《智能信息处理》课程作业

形式概念分析的知识扩展与应用研究

刘丙富

作业	分数[20]
得分	

2020 年 11 月 1 日

形式概念分析的知识扩展与应用研究

刘丙富

摘 要 形式概念分析不仅具有良好的数学性质，可以从本质上对于对象与属性间的联系进行描述，还能够从直观上展现概念间的泛化和特化关系，因此，形式概念分析已作为一种有效的工具在数据挖掘、信息检索和软件工程等诸多领域有广泛的应用。本文主要对形式概念分析理论进行了知识扩展研究，从理论上进一步丰富了经典的形式概念分析。

关键词 形式概念分析，泛化关系，知识扩展，概念格

Research on knowledge extension and application based on formal concept analysis

Liu Bingfu

Abstract Formal concept analysis not only has good mathematical properties, it can describe the relationship between objects and attributes in essence, but also show the generalization and specialization relationship between concepts intuitively. Therefore, formal concept analysis has been widely used in data mining, information retrieval and software engineering as an effective tool. This paper mainly studies the knowledge extension of formal concept analysis theory, which enriches the classical formal concept analysis in theory.

Keywords Formal concept analysis, Generalization relation, Knowledge extension, Concept lattice

1 绪论

随着数据库技术的迅速发展，各个领域数据与日俱增，如何从海量的原始数据中获取潜在的和有价值的信息，成为人们日益关注的焦点，因此，很多知识发现方法研究也应运而生。形式概念分析理论，即概念格，因其具有良好的数学性能，不仅可W从本质上描述对象与属性间的联系，还能够直观展现概念间的泛化和特化关系，已经成为数据分析的一种有力工具。

目前，形式概念分析理论已经在数据挖掘、信息检索、软件工程、模式识别等领域有广泛的应用研究，而在处理实际应用问题时，有更多不完备或者不确定的信息存在，单纯的使用一种理论已经不能够满足解决实际应用的需求，因而如何将形式概念分析与一些软技术分析工具进行有效的融合，以适用于不同领域的应用研究，湿得尤为重要。粗糖集和模糊集都作为软技术数据分析工具，能有效地处理不精确性、不确定性和模糊知识，并且能够更加准确的描述现实中存在的应用问题。在文本中，将形式概念分析理论分别与粗糙集、模糊集理论进行有效融合，研究不同理论的异同，各取所长，并且构建两者间的联系，一方面对经典的形式概念分析理论进行了扩展研究，另一方面也拓展了形式概念分析理论应用研究的范围。

2 形式概念分析概念

形式概念分析理论的主要思想源于哲学中对概念的定义。在哲学体系中概念是由外延和内涵两个部分组成的思想单元，外延被定义为属于这个概念的所有对象的集合，而内涵被定义为属于这个概念的所有对象所共同具有的属性的集合。形式概念分析理论从概念的哲学定义中得到启发，根据数据集中对象与属性之间的二元关系来获取数据中隐含的概念及其结构。所有的概念连同他们之前的泛化与例化关

系可以构成一个概念格。概念格是形式概念分析的核心数据机构，其本质上描述了对象和特征之间的联系。一般认为外延是概念覆盖的实例，而内涵则是对于概念的描述，概念进一步可以通过 Hasse 图来实现可视化，通常 Hasse 图的每一个节点就代表一个概念。

早在 1940 年 Birkhoff 就已为该方法提供较好的数学理论基础；之后 Ganter 等人将其作为一个较好的数据分析方法，深化、完善该理论基础，并将它们扩展到各种现实应用中。形式概念分析提供了一种较好的层次化（形式）对象的分析方法，它能够识别那些具有共同（形式）属性的一组（形式）对象的组合。在应用形式概念分析方法的过程中，线路图的制定是非常重要的一个环节，其本身也是对于概念化的图形化表示。通过线路图能够对语境中所包含的对象和属性关系进行展示，在一些特定的语境下还包含有继承以及发展的关系，因此说形式概念分析其本质是一种准确性高以及使用范围广泛的分析模式。

我们已经知道概念的内涵与外延是关于概念的对象与属性的两个基本特征，但是它们同对象的属性和对象本身既有联系又有区别，。对于内涵来说，对象的各种特有属性或者本质属性都可以反映在特定的概念中而成为该概念的内涵，任何概念的内涵也都是反映特定兑现一定方面的特定属性或本质属性。但是，并非对象的特有属性或本质属性就是概念的内涵，而是只有当对虾干的特有属性或本质属性被反映到概念之中时，才转化为概念的内涵。对外延来说，任何事情都可以反映在特定概念中而成为概念的外延，概念的外延就是指适用于该概念的对象。同理，与概念的内涵一样，并非一般客观事物都是概念的外延，而是只有当客观事物被反映到概念之中成为其对象时，才转化为概念的外延。

3 形式概念分析应用领域研究

3.1 本体论领域的应用

在本体研究领域，以形式概念分析为基础构建的概念格凭借其在潜在对象和属性的探索功能、概念及概念间关系的呈现功能以及可视化方面的优势，被广泛地应用于本体构建、本体映射和本体合并等方面，促进了本体的重用和知识的共享。特别是将概念格理论引入数字图书馆知识组织体系构建，不仅拓宽了数字图书馆知识组织相关研究领域的视野，还开辟“概念格与本体的互补融合”等新的研究途径。

在形式概念分析领域内，本体（Ontology）是目前公认的共享概念模型的明确形式化规范说明，同时人工智能和知识工程领域的相关研究也越来越重视形式概念分析在知识共享和知识重用方面的优势，同时也反作用于形式概念格理论自动化构建本体算法的改进和提升。Obitko 等人认为，语法结构和语义描述对本体构造的定义都很重要，而形式语义能够实现本体推理的自动进行，使得良好语义的本体在不同领域能够更好的共享和重用，因此利用形式概念分析来探索概念格中潜在的对象和属性，从而更好的实现自动化的数据可视并表现出相比于单纯的数据分类所带来的更大优势。Formica 对现有本体的分析进行改造，将形式概念分析和概念格更好的为本体的重用进行服务。通过对概念格的构造和对形式背景下不同概念的相似性分析来实现概念的相似性推理。

3.2 软件工程领域的应用

同样，软件工程也是形式概念格分析的一个重要应用领域，也是促进形式概念格相关技术发展的一个重要组成元素。相比于其它领域，软件工程更贴近于社会科技文化生活，因此概念格与软件工程所交叉领域的发展也是迅猛而又突出。Tilley 等人在软件工程国际标准 ISO12207 的基础上对形式概念格与软件工程相结合的应用领域进行了相关应用的分析，同时对概念格的构造在软件工程编码中的应用进行了研究，其结论涉及形式规范、需求分析、组件重用、功能完善等。

其中，最具有代表性的应用研究集中在代码特征定位、类层次再购、模块结构

调整和切片影响分析这几个领域。

而在国际社会中，对于概念格在软件工程领域的应用与研究远不止于上述领域。在组件重构、故障分析与定位、设计流程检验等诸多方面均有使用。随着形式化特征的逐步加强，软件工程领域标准化趋势的逐步明显，开发人员对工程整体的认识的依赖主播减弱，因此高效准确的软件设计、开发、维护和定位系统也将变得越来越重要，在领域内的应用也将更加广泛。

3.3 知识发现领域的研究

德国著名学者 Gerd Stumme 最早将形式概念分析和概念格理论应用于知识发现领域。而在最近的１０多年内，该领域的核心研究成果均出在该团队。Stumme 最早在 1998 年就提出了数据库概念化知识(Conceptual Knowledge Discovery in Databases,CKDD)的思想，同时在概念分析的知识处理(Concptual Knowledge Processing)的基础上建立了形式概念分析的知识挖掘系统。之后 Stumme 的团队依托概念格理论开发出一套概念分析的软件：托斯卡纳系统(TOSCANA)，从而实现对大型数据库中的概念化知识进行挖掘。1999 年，Stumme 的团队又在概念格的基础上研究出了一套用于呈现知识的概念化信息系统，以关联规则为目的，并极大的简化了概念信息可视化过程的复杂计算。之后 Stumme 团队将概念格理论在知识发现领域进一步深入，于 2000 年提出了 TITANIC 算法。之后的研究论证了形式概念分析与潜在规则和关联规则挖掘的相关性，并采用概念格提高了挖掘的效率，同时 Stumme 团队针对概念格中的大量冗余规则进行分析，进一步简化了概念格的构造过程。之后，该团队在之前的研究基础之上提出了冰山概念格(Iceberg Concept Lattices)特大型分析数据库，该数据库能够在不损失信息的情况下呈现出关联规则，并对数据进行可视化。但该系统开销极大。2006 年以后，随着概念格相关算法的不断完善，形式概念分析的应用领域也在逐步扩大，目前已经延伸到了语义识别、垃圾识别和系统评估等。纵观上述算法，关联规则挖掘已经成为知识发现领域的重要内容。

在知识发现领域 ，形式概念分析与概念格不但能够提高知识挖掘的响应效率 ,还能在没有信息损失的前提下以直观的视图呈现规则 ，因此适合于大型及特大型数据库的知识发现。

3.4 Web 数据挖掘领域的研究

由于 Web 技术全面普及较晚，且 Web 领域的数据杂乱，且多为非结构化数据，因此早期的概念格理论在 Web 领域的应用都无法深入，直到 1993 年，Godin 的团队将概念格结构和传统的信息检索相结合，才打开了该领域与概念格联系。之后 Carpineto 等人通过概念格聚类算法为文本建立索引，同时混合了基于概念格的导航系统，从而提升了布尔查询的效率。经过实践检验，在概念格的支持下导航系统能够具有更高的适性和更好的检索性能，因此近 10 年来，许多专家学者开始致力于将概念格更好的应用于 Web 语义检索领域，这也使得两者成为最为直接和活跃的领域之一。

针对 Web 语义检索，Kim 团队通过形式概念格的构造来增强知识获取浏览机制，从而提升特定领域的信息检索能力。该机制能够动态升级特定领域的文档结构，同时随着时间的增加，多用户文档之间可以建立相互协作的文档浏览方案，将文档管理中的耦合性降低到最小。但特定领域的检索必定无法适应当前海量数据的 Web 领域，因此 Cigarran 团队借助概念格技术将其拓展到自由文本领域，通过自动术语抽取和权重词库的构建来自动分配描述符，帮助 Web 语义检索系统扩展到非特定的领域和词库。

概念格本身就具有概念间的层级关系，因此 Rome 团队就借助这个特性对语义知识库进行定位和组织，同时借助群落来降低 Web 知识库的复杂度。这种层级关系也可以理解为概念之间的依赖关系，因此 Messai 团队从形式概念分析中的形式背景作为出发点，通过这种依赖关系来反映属性之间的相关程度，生成能够更好驾驭

的导航信息。

4 概念格构造方法的现状

目前概念格的构造方法主要分为两类：批处理算法和渐进式算法。且目前的概念格理论都是针对二元布尔关系所提出的。

批处理算法的主要算法思想是先生成概念集合，然后再在所生成的所有概念之间构造概念之间的相关关系，而由于概念格具有层次结构，因此我们通常也将这种关系称为概念之间的父子关系。而先生成概念还是概念生成的同时将概念间的父子关系也同时生成是区别批处理算法和之后要介绍的渐进式算法的根本区别。常见的批处理算法有 Bordat 算法，Chein 算法。

渐进式算法与批处理算法的不同在于渐进式算法在生成新的概念格的同时就为其构造概念之间的父子关系，即同时生成概念和构造格结构。最典型的渐进式算法有 Godin 算法，Capineto 算法等，但这些算法大体思想类似，所有的改进只是在对于边的连接处理上。

目前随着高性能计算的发展，分布式计算日趋成熟，概念格构造和应用与并行化技术的结合也提上了日程，目前概念格在构造过程中的并行化处理技术仍然不够完善，主要是因为概念格中的概念之间具有复杂的父子关系，且这种关系受冗余节点等的影响，使得构造过程中概念之间的关系难以确定，从而难以划分子过程，使得并行划分效能无法达到理想状态

5 总结

事实上，形式概念分析与概念格理论在以上 4 个方面的应用并非是绝对孤立的，而是相互交叉、相互融通的。本体的构建、映射与合并，在本质上正是借助于形式概念分析与概念格理论在概念化知识呈现与处理方面的优势，而这一优势同时又为其在知识发现领域中的应用奠定了基础，软件工程中的代码定位、组件重构等应用一定程度上体现的就是面向软件开发与维护领域的知识重用。

随着对形式概念分析与概念格理论应用研究的不断深入，其发展前沿和研究热点也不会一成不变，本文所列的 4 个方面只是形式概念分析与概念格理论应用研究中诸多分支中的一部分，并不足以囊括国际上这一领域研究成果的全貌。在今后的研究中，仍然需要不间断跟踪和把握国际学术界的发展前沿和研究热点。

参考文献

[1]张云中,柳迪,张原铭.基于形式概念分析的知识发现研究态势[J].情报科学,2018,36(09):153-158.

[2]李想. 基于形式概念分析的知识扩展与应用研究[D].山西大学,2014.

[3]胡鑫.形式概念分析在软件工程中的应用 [J]. 电脑迷， 2016（5）：39.

[41]李海霞,聂东明,汪慧,王兴龙.概念格的一种并行构造算法[J].河南科技学院学报(自然科学版),2020,48(02):59-64.

[5]钟迪. 概念格构造算法改进[D].华南理工大学,2018.

[6]李金海,魏玲,张卓,翟岩慧,张涛,智慧来,米允龙.概念格理论与方法及其研究展望[J].模式识别与人工智能,2020,33(07):619-642.

[7]孙雨生,付荣荣,郭隆敏.国内本体研究进展：载文分析和知识基础[J].计算机与数字工程,2020,48(06):1314-1323+1378.

[8]盛秋艳，刘群，一种基于本体的叙词语义描述方法,情报科学第 25 卷第 9 期，2007 年 9 月

[9]贺晓丽，刘华丽，刘瑶瑶.多粒度数据的区间形式概念分析方法[J].计算机工程与应用，2019,55(19)，52-57.

[10]Zeinab Javidi,Reza Akbari,Omid Bushehrian. A new method based on formal concept analysis and metaheuristics to solve class responsibility assignment problem[J]. Iran Journal of Computer Science,2020(prepublish).