基于形式概念分析的机器学习应用

高帅

作业	分数
得分	

2020 年 11 月 13 日

基于形式概念分析的机器学习应用

高帅

（大连海事大学 信息科学技术学院，辽宁 大连 116026）

摘 要

随着科学技术的进步，现代信息检索所处理的对象和规模都有了很大的变化。人们迫切需要一种能够快速而且准确地查找访问图像的技术，这就是图像检索技术。图像检索的过程反映的是从图像数据中抽取出特征，形成概念，来研究这些概念之间的关系。同时，形式概念分析理论也提供了一种以概念格让数据有机地组织起来的形式，概念格节点体现了概念内涵和外延的统一，非常适合于发现知识。因此在许 多领域获得了广泛的应用，如信息检索、数字图书馆、软件工程、知识发现等。本文系统研究了国内外图像检索的现状以及概念格在相关领域的应用 ，提出了一种基于形式概念分析的图像检索方法。

关键词 形式概念分析 概念匹配 相似度 隶属度函数 图像检索

中图法分类号 TP311 **文献标识码** A

Image retrieval based on formal concept analysis

Gao Shuai

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026，China)

Abstract With the development of the science technology,the object and scale whcih the modern information retrieval deals with have ahuge change.Therefore,people need a kind of technique urgently that can seek an interview image quickly and accurately.This is the image retrieval technique.The process of image retrieval reflects the relationship of concepts which are formed by the feature extracted from the image data.

Meanwhile,the Formal Concept Analysis theory provides one kind of model to organize the data by the concept content and extent,suited very much to discover knowledge.Therefore index,the numerical library,software engineering,knowledge discover and etc...

Keywords Formal concept analysis,Concept lattice;Similarity degree;Membership degree;Image

1. 引言

概念格(Concept Lattice)是形式概念分析理论(FCA)中的一种数据结构 。形式概念分析理论是由德国的数学家 Wille 于 1982 年年提出来的。形式 概念分析中的“形式 ”一词表示我们正在处理领域的工作，通过与这些工作相联系的结构化的概念的联系，发现可理解 的，有意义的知识 “概念 ”是对哲学中概念的一种数学表示，是对人们认知的知识的一种数学化描述 。在形式概念分析中，概念的外延被理解为属于这个概念的所有对象的集合，而内涵则被认为是所有这些对象所共有的特征或属性集，这实现了对概念的哲学理解的形式化 。所有的概念连同它们之间的泛化与例化关系构成一个概念格。概念格的数学基础是序论和格论。概念格结构模型是形式概念分析理论中的核心数据结构，它本质上描述了对象和特征之间的联系表明了概念之间的泛化与例化关系。它是一种用于识别数据集中概念结构的数据分析理论，用来研究特定领域的概念以及概念之间的关系，概念和概念之间的关系通过概念格表示出来，概念格对应的图则实现了对数据的可视化。概念格自

从提出以来，其内在的优势得到越来越多的科研工作者的注意，并迅速在多个领域得到发展。概念格以其独特的优势引起越来越多的研究人员的关注，并在许多领域获得了广泛的应用，如信息检索、数字图书馆、软件工程、知识发现等。然而，概念格的研究仍是一个极其年轻的领域，有很大的发展前景和应用潜力 。在过去的十年里，FCA 的应用领域发生了巨大的变化，其主要应用领域逐步由数学转向计算机科学。国内外的研究主要有基于概念格的分类系统、关联规则和聚类。由于形式概念分析以概念格的形式使数据有机地组织起来，概念格节点体现了概念内涵和外延的统一，因此非常适合于用来发现知识。概念格从关系数据中构造出来，然后从它提取各类型的知识，如关联规则、分类规则等。

2. 形式概念分析

形式概念分析[1] 是应用数学和格论的一个分支,它建立在概念和概念层次的数学化基础之上。一个概念就是最大限度地收集对集合中共同特点有帮助的元素，并且运用形式概念分析的方法，可以发现、构造和展示由属性(Attributes)和对象(Objects)构成的概念(Concept)及其之间的关系。因而，形式概念分析的方法已经运用在软件开发等众多环节之中。

2 .1 相关概念及定义

定义 1 :一个形式化的语境(Context)k =(G,M I)，包括两个稽核(G 和 M)和一个二元关系。在这个语境中，G 中的元素称为对象，M 中的元素称为属性。一般用 gIm，或者(g，m)∈ I 来表达对象 g 和属性 m 的关系，读作“对象 g 具有属性 m ”。根据定义 1 ,可以通过矩阵来表示语境。每行的开头是对象名,每列的开头是属性名。行 g 和列 m 的交叉表示对象 g 具有属性 m。

定义 2:对一个对象集 A ,定义 A’ ={m ∈ M |gIm ,对所有的 g ∈ A}(即 A 中所有的对象共有的属性集合)。相应地，对一个属性集 B ，定义 B’ ={g ∈ G |gIm ，对所有的 m ∈ B}(即包含 B 中所有的属性的集合)。

定义 3 :语境(G,M,I)中的形式概念(Formal Concept)是个集合对(A,B)，其中 A G ，B M ,并且 A’=B ，B’ = A 。A ,B 分别称作概念(A,B)的外延(Extent)和内涵(Intent)。β (G,M,I)表示语境(G,M,I)中的所有概念集。

定义 4 :如果(A1,B1),(A2,B2)都是语境中的概念，并且 A1 属于 A2,那么(A1,B1)被称作(A2,B2)的子概念(Sub concept),(A2,B2)则是(A1,B1)的超概念(Super concept)，记为(A1,B1)≤(A2,B2)。“ ≤”反映了概念间的层次关系。由层次关系搭构的所有(G, M , I)的概念记作 β (G, M,I)，被叫作概念格[2](Concept Lattice)。

2.2 概念格的构造算法

在应用概念格的过程中涉及到概念格的构造，概念格的构造效率是一个很重要的问题，所以概念格的构造算法也成为了一个重要的研究课题。建格的过程实际上是概念聚类的过程，是一个从低层概念进行综合从而得到高层概念的过程，这个过程体现了从数据中提取出隐含的概念的过程。因此，在概念格中，建格的算法具有很重要的地位。对于同一批数据，所生成的格是唯一的，即不受数据或属性排列次序的影响，这也是概念格的优点之一。概念格的建造算法可分为两类：批处理算法和增量算法。概念格可以添加背景知识，这些知识以 if...then 的规则 形式出现。概念格甚至可以只用形式背景知识建造。

概念格的典型构造算法主要分两大类：批处理算法和渐进式构造算法。

2.1.1 批处理算法

批处理算法(Batch Algorithm)构造概念格要完成两项任务:一是要生成所有的格节点，即所有概念的集合;二是要建立这些格节点间的直接前驱/直接后继关系。 根据其构造格的不同方式，可将批处理算法分

为三类:自顶向下算法、自底向上算 法及枚举算法。

自顶向下算法首先构造格的最上层节点，再逐渐往下生成概念节点。自底向’ 上算法则相反，首先构造底部的节点，再向上扩展。枚举算法则是按照一定顺序 枚举格的所有节点，然后再生成各节点之间的关系，此算法有 Ganter1141 的算法等。

批处理算法的操作相对于渐进式算法较简单，因为它不要进行边的修改，但 缺点是不能实现增量操作，当加入新对象时，需要重新对所有对象进行操作来生 成新的概念格。

对于形式背景 K = (G,MJ),其概念格的批生成算法的一般框架如下所示：

概念格的批生成算法：

1. 初始化格 Z={(G, XG)));
2. 队列 F={(G, XG))};
3. 对于队列F中的一个概念C,产生出它的每个子概念 G;
4. 如果某个子概念 G 以前没有产生过，则加入到▲中；
5. 增加概念 C 和其子概念 G 的链接关系；
6. 迭代⑶-(5),直至队列尸为空；
7. 输出概念格 Z。

2.1.2 渐进式算法

为了解决数据的集中式存储和算法串行之间的矛盾，研究者又纷纷提出了概 念格的并行构造算法。渐进式构造算法是比较有应用价值的一类构造算法。国内 研究者后来也相应地提出了一些改进算法以提高概念格的构造效率，例如基于属 性的概念格渐进式生成算法网，利用数据库技术对概念格构造算法的改进，基于 剪枝的概念格渐进式构造等。

渐进式算法(Incremental Algorithm)也叫增量算法，其基本思想是将当前要插.入的对象与格中的概念求交，根据交的结果进行不同的操作。与批处理算法不同 的是，渐进式算法主要是通过利用新增对象对原始概念格的进行更新操作来完成 新概念格的生成。这种更新操作体现了两种变化:概念的变化和边的变化。因此， 渐进式算法生成概念格的过程主要解决的问题有新概念的生成和边的更新。

渐进式算法是指在给定原始形式背景 K = (G,MJ)所对应的原始概念格 L 以及 新增对象¥的情况下，求解形式背景 K’ = (G<jx’,M,I)所对应的概念格〃。

由此可见，它主要是针对属性集 M 不变的情况下，完成对原始概念格的更新 操作，可以将这种渐进式算法称为基于对象的渐进式算法。典型的算法有 Godin[3] 的算法。实际上，形式背景的变化包括两个方面:对象的变化和属性的变化。基于 属性的渐进式算法通过不断地渐增属性来构造概念格。当然，还可以对新对象和 新属性同时增加的情况进行进一步研究。在实际问题中，应根据问题的具体情况 来选择采用何种方式的渐进式算法。

渐进式算法是比较有应用前景的一类概念格构造算法，由于它的增量操作使 得当加入新对象时不需要重新生成概念格，只需在原有概念格的基础上进行更新， 所以能很好的解决实际问题中所遇到的增量问题。

对于形式背景 K = (G,M,I),其概念格的渐进式生成算法的框架如下所示：

概念格的渐进式生成算法：

(1) 初始化格 Z 为一个空格；
(2) 从 G 中取一个对象 g;
(3) 对于格 L 中的每个概念弓=(4，耳)，如果 BQ(g),则把 g 并到也中；
(4) 如果同时满足：和不存在(4,4)的某个父节点“2,82)满足 BE(g),则要产生一个新节点；
(5) 对新产生的节点加入到 Z 中，同时调整节点之间的链接关系；
(6) 迭代(2)-(5),直至形式背景中的对象处理结束；
(7) 输出概念格 Z。

3. 概念格的应用

作为数据分析和知识处理的形式化工具，形式概念分析已经获得了广泛而成 功的应用。在软件工程领域，形式概念分析为再工程、软件重用、面向对象程序 设计等领域中某些问题的解决提供了理论支持，并已经取得了一系列的应用成果。 在数据挖掘领域，由于形式概念分析以概念格的形式使数据有机地组织起来，概 念格节点体现了概念内涵和外延的统一，因此非常适合于用来发现规则型知识。 除了在数据挖掘和软件工程领域获得的研究成果外，概念格还被成功地应用于信 息检索、知识库组织等诸多领域[4]。

3.1.1 概念格在软件工程中的应用

在面向对象的程序设计过程中，类的层

次结构对于理解和分析类库是非常重 要的。Arfi 等[5]描述了一个用于生成 Smalltalk-80 类层次的工具，该工具从 Smalltalk 代码中直接提取类的接口信息，并生成类层次，同时提供一个简单的图 形用户界面用于对类层次进行浏览。Godin 等 Ml 开发了一个原型工具用于从类的 规范说明中计算类层次，它通过逐个地插入新类来生成概念格，最终生成的类层次可以通过图形浏览器来进行交互式研究和提炼。为了更好地评估生成的类层次 质量，该工具还计算了类层次的若干个面向对象的度量指标。在类库的开发过程 中类层次的设计往往存在着某些缺陷 oSnelting 等 卩句基于概念分析提出了一个用于 检测和补救这些设计问题的框架结构。它通过分析一个类库以及一系列使用它的 应用程序来构建出格结构，从而提供了关于类层次使用情况的有用信息。在此基 础上说明了如何根据格结构生成重构的类层次，以及格结构如何用作对类层次进 行重新设计和重新组织。

在软件再工程过程中，对旧代码进行模块化、从遗产代码中识别出对象及配 置的再工程是其中一些重要的课题。Lindig 等[6]通过分析过程和全局变量之间的 关系构造出概念格，并从格结构得到模块结构，进一步使用格结构评估模块候选 项之间的内聚度和耦合度。Deursen 等〔浏通过重组遗产数据结构来识别对象，他们 将聚类分析和概念分析技术用于对象识别，并实际应用于一个十万代码行的 COBOL 系统中。Snelting 等同采用概念分析技术对现有源代码进行分析，并生成 概念格，这样可清晰地显示出可能存在的配置结构和性质，不仅可以展示出配置 线之间细致的相关性，而且可以形象化地显示出配置结构的整体质量。概念格也 被用于软件重用。文[7]利用概念形成方法以两种方式来支持软件重用：建立一个导航空间，即概念层次，来对库中的产品进行组织和检索；以及通过建议更好的 抽象为重用打包活动(Reuse Packaging Activity)提供支持。文[8]则针对可重用软件 构件的检索，允许用户渐进式地用一系列的关键字来检索用户所需要的构件。每个步骤之后被选择的构件以及用于进一步精炼该查询的关键字集合被提供给用户，从而确保至少可以找到一个构件且用户不能指定互相冲突的关键字。试验结 果也表明该方法可以使用户快速而精确地选择出所需要的构件。

3.1.2 概念格在数据挖掘中的应用

已知的一些基于概念格的分类学习系统有：Sahami 网开发的 RULEARNER 系 统，它首先根据条件属性构造出概念格，然后从格中提取出分类规则用于支持对 象的分类。而 Njiwoua 等[9]设计的 LEGAL-E 系统使用学习参数来生成半格，从而 有效地控制了概念格的空间复杂性，然后采用投票的方式对新对象的分类进行群 体决策；通过将特征选择方法应用于 LEGAL-E 系统，又设计得到了 LEGAL-F 分类系统。将简单的基本的分类器(朴素贝叶斯或最近邻)与概念格中的每个节点结合，从而形成新的合成分类器。开发出两个分类系统 CLNB 和 CLNN, 其中 CLNB 是将朴素贝叶斯分类与概念格节点结合，CLNN 是将最近邻分类与概念格节点结合。实验结果表明这两个合成分类器在很大程度上提高了相应基本分 类器的分类准确率 0 Godin 等 [10] 描述了基于概念格模型的概念形成方法，主要提 出了从概念格中提取出蕴涵规则的算法，并使用了关系数据库中函数依赖的理论 结果来处理规则的蕴含问题，但是这种蕴涵规则是一种确定性规则，不具备描述 概率规则的能力和抗噪音能力。为了提高规则发现的鲁棒性，Missaoui 等 SI 对它 进行了扩展，提出了从概念格中提取近似规则(又称为概率蕴涵规则)的算法。 Pasquier 等[岡研究了关联规则的提取问题，他们的工作是以已经发现了的所有频 繁项集作为基础的，文中提出了用于提取确定性关联规则的 Duquenne-Guigue 基, 以及用于近似关联规则的适当基(Proper Basis)和结构基(Structural Basis)。

Ho 等研究了基于概念格的概念聚类方法，并实现了一些学习系统，包括 OSHAM[11] 和 INCOSHAM 网，其中 INCOSHAM 在 OSHAM 系统基础上增加了渐 进式学习的能力。OSHAM 系统将三种关于概念的不同观点，即经典的、原型的和 示例的观点组合起来，并使用了一种灵活的匹配过程，组合了逻辑的门限的和最 近邻的条件，允许系统改进它自身的推理性能，从而可以对未知的实例进行灵活 的预测。Wille 在文 [12] 中提出，在

某些情况下仅考虑一个形式背景是不够的，并 引入背景网络和多背景(Multi-context)的形式方法。文中还给出了不同形式背景之间的四种操作：并置(Apposition)、下置(Subposition)、融合(Fusion)和级连(Concatenation)o 以多背景为理论依据，文[13]以不同形式背景的对象集和特征集之间的关系为基础，研究了在概念格框架结构下对结构化(复杂)对象所进行的 概念学习和规则提取，文中还对学习所得到的概念和规则进行了解释。Burusco 研究了 L-模糊概念集的格结构，从而推广了形式概念理论。同时给出了计算这 种格结构的方法，并将它应用于一个简单的例子。文[14]则研究了和蕴含算子相关 联的模糊概念，以及它们所形成的格结构。文中还说明[15]中的模糊概念格对应于 模糊 Kleene-Dienes 蕴含，Wille 的形式概念是本文的一个特例。文[16]研究了模糊 概念格中的相似性关系，定义了三个不同层次上的相似性：对象相似性(和属性 相似性)、概念的相似性和概念格的相似性。文中还给出了一种根据概念相似性来 对概念格进行简化的方法。此外，Girard 等网研究了由模糊量词所描述的数据的 概念格构造问题。Oosthuizen 等非形式化地讨论了粗糙集合与概念格之间的联 系，并介绍了其开发的 GRAND 系统和 DATA-MAP 系统。Kent 等阚探讨了在概 念格模型上几种可能的偏大近似与偏小近似算子的定义及其相互关系。基本思想 是：给定近似空间，采用两个形式背景来对它进行近似表示，从这两个形式背景可以分别构造出一个概念格并进行知识提取。

3.1.3 概念格在信息检索领域的应用

概念格还被应用于信息检索方面。Godin 等 Ml 对使用概念格结构的信息检索进行了实验，并和两种较为传统的检索方法：在手工建立层次分类系统中导航和 使用索引项的布尔查询，做了比较实验。结果表明，在布尔查询和概念格检索方 法之间并没有显著的性能差异；然而层次分类系统检索的查全率要明显低于其它 两种方法。因此得出结论，基于概念格结构的检索是非常有吸引力的，因为它将 主题搜索的良好性能和浏览的潜力结合在了一起。对基于概念格的 文本数据库的自动组织和混合导航进行了研究，设计了一个检索系统 ULYSSES。

它首先建立数据的格结构，即先对文本进行索引，再对索引后的文本进行聚. 类。格结构为系统的导航阶段提供支持。文章最后将基于格结构的信息检索与传 统的布尔查询在两个数据集上进行了比较试验，结果表明格检索的性能优于布尔 检索。概念格在其它方面也得到了广泛的应用 0 Richards 等利用概念格对 Ripplewn Rule 进行有机的组织;Cole 的 CEM 电子邮件管理系统 卩将 Email 存储在概念格 中，而不是常用的树状结构中，使得在检索电子邮件时获得了更大的灵活；文[17] 则将概念格应用于智能帮助系统的领域建模。

概念格还被应用在图像检索领域。Imagesleth1711 实现了使用 FCA 方法在图像 中进行信息检索的功能。该系统使用的是 Thumbnail 图像，每一个都是相同大小，并且按照不同的特征存放起来。将图像作为对象，图像所对应的注解作为属性， 这些注解可以是手工输入，也可以是根据图像信息自动提取。并提出了半格的概 念，在概念格中，不同的路径可以通向一个共同的结果，因此用户可以通过不断 输入查询图像的特征来一步步得到想要的搜索结果

参考文献

[1] Vailaya A, Zhong Y, Jaing A K. A hierarchical system for efficient image retrieval. In: Proc.of the 13th Int Conf on Pattern Recognition. Washington: IEEE Computer Society, 1996, 356-360

[2] Rui Y, Huang T S, Chang S F. Image retrieval: current techniques, promising directions, and open issues. Journal of Visual Communication and Image Representation, 1999, 10(4): 39-62

[3] 齐红.基于形式概念分析的知识发现方法研究[博士学位论文]. 吉林:吉林大学. 2005, 16-18

[4] Arfi A, Godin R, Mili H, Mineau G, Missaoui R. Generating the

interface hierarchy of a class library. In Alagar V S and Missaoui R(Eds.), Technology of Object-Oriented Databases and Software Systems, World Scientific Publishing Company, Singapore, 1995, 42-57

[5] Snelting G, Tip T. Reengineering class hierarchies using concept analysis. In: Proceedings of the 6th ACM SIGSOFT Symposium on the Foundations of Software Engineering, 1998, 99-110

[6] Lindig C, Snelting G. Assessing modular structure of legacy code based on mathematical concept analysis. In: Proceedings of the International Conference on Software Engineering, Boston, USA, 1997, 349-359

[7] Deursen A, and Kuipers T. Identifying objects using cluster and concept analysis. In: Proceedings of the 21st International Conference on Software Engineering (ICSE'99), Los Angeles, California, 1999, 246-255.

[8] Snelting G. Reengineering of configurations based on mathematical concept analysis. Technical Report TR-95-02, Software Technology Department, Technical University of Braunschweig

[9] Godin R, Mineau G, Missaoui R, St-Germain M and Faraj N. Applying concept formation methods to software reuse. International Journal of Knowledge Engineering and Software Engineering, 5(1):119-142

[10]Lindig C. Concept-based component retrieval. In Proceedings of IJCAI-95 Workshop on Formal Approaches to the Reuse of Plans, Proofs, and Programs, Montreal, 21-25

[11]Sahami M. Learning classification rules using lattices. In: Proceedings of the 8th European Conference on Machine Learning, Berlin, Germany, 1995, 343-346

[12]Njiwoua P, Mephu N, Guifo E. Back from experimentation: a study of learning bias in LEGAL-E. In: Proceedings of Benelux Conference on Machine Learning (BENELEARN), Maastricht, 57-68

[13]Xie Z, Hsu W, Liu Z, and Lee M. Concept lattice based composite classifiers for high predictability. Journal of Experimental and Theoretical Artificial Intelligence, 14:143-156

[14]Missaoui R, Godin R. Extracting exact and approximate rules from databases. In Alagar V S, Bergler S, Dong F Q (Eds), Incompleteness and Uncertainty in Information Systems.London, Springer-Verlag, 209-222

[15]Pasquier N, Bastide Y, Taouil R, and Lakhal L. Closed set based discovery of small covers for association rules. In: Proceedings of BDA Confernece, 361-381 .

[16]Ho T B. Incremental conceptual clustering in the framework of Galois lattice. In: Lu H, Motoda H and Liu H, KDD: Techniques and Applications. World Scientific, 49-64