增属性所在的文档集合。那么,根据术语共现理论"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"的论断[20],即术语"治疗"是"基因疗法"的上位术语。同理可得,"白血病"是"细胞"和"治疗"的上位术语,"细胞"是"粒细胞"和"干细胞"的上位术语等。于是,用于构建概念层次结构的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,而且根据"概念格中子概念新增的术语属性是父概念新增术语属性的下位类"的结论,利用术语属性在概念格中的分布以及主题概念的层次结构,可推导出所有术语间的上下位直接关联。

综上所述,基于FCA的概念格以及术语的文档共现理论可以自动推导出作为属性的术语的层次结构,具体过程分为4个步骤:①根据术语在文档中出现状况,建立领域形式化背景,即文档×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关系推导出术语属性之间的层次语义关联。需要说明的是,概念格中属性标签所在的概念一般以属性名称命名,而其他没有属性标签的概念以数字标识命名;概念格保存了所有的直接父子关系,即父->子,那么可对该关系进行自连接计算,直到父端和子端均为术语名,或者子端为最下层概念标识为止,可自动获得所有的术语上下位关联。④术语层次结构可用OWL语言形式化,并进行可视化展示。

3.2 词汇-术语空间在FCA中的应用

基于FCA建立术语层次结构的关键在于术语所在文档集合之间的包含关系,即如果一个术语出现在较多文档中或其分散度较大,那么其泛化程度较高。该理论最初是在研究英文的文档术语空间时发现的一个在文本中具有较强普适性的规律,后来被广泛作为英文术语层次关联构建的依据。但是,将这一规律直接应用于本文数据中,却存在较大问题。术语共现原是指两个术语同时出现在一个文档中,而文档一般要求具有一定的篇幅以保证术语共现具有较大的概率;而本文数据中文档指关键词文本,一般认为具有父子或相似关系的术语同时作为一个文档的关键词是不合适的,例如表1中,"治疗"和"基因疗法"一般来说不会同时出现在一个文档的关键词文本中,因为关键词的选择需要具有一定代表性,不能语义重复或较大交叉。因此,作为修正,笔者把共现范围扩展到了题名,即认为两个术语只要在题名或关键词文本中同时出现,即认为存在共现关联,这在一定程度上强化了术语之间的语义关系,但是由于题名的篇幅较短,修正的效果并不理想。可见,上述术语共现规律直接应用于题名和关键词等短文本中,特别是在构词极其复杂的中文环境下,由于术语共现概率较小使得所获得术语层次结构具有较低的召回率。

在原始的文档术语空间中,术语共现的概率较小,因此术语之间的层次关系召回困难;那么将短文本分解成粒度较小的词汇,是否可以增加术语共现的可能性呢?如图3所示,上层为文档级的术语共现,假设术语A出现在$D1$和Dm中,术语B出现在$D2$中,术语C出现在Dm中,很明显A和B以及B和C均不存在共现;现在将文档全部分解为词汇(下层),若$D1$和$D2$中均存在词汇$W1$和$W2$,那么术语A和B因为与$W1$和$W2$均存在共现关联,而致使A与B在词汇级上共同出现了2次,同理B和C由于$W2$、Wi和Wk的关系,共现了3次;原本存在共现关联的术语A和C与Dm分解后的词汇均存在关联,因此两者将保持原有相关状态。可见,随着文本粒度的降低,作为共现桥梁的词汇的出现率比文档要高得多,这将会明显提高共现空间的稠密度,并在保持原有共现关联的基础上进一步强化了术语之间的语义关系。于是,上述的术语包含规律将被改造为:当且仅当术语A的相关词汇集合是术语B的相关词汇集合的超集时,术语A包含术语B;与术语相关的词汇量越大,说明术语的语义分散度越大,其泛化程度也越高。

利用词汇-术语空间代替文档-术语空间的主要目的是增加术语之间的共现关联,其实质是用一个词汇集合来描述术语。但是这种方法很显然也会极大增加词汇与术语之间无意义的关联,如图3中<A, Wj>和<B, Wl>,导致相关术语如A和B之间的交叉关联也大量增加,这会严重干扰术语层次关联的计算。因此,词汇与术语之间语义关联的有效性需要通过一定的测评指标如关联频次等进行验证。表2为"白血病"领域词汇与术语关联的一个示例,根据该形式化背景同样可以计算出对应的概念格,如图4所示。对比图2,同理,根据术语相关词汇集合之间的包含关系可以推导术语之间的上下位关联,反映在图4中即节点上方的术语属性相互之间呈现出了一种层次结构;而不同点在于作为主

题概念的外延变成了词汇，而且由于词汇与术语之间的共现关联明显增强了，即表 2 的稠密度大于表 1，使得图 4 概念格的层次性更为明显，内涵交叉的概念（即图中上半部分空白的圆）也更多。

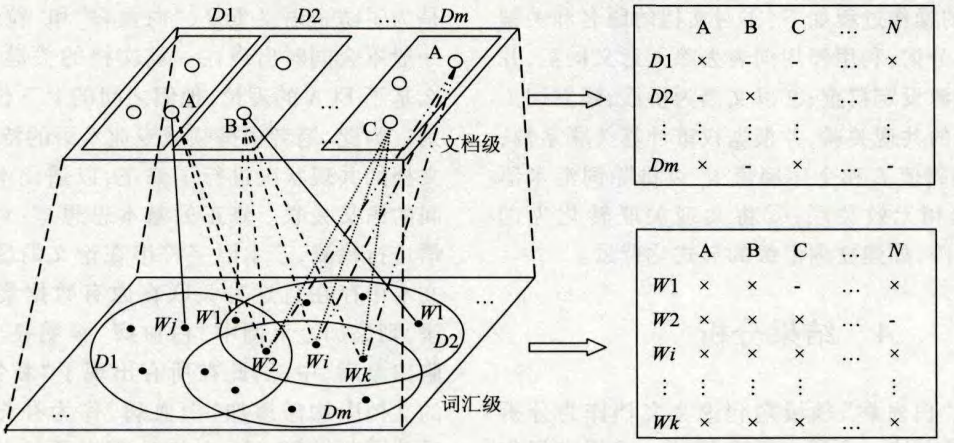

图 3　文档级的术语共现转化为词汇级的术语共现

表 2　“白血病”领域词汇与术语关联示例

W \ T	白血病	髓系白血病	基因	髓系	急性髓系白血病	基因疗法	基因融合
白血病	√	√	√	√	√	√	√
基因	√	√	√	√	-	√	√
粒细胞	√	√	√	√	√	√	√
治疗	√	√	√	√	√	√	-
信号	√	-	√	-	-	√	-
移植	√	√	√	√	√	-	√
融合	√	-	√	√	-	-	√
蛋白质	√	√	√	√	√	-	-
急性	√	√	-	√	√	-	-
淋巴	√	√	-	√	-	-	-

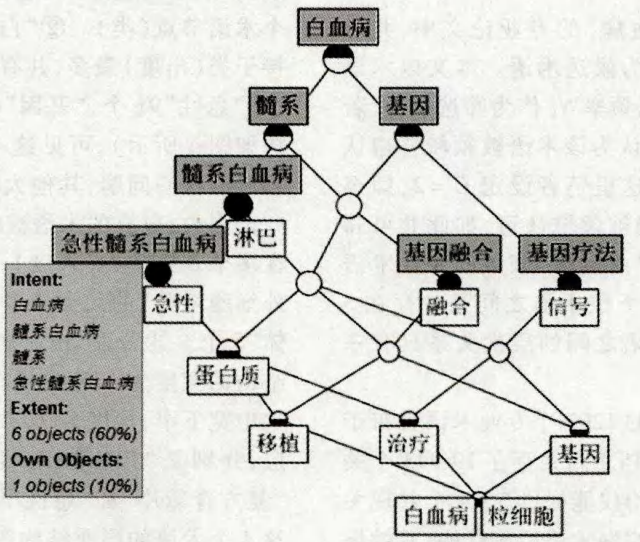

图 4　根据表 2 所示词汇 - 术语关联示例所生成的概念格

综上所述,用词汇代替文档作为形式化背景中的对象,能够在一定程度上强化术语之间的共现概率,其具体的操作过程如下:①对文档的题名和关键词文本进行分词,利用停用词表去除无意义词汇,并根据词汇来源设定权重;②以文档为桥接,建立词汇与术语之间的共现关联,并根据权重计算关联系数;③设立全局阈值 L 和个体阈值 K_i 以排除词汇术语之间的偶然和无效关联;④将共现关联转化为词汇-术语矩阵,即建立词汇级的形式化背景。

4 结果分析

本文以"白血病"领域期刊论文文档作为分析对象,分别以文档-术语空间和词汇-术语空间作为形式化背景进行了FCA,并对由此获得的术语层次体系进行了详细的比较论证。

4.1 数据预处理

笔者从2009~2012年万方数据收录的医学专业期刊论文数据集中检索出题名中含有"白血病"的6194条记录,以其题名和关键词文本建立了该领域的文档集合。本文重点讨论中文术语层次体系的构建及比较,因此直接以论文中提供的关键词作为候选专业术语。但是,由于论文中的关键词特别是中文关键词多由作者给出,没有经过规范化处理,存在较大的随意性。为此,有必要对关键词进行预处理,从中筛选出具有一定领域认可度的词汇作为专业术语,同时对专业术语与文档之间的关联性做初步修正。

在6194篇关于"白血病"的专业论文中,共可抽取出6454个关键词作为候选术语。本文以术语在整个文档集合中的出现频率 N_k 作为筛选条件,若 N_k 大于词频阈值 C,那么认为该术语被领域普遍认可,可作为专业术语。在这里笔者设定 $C=2$,即至少出现3次的关键词才能被领域认可,如此共可筛选出1203个关键词作为"白血病"领域的专业术语集合;而且由此获得的上下位术语之间至少存在3个文档的重合,可保证术语之间的层次关系具有一定可信度。

经过词频筛选,获得的1203个专业术语分布于6116篇论文中,术语与文档之间共存在19 943个关联,也就是说每个文档平均仅能提供3~4个共现关联,在如此稀疏的矩阵中挖掘术语层次结构,其效果可想而知;再加上关键词之间一般语义不重复,例如"粒细胞白血病"很明显是"白血病"的下位术语,但是为了防止语义重复,"白血病"和"粒细胞白血病"一般不会同时出现在一篇文档的关键词集合中,那么基于FCA的理论,他们之间的上下位关系很难识别。为此,笔者根据领域专业术语的特点,对术语与文档的共现状况进行了修正,以强化术语与文档之间的语义关联。修正的基本思想是:对所有专业术语进行检测,若术语字符串在论文的题名或关键词文本中存在且这种关联在原有数据集合中没有记录,则添加。上例中"白血病"很明显出现在"粒细胞白血病"中,因此在所有出现了"粒细胞白血病"的文档中均需添加"白血病"作为补充。如此修正后共可获得70 436个共现关联,而这些语义关联将成为构建术语间层次关系的主要依据。

4.2 基于文档-术语空间的FCA结果分析

将文档术语间关联转化为文档×术语矩阵(Documents-Terms Matric, DTM)作为形式化背景 FM1 = {D, T, I}。其中 D 集合中共有6116个对象,T 集合中有1203个属性,I 集合中为70 436个关联。对FM1进行形式概念分析,共生成主题概念81 323个,并根据属性首次出现的概念之间的上下位关系可以推导出术语间的上下位关系共1688对,不包括传递包含,笔者将该术语层次体系记为HR1,并采用OWL语言进行形式化描述。在HR1中,①共涉及术语1203个,其中有11个术语完全共现,形成了5对同义术语,如"FLAG"和"FLAG方案"、"高压液相"和"色谱法"等,因此共计有1197个术语节点(类)。②"白血病"术语为根节点,其直接子类(出度)最多,共有374个,其次为"细胞"193个、"急性"91个、"基因"41和"治疗"39个(层次结构如图4所示),可见这4个大类是目前"白血病"领域的核心问题,其他大部分热点术语均处于其研究范围内;也有的术语被纳入到了不同类目下,即存在多个直接父类(入度),其中"四氢叶酸钙"和"体外细胞培养"的父类最多,有6个,其次是"VLA4抗体"5个。③若以"白血病"为第1层,那么整个层次结构的宽度为675,出现在第3层,整体形状呈现上尖中宽下窄;深度为7,共有4个术语出现在了第7层,分别是"PMLRARα融合基因"、"儿童ALL"、"复方青黛片"和"急性早幼粒细胞白血病(APL)",这4个术语的层次结构图如图5所示。

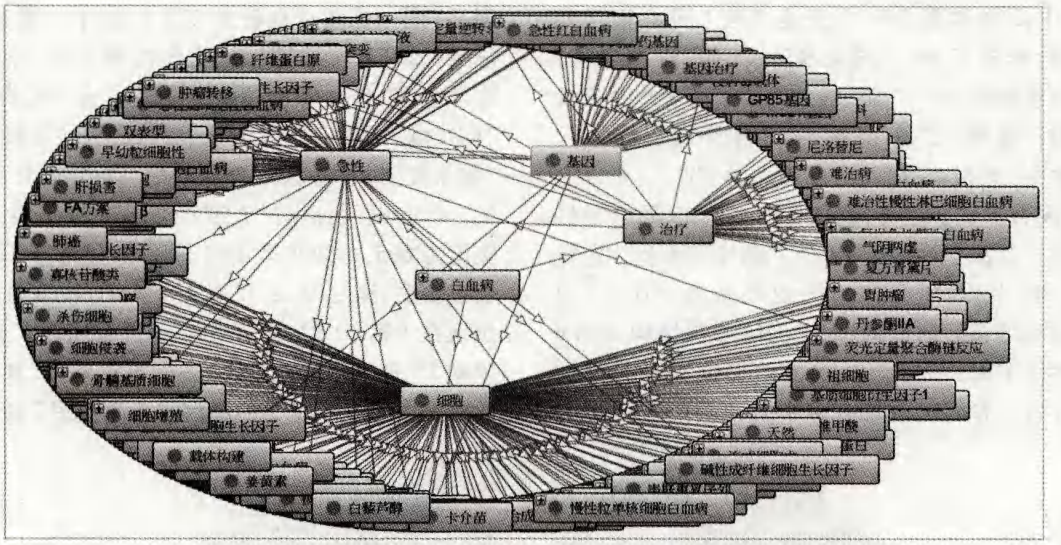

图 5 "白血病"领域中子类最多的 4 个术语的层次结构

图 6 "白血病"领域深度最大(7 层)4 个术语的层次结构

4.3 基于词汇 - 术语空间的 FCA 结果分析

笔者将 DTM 转化为词汇 - 术语矩阵(Words-Terms Matrix, WTM),重新构建了"白血病"领域的形式化背景 FM2 = {W, T, I}。①对 6116 个题名(权重为 1)及 70 436 个关键词(权重为 2)文本进行分词,然后利用中文停用词表对获得的中文词汇进行过滤,可得 4091 个非停用词;②以文档作为连接依据,建立词汇与术语之间的共现关联并计算关联权重,可形成 188 714 个 < 词汇, 术语, 权重 > 3 元组;③排除词汇术语关联中的偶然和无效关联,它们可能导致术语之间出现无意义交叉而影响术语层次结构的建立。对于前者,笔者以关联阈值 L 在整体上进行排除,本文中选择 L = 4,即认为词汇与术语之间的关联权重至少要达到 5 以上才是可信的;对于后者,笔者采用权重均值 K_i 为每个术语选择最重要的词汇作为其有效描述特征,以减少术语之间无意义的词汇交叉。需要说明的是,为了保证"白血病"作为所有术语的父节点,笔者假设其与所有词汇之间存在关联。通过上述筛选,可保证术语量 T = 1203 个,词汇 W = 872 个,他们之间的关联数量 I = 16 199 个。相比之下,WTM 比 DTM 的体积大大缩小了,但稠密度却极大提高了。

对 FM2 进行形式概念分析,共可生成主题概念 212 832个,同理可推导出术语层次结构 HR2 共包含直接上下位关系 4911 对。①FM2 生成的主题概念约为 FM1 的 2.6 倍多,HR2 中的术语层次关系约为 HR1 的 3 倍,可见 FM2 稀疏程度的减弱使得对象的属性集合之间相互重叠的现象更为明显,导致大量交叉主题的产生,术语间的内容包含也更为复

杂。②HR2中涉及100个完全共现(词汇集合相同)术语,形成了46个同义术语类,如"棉絮状斑"和"视网膜内出血"、"NK细胞白血病"和"侵袭性NK细胞白血病"、"免疫性"和"自身免疫性"等。③"白血病"为根节点,其出度最大为605,一半以上的术语成为了其下位类,就这点而言HR2的合理性弱于HR1;其次是"细胞"79个、"急性"64个、"急性白血病"56个和"研究"52个,这些术语对应了"白血病"领域的核心研究内容,它们的层次结构如图6所示,对比HR1,一方面各术语的下位类分布更加均匀了,另一方面术语具有多个上位类的现象更为明显(图中连线存在较多交叉),其中入度最大的术语为"CXCR7、脓毒症、强直性脊柱炎、新城疫病毒",竟有154个上位类,其次还有"阿奇霉素"、"CD184"、"PAKT"、"先天性白血病"、"质性研究"和"靛玉红、吲哚胺2、杂合性缺失"等均有上百个上位类。④若以"白血病"为第1层,那么整个层次结构的宽度为605,出现在第2层,此后逐层缩小,呈现为倒金字塔型;深度为12,仅有1个术语"先天性白血病"出现在了第12层,另有3个术语出现在了第11层,分别为"多毛细胞白血病"、"幼淋巴细胞"和"先天性白血病",这3个术语的层次结构图如图7所示。

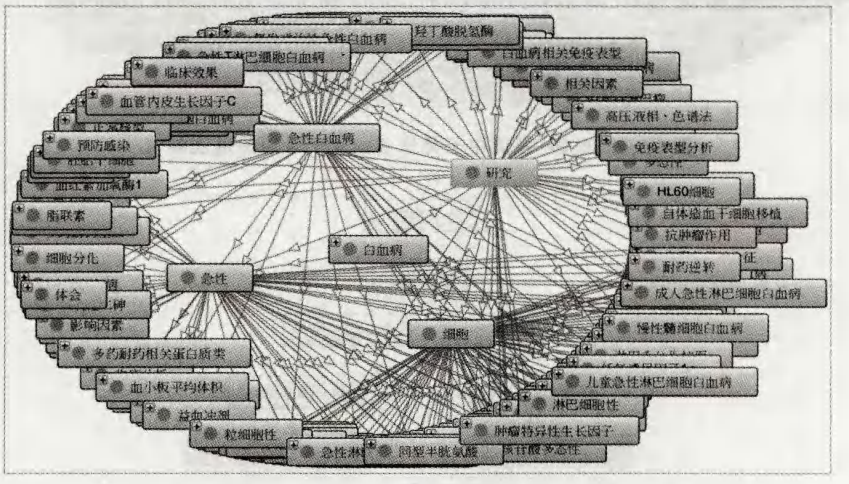

图7 "白血病"领域中子类最多的4个术语的层次结构

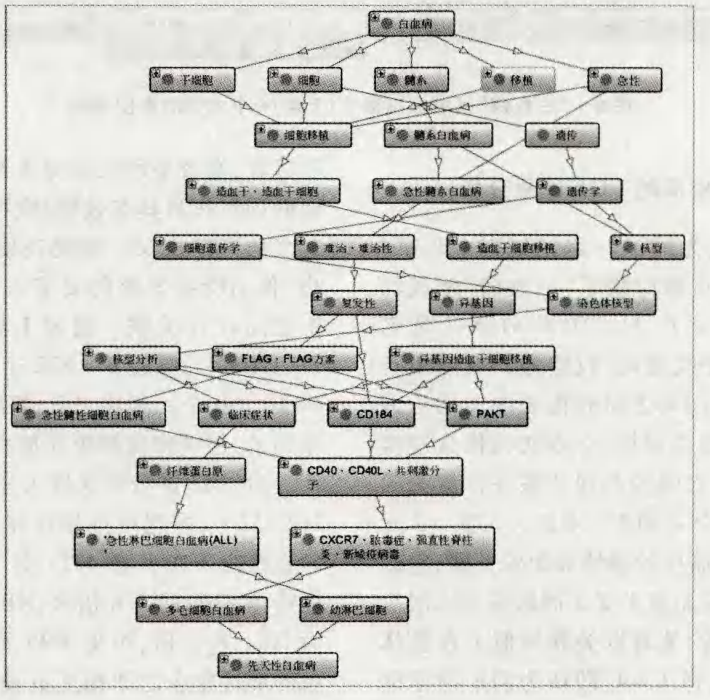

图8 "白血病"领域深度最大(12层)术语的层次结构

5 结 语

本文基于术语共现理论,利用 FCA 中概念格的自动生成来推理作为属性的领域专业术语的层次结构并进行可视化展示,进而提出了一整套用于实现领域本体概念层次关系构建的解决方案,具体包括文档/词汇与术语语义关联的识别、领域形式化背景的建立、基于形式概念分析的主题概念的生成、基于主题概念格的术语层次关系抽取、术语层次体系的 OWL 描述和图形展示等。笔者以"白血病"领域为例,详细论证了无知识库支持环境下面向中文文本的专业术语层次关系的抽取过程,并对以 DTM 和 WTM 为形式化背景生成的术语层次结构进行了充分比较研究。

从总体上来看,将 DTM 转化为 WTM,描述术语的特征粒度变小了,术语间关联的概率变大了,因此术语间上下位关系更为丰富,而整个术语层次结构也变得更为复杂。基于 DTM 获得的术语层次关联主要依赖于术语的字面包含,这对专业领域而言具有一定的合理性;而基于 WTM 获得的术语层次结构更多的依赖于术语的语义包含,从理论上讲更加可信,但是大量噪声词汇的存在使得层次关联抽取的计算量呈指数增加,而且极易导致无意义层次关系的生成,这就需要对词汇进行严格控制和筛选。本文采用的筛选方式主要注重处理的全局性和计算的简便性,存在不合理之处有待今后研究改进。

参 考 文 献

[1] 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.

[2] 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.

[3] Maio C De, 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.

[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, 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] 谷俊, 严明, 王昊. 基于改进关联规则的本体关系获取研究[J]. 情报理论与实践, 2011(12): 121-125.

[8] 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.

[9] 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.

[10] 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, 1999: 206-213.

[11] 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.

[12] 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.

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

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

[15] Ahmed K B S, Toumouh A, Malki M. Effective ontology learning: concepts' hierarchy building using plain text Wikipedia [C]//Proceedings of CEUR Workshop, ICWIT, Hawaii, USA, Nov. 11-16, 2012, 867: 170-178.

[16] Lee S, Huh S Y, McNiel R D. Automatic generation of concept hierarchies using WordNet [J]. Expert Systems with Applications, 2008, 35(3): 1132-1144.

[17] Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts [A]// Rival I. Ordered Sets, 1982: 445-470.

[18] Priss U. Formal concept analysis in information science [A]//Blaise C. Annual Review of Information Science and Technology, ASIST, 2006, 40(1): 521-543.

[19] Pei Z, Ruan D, Meng D, et al. Formal concept analysis based on the topology for attributes of a formal context [J]. Information Sciences, 2013, 236: 66-82.

[20] 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.

[21] Kolozali S, Barthet M, Fazekas G, et al. Automatic ontology generation for musical instruments based on audio analysis[C]. IEEE Transactions on Audio Speech and Language Processing, 2013, 21(10): 1-14.

[22] 张云中, 徐宝祥. 基于形式概念分析的领域本体描述模型研究[J]. 图书情报工作, 2010(14): 111-115.

[23] Weng S S, Tsai H J, Liu S C, et al. Ontology construction for information classification [J]. Expert Systems with Applications, 2006, 31(1): 1-12.

[24] 黄美丽, 刘宗田. 基于形式概念分析的领域本体构建方法研究[J]. 计算机科学, 2006, 33(1): 210-212,239

[25] 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.

[26] 刘萍, 高慧琴, 胡月红. 基于形式概念分析的情报学领域本体构建[J]. 图书情报知识, 2012(3): 20-26.

[27] Kim IC. FCA-based ontology augmentation in a medical domain[C]//Karagiannis D, Reimer U. Lecture Notes in Artificial Intelligence, 2004, 3336: 408-413.

[28] 徐坤, 曹锦丹, 毕强. FCA在医学领域文本分类中的研究和应用[J]. 现代图书情报技术, 2012(3): 23-26.

[29] 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

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

(责任编辑 魏瑞斌)