《智能信息处理》课程作业

形式概念分析在知识结构探测中的应用

姚子晗

作业	分数[20]
得分	

形式概念分析在知识结构探测中的应用

姚子晗

(大连海事大学 计算机科学与技术 辽宁省大连市 中国 116026)

摘　要　在图书情报学的大背景下，对比作者共被引分析、作者文献耦合分析和共词分析的传统知识结构探测方法，文章研究基于形式概念分析(FCA)的学科知识结构探测方法。为了更加清晰合理地描述复杂概念及概念之间的多层次关系，同时力求将聚类和关联特征客观地显示出来，文章利用 FCA 方法研究学科知识结构的定义，以及运用及定义和分析表示结构模型，详细的阐述基于 FCA 的学科知识结构的构建方法的理论及应用。研究探测到新世纪以来图书情报学的 9 个主要研究主题，并揭示每个研究主题的核心关键词和活跃作者。与传统的知识结构探测方法相比，基于形式概念分析的学科知识结构探测方法更胜一筹。

关键词　形式概念分析　概念格　知识结构　图书情报学

Application of formal concept analysis in knowledge structure detection

Yao zihan

(Dalian maritime university, LiaoningDalian,116026,China)

Abstract　*Abstract In the background of Library and information science, comparative author co citation analysis, author literature of traditional knowledge structure coupling analysis and co word analysis method of detection, this paper based on formal concept analysis (FCA) method for detection of knowledge structure. In order to more clearly and reasonably describe the multi-level relationship between complex concepts and at the same time, in order to clustering and association characteristics displayed objectively, this study defined knowledge structure by using the FCA method, and the use and definition and analysis structure model, elaborated the theory and application of the construction method of knowledge structure of the FCA based on. 9 main research topics of Library and Information Science in the new century are studied, and the key key words and active author of each research topic are revealed. Compared with the traditional detection method of knowledge structure, knowledge structure better detection method based on formal concept analysis.

Key words　Formal concept analysis; Concept lattice; knowledge structure; library and Information Science;

1 引言

科技文献是科学研究活动最直接的表达形式，是科研人员研究活动的产物，具有重要的研究意义。其中，对学科知识结构的探测一直是文献计量学、科学计量学和情报学的研究主题。目前对知识结构的探测多以科技文献出版物的相关特征为基础。常用的探测方法主要分为两大类:一类是基于引文的方法从宏观的角度探测知识结构; 另 一类是基于内容分析的方法从微观的角度描述知 识结构。知识结构包括两个要素:内容和结构，内容就是知识节点，结构则体现知识节点的相互关系[1]。通常来说，学科知识结构是指特定学科所包含的知识元素及其相互关联所形成的具有层次结构的知识体系，能够系统地体现该领域知识的基本构成和不同知识之间的关联。J. F. Nicolai 和 P. Torben 强调知识元素不仅仅指显性知识元素(概括其主题的某些基本概念) ，也包 括隐性知识元素(作为知识生命载体的人)[2]。B. Kedrov 在“学科的组织和发展中的知识结构”[3]一文中 明确指出科学知识的一般结构应包括下述几个方面: 作为一切科学家共同从事的一项活动而言的一般科 学、科学的各个具体分支，以及每个科学家进行科学活

动的狭窄领域。这样便形成了一个简单的三分系列：一般科学—具体分支—作为个体的科学家。个体科学家在单独研究过程中的发现融化在特定学科内，该学科内各科学家的成就决定着学科整体知识的发展。因此，学科知识结构是不同类型知识元在学科发展中相互融合、相互作用形成的复杂关联关系。不仅仅包含 主题概念和它们的层次关系，还应包括学者以及他们所关联的学术主题。基于“概念由外延和内涵组成的思想单元”这一 哲学上的理解，德国学者 R．Wille 教授于 1982 年开创了形式概念分析(formal concept analysis，FCA)研究领域[4]。形式概念分析理论是一种建立在数学基础之上，用于对数据集中的概念结构的识别、排序和显示的数据分析理论。形式概念分析强调以人的认知为中心[5]，提供了一种与传统的、统计的数据分析和知识表示完全不同的方法，使得它在知识发现过程中具有独特的优势。但利用 FCA 来探测学科知识结构的研究很少，虽然本体和知识结构的目标都是捕获相关领域的知识，形成对该领域知识的共同理解，但在知识结构探测中，概念的外延(学者)和内涵(主题)是同样重要的两方面，而本体则更强调概念的内涵部分。本文拓展经典形式概念分析理论并将其应用于学科知识结构的探测，以揭示学科内不同层次的主题概念、学者以及概念和学者之间的复杂关联关系。具体来说，着重研究以下问题:①基于 FCA 的学科知识结构定义和表示模型;②基于 FCA 的学科知识结构构建方法;③以图书情报学科为例，对基于 FCA 的学科知识结构探测进行实证研究。

2 基于形式概念分析的学科知识结构定义及表示模型

2.1 形式背景和概念格

定义 1：形式背景是一个知识学科结构的三元组 KS=(A，K，R)，其中 A 是作者(对象)的集合，K 是主题关键词(属性)的集合，R 是 A 和 K 之间的一个二元关系，即 $R \supseteq A \times K$。aRk 表示 $a \in A$ 与 $k \in K$ 之间存在关系 R，读作作者(对象)a 具有关键词(属性)k。表 1 是一个由 8 个作者和他们在论文中使用过的 9 个关键词所构成的形式背景。其中，“×”表示作者 a_i 标注了关键词 k_j，空格表示作者 a_i 未使用关键词 k_j(下同)。

表 1 形式背景举例(由 8 个作者和 9 个关键词生成)

	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9
a_1	×	×					×		
a_2	×	×					×	×	
a_3	×	×	×				×	×	
a_4	×		×				×	×	×
a_5	×	×		×		×			
a_6	×	×	×	×		×			
a_7	×		×	×	×				
a_8	×		×	×		×			

定义 2:设 P 是作者集合 A 的一个子集，定义 $f(P)=\{k \in K \mid \forall a \in P, aRk\}$，表示 P 中作者共同关键词的集合;相应地，设 T 是主题关键词集合 K 的一个子集，定义 $g(T)=\{a \in P \mid \forall k \in T, aRk\}$，表示具有 T 中所有关键词的作者集合。以表 1 为例，设 $p_1=\{a_1, a_2\}$，则 $f(p_1)=\{k_1, k_2, k_7\}$;设 $T_1=\{k_1, k_2, k_7\}$，则 $g(T_1)=\{a_1, a_2, a_3\}$。

定义 3:形式背景(A，K，R)上的一个形式概念(formal concept)是二元组(P，T)，其中 $P \subseteq A$，$T \subseteq K$，且满足 f(P)=T 和 g(T)=P。我们称 P 是形式概念(P，T)的外延，T 是形式概念(P，T)的内涵。在定义 2 的例子中，$f(p_1)=T_1$，但 $g(T_1) \neq p_1$，所以 (p_1, T_1)不是形式概念。但若设 $p_2=\{a_1, a_2, a_3\}$，$T_2=\{k_1, k_2, k_7\}$，则 $f(p_2)=T_2$，$g(T_2)=p_2$，所以 (p_2, T_2)是形式概念。

定义 4:若 (p_1, T_1)、(p_2, T_2)是某个形式背景(A,K,R)上的两个概念，如果 $p_1 \subseteq p_2$(或 $T_2 \subseteq T_1$)，那么 (p_1, T_1)被称为 (p_2, T_2)的子概念，(p_2, T_2)被称为 (p_1, T_1)的超概念，并将其记作 $(p_1, T_1) \leqslant (p_2, T_2)$。关系≤被称为形式概念之间的偏序关系。超概念与子概念的关系是所有形式概念集合上的偏序关系。例如在表 1 中，以概念 $C_1=(\{a_2, a_3\}, \{k_1, k_2, k_7, k_8\})$和概念 $C_2=(\{a_1, a_2, a_3\}, \{k_1, k_2, k_7\})$为例，因为概念 C_1的外延 $\{a_2, a_3\} \subseteq$概念 C_2的外延 $\{a_1, a_2, a_3\}$，而同时概念 C_2的内涵 $\{k_1, k_2, k_7\} \subseteq$概念 C_1的内涵 $\{k_1, k_2, k_7, k_8\}$，所以概念 C_1是概念 C_2的子概念，概念 C_2是概念 C_1的超概念。

定义 5:按上述方式，有序的(A，K，R)所有形式概 念的集合被表示为 β(A，K，R)，并且被称

为形式背景(A，K，R)上的概念格。由于所有概念按偏序关系排列，概念格可以通过线性图可视化地呈现出来，图 1 所展示的是由表 1 的形式背景所生成的完整概念格。

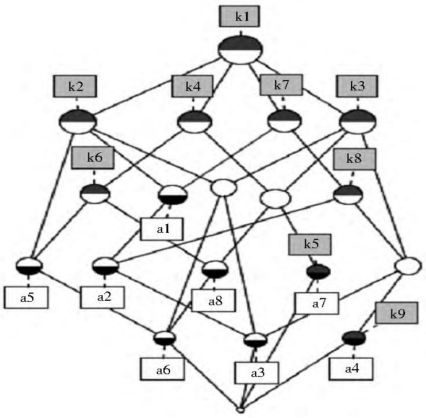

图 1　表 1 对应的概念格 Hasse 图

3 基于 FCA 的学科知识构建的构建方法

基于 FCA 的学科知识结构构建模型见图 2，该模型体现了学科知识结构的整体构建流程，其中包括:一开始对学科代表期刊的文献进行收集和预处理; 其次从预处理过的期刊文献中识别核心作者; 针对核心作者所标注的关键词，进行同义词合并以及高频词的筛选; 以作者为形式概念的对象，以作者标注的关键词为形式概念的属性，根据对象与属性的关联关系构造学科领域形式背景，进而生成概念格; 最后利用 Hasse 图展示学科知识结构。

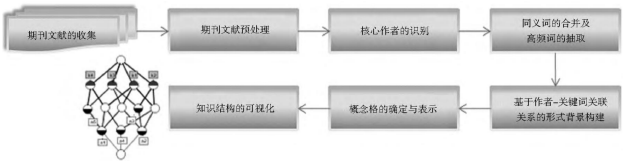

图 2　学科知识结构的构建模型

3.1 期刊文献的收集

开始选取某个数据库(如 Web of Company)作为数据来源，在检索平台中设置检索条件(如来源期刊、文献类型、入库时间等)，检索得到相应的文献; 然后选定所需的题录内容(如题名、作者、摘要、关键词、出版期刊、出版年、参考文献等)下载期刊题录数据。

3.2 期刊文献预处理

在收集到的文献数据集中，会出现同人不同名或 同名不同人的姓名著录不规范的情况，例如: Ortega．JL、Luis Ortega．Jose 和 Ortega．Jose Luis 指向的是同一位作者; Brown．C 指的是不同作者 Brown．Caroline 和 Brown．Cecelia。因而需要借助于文献中的作者姓名全称以及机构、邮箱地址等信息来判断作者确切身份信息，并进行作者名的统一。

3.3 核心作者的识别

在学科核心作者选取方法上目前还没有一个统一 的标准，主要选择指标包括作者发表文章的数量以及作者文章的被引频次。D．J．S．Price[6] 提出高产作者 发文量的阈值可以利用公式来计算: $N = 0.749*(\eta_{max})^{1/2}$，其中 η_{max} 表示发文量最多的作者的论文数。而作者文章的被引频次则是指某位学者在选定的数据集中所有文章被引用频次的总和。核心作者的选取以发文量作为主要指标，以作者被引频次作为辅助指标。

3.4 同义词的合并及高频词的抽取

在确定学科核心作者后，得到这些作者标注的关键词集合。由于不同的关键词可能表达相同的含义，因此需将同义词进行合并，还应将单复数、全称与缩写、连字符进行合并统一处理。另外，要删除不足以代表学科主题的词汇，比如地名。最后从词汇中删除出现频率低的词，保留高频词，作为关键词集合。

3.5 基于作者 －关键词关联关系的形式背景构建

通过上述两个步骤所识别的核心作者与提取的关 键词集合构成了对象集合 $A=\{a_1, a_2, \cdots, a_m\}$ 和属性 集合 $K=\{k_1, k_2, \cdots, k_n\}$，任一作者标注的关键词集合代表了该对象拥有的属性，于是通过作者关键词关联关系矩阵生成学科知识结构的形式背景，如表 2 所示:

表 2 作者关键词关联矩阵(形式背景)

	k_1	k_2	…	k_n
a_1	×		…	×
a_2		×	…	×
…	…	…		…
a_m	×	×	…	

3.6 概念格的确定与表示

概念格作为形式概念分析中核心的数据结构，本质上描述了对象和属性之间的联系，表明了概念

之间的泛化和特化关系。生成概念格的过程实质上就是概念聚类的过程，选用合适的概念格构造算法将形式背景中的对象及其属性转换成概念格中具有偏序层次的概念节点。现有的概念格构造算法大致分为两种: 批 处理构造算法和渐进式构造算法[7]。无论选用哪一种算法，同一个形式背景生成的概念格是唯一的，即概念格的构造结果不受数据排列次序的影响。

3.7 知识结构的可视化

为有效地显示概念格的结构，为读者呈现出美观、可读性强的格图，通常用 Hasse 图作为概念格的图形化表示。概念格 Hasse 图绘制为概念层次结构的萃取提供了一个通用的机制，是概念节点偏序关系的简明而有效的表示，实现了对知识结构的可视化描述。Hasse 图中每一个节点代表了知识结构的每一个概念，每个概念都由外延(学者)和内涵(关键词)组成，每条边揭示了概念之间的层次关系。Hasse 图中每个节点的外延涵盖了其下层节点中的所有对象，同时节点的内涵继承了其上层节点的所有属性。随着层次的升高，概念越泛化，其节点包含的子概念越多，就越具有概括性，也就是被越多作者所拥有的关键词越处于概念格的上层。反之，随着层次的降低，概念得到特化，越下层的概念继承的属性越多，就越具有特殊性。拥有越多关键词的作者相对就越少。最上层节点包含了所有的概念，最底层节点包含了所有的内涵。相对传统的树图，概念格 Hasse 图能更合理地展示知识结构的复杂性和交互性。

结束语

对图书情报学的学科知识结构的探测是一个良好的可使用形式概念分析的运用实例。目前常用引文分析方法、共词分析方法针对知识结构进行探测分析。本文创新性地运用形式概念分析对知识结构进行探测，形成独到的知识结构的探测方法：以人的认知为中心，以作者(集)作为概念的外延，以作者集共享的关键词作为概念的内涵，对学科的概念结构进行识别、排序和展示。根据对图书情报学的实证分析，证实基于 FCA 的知识结构探测能更好地揭示学科内不同层次的主题概念、学者以及概念和学者之间的复杂关联关系。相较于作者共被引分析、作者文献耦合分析和共词分析等传统的研究方法，FCA 在揭示学科知识结构上的优势在于: ①通过同时对作者和关键词进行形式 背景构建，可以将作者及其关键词分解到不同的子模 块间，能揭示细粒度的知识结构; ②通过分析概念格中不同部分的相似性，可以在不同的层次上研究作者研究领域间的异同; ③所构建的学科结构客观真实，减少了人为主观因素，同时避免了引用的时滞局限，在揭示学科新兴研究主题方面更具优势。本文的研究仍存一些不足: ①大量的概念使得概念格结构相对复杂，需利用粗糙集理论和概念稳定性判断等方法对概念进行约简; ②针对对象的属性选取问题，目前采用的是一旦某个作者标注了某个高频关键词，则形式背景中对应值设为 1，没有考虑作者标注该词的频率。下一步将比较多值形式背景和单值形式背景的异同，以选择更加合理的形式背景构建方式。

参 考 文 献

[1] 念闯玲．基于组织知识结构的知识缺口识别方法研究[D]．大 连: 大连理工大学，2010．

[2] Nicolai J F，Torben P． The MNC as a knowledge structure: The roles of knowledge sources and organizational instruments in MNC knowledge management［J］． Danish R esearch Unit for Industrial Dynamics，2003(5) : 1 －33．

[3] 克德罗夫．学科的组织和发展中的知识结构[J]．徐瑞方，译．国外社会科学文摘，1987(9) : 43 －45．

[4] Wille R． R estructuring lattice theory: An approach based on hierarchies of concepts[C]/ /The Ordered sets．Dordrecht-Boston: R eidel Publishing Company，1982: 445 －470．

[5] Wille R． Formal concept analysis as mathematical theory of concepts and concept hierarchies［M］/ /Formal Concept Analysis． Berlin Heidelberg: Springer-verlag，2005: 1 －33．

[6] Price D J S． The scientific foundations of science policy[J]． Nature，1965，206(4) : 233 －238．

[7] 王绍斐．概念格构造算法的研究及其在本体中的应用[D]．大 连: 大连交通大学，2010．

[8] Three-way granular computing, rough sets, and formal concept analysis[J] Yao YiYU Volume 116, January 2020, Pages 106-125