《智能信息处理》课程作业

概念格在阅读推荐中的应用研究

吴继萌

作业	分数[20]
得分	

2020 年 11 月 12 日

概念格在阅读推荐中的应用研究

吴继萌

摘 要：概念格是一种广泛应用在机器学习、软件工程 和信息检索等领域的形式化分析工具，具有精确性和完备性等特点。文章介绍了形式概念分析和概念格的基本概念并对概念格在阅读推荐的应用进行了研究，对用户阅读偏好进行形式概念分析，给出了从概念到形式概念，从背景到形式背景，从形式背景到约简形式背景，再然后形成单值形式背景，绘制 Hasse 图，最后构造出概念格的全过程。

关键词：概念格，形式概念分析，单值形式背景，Hasse 图

Research on the Application of Concept Lattice in Reading Recommendation

Wu Jimeng

Abstract: Concept lattice is a formal analysis tool widely used in the fields of machine learning, software engineering and information retrieval. It has the characteristics of accuracy and completeness. The article introduces the basic concepts of formal concept analysis and concept lattices, and studies the application of concept lattices in reading recommendation. It conducts formal conceptual analysis of users' reading preferences, and presents concepts from concept to formal concept, from background to formal background, from From the formal background to the reduced formal background, and then form the single-value formal background, draw the Hasse diagram, and finally construct the whole process of the concept lattice.
Keywords: concept lattice, formal concept analysis, single value formal background, Hasse diagram

1 引言

怎样从海量的信息中检索用户想要的结果成为许多学者热切研究的问题。传统的检索方法可以分为布尔逻辑检索和分类检索，布尔逻辑检索适用于用户明确自己查找的内容并能用计算机语言准确描述。但是大多数情况下，用户并不十分确认自己查找的内容，导致普通用户使用布尔逻辑检索比较困难。分类检索适用于用户熟悉分类信息，能够根据分类名称检索到目标结果。但是对某些在分类上有交叉的信息，将不能进行有效的筛选。考虑到传统检索方法存在的问题，有学者提出基于本体的信息检索方法，这种方法虽然在一定程度上克服了概念信息检索不能对概念关系进行处理基于概念格的信息检索及其树形可视化的局限性，但是因为缺乏构建领域本体的成熟流程、方法和标准规范，使得研究本体构建的技术体系与研究检索的技术体系之间存在脱节问题。随着形式概念分析理论的日益完善，有学者提出基于概念格的信息检索。

为了更好的为用户筛选符合其心意的书籍或文章，本文选择使用概念格来进行信息检索和筛选，使得推荐的书籍更加准确，更加符合读者的阅读口味。

2 形式概念分析

2.1 形式概念分析

形式概念分析（Formal Concept Analysis，FCA）理论由德国数学家 Wille 教授在 1982 年的论文中提出，在 1999 年 Ganter 出版的学术著作 Formal Concept Analysis:Mathematical Foundations 中，对形式概念理论的早期成果做了总结。FCA 技术通过数据集中对象和属性之间的二元关系建立概念层次结构，可以描述对象、属性和概念之间蕴含的关系，使其成为一种重要的知识表示方法。FCA 使用概念格结构表示概念之间的关系，这种结构生动直观，方便用户进行信息检索。

现实世界是由各种各样的对象组成的，每个对象都有自己的一组属性或者特征。概念就是指对象、属性以及它们之间的关系，是反映对象的特有属性的思维方式，是从对象的属性中抽出特有属性概括而成。分为两部分：一部分是对象，一部分是属

性集。因此，概念也可以表示为（对象，属性集）的二元组形式，则形式概念可以表示为（形式对象集，形式属性集）的二元组形式。

定义：设形式对象集U，$X \subseteq U$，形式属性集A，B $\subseteq A$，二元关系 R$\subseteq U \times$ A。若 X={x|x$\in U$, $\forall a \in B$，xRa}，B={$a|a \in A$，$\forall x \in X, xRa$}。则二元组（X,B）被称为形式概念。

2.2 形式背景

背景是概念的集合，也就是对象集合及其具有的属性的集合。任何一个概念都是从背景中 提取出来的一个子集，通常以对象-属性集的二维表表示一个背景，用 1 表示某个对象具有某个属性，而用 0 表示某个 对象不具有某个属性。形式概念分析是做为一种数学理论被提出的，是人们组织和分析数据的一种方法，将数据及其结构、本质以及依赖关系进行形象化的一种描述。那么，对现实世界中的概念和背景在形式概念分析时就会形成形式概念和形式背景。

形势背景需要从数据源中提取，并不是直接存在的，即对现有概念中对象和属性进行约简。将具有一样属性的对象进行合并成为一个形式对象叫做对象的约简，将所有对应于同一个对象集的几个属性合并为一个形式属性被称作属性的约简。不能约简的对象和属性会转换为相应的形式对象和形式属性。将经过对象和属性约简后得到的形式概念与形式属性以形式对象——形式属性集的形式置于二维表中，得到的就是形式背景，如表 1 所示。

表 1 形式背景范例

	1	2	3
a	0	1	0
b	0	0	1
c	1	0	1

2.3 概念格

概念格作为近来引起广泛关注的一种数据分析和知识处理的形式化工具,目前已经在机器学习, 信息检索, 数字图书馆, 软件工程, 知识分类和数据挖掘等领域显示出一定的应用价值。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。随着计算机和数据库技术的发展以及各种电子设备的大量使用, 人类收集数据的能力得到了极大的增强, 数据信息日益膨胀, 但是堆积如山的积累数据对于人类是难以处理的, 真正有价值的是埋藏于数据中的知识, 因此数据挖掘技术已经得到了广泛的研究 。而形式概念分析以概念格的形式使数据有机的组织起来, 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。已有不少学者讨论了从概念格上提取规则或函数依赖的问题。

定义：假设给定形式背景(context)为三元组 T=(O，D，R)，其中 O 是事例集合，D 是描述符（属性）集合，R 是 O 和 D 之间的一个二元关系，则存在唯一的一个偏序集合与之对应，并且这个偏序集合产生一种格结构，这种由背景(O，D，R)所诱导的格 L 就称为一个概念格。格 L 中的每个节点是一个序偶（称为概念），记为(X，Y)，其中 X∈P(O)称为概念的外延；Y∈P(D)称为概念的内涵。

每一个序偶关于关系 R 是完备的，即有性质：

1）X= $\alpha(Y)$={x$\in O|\forall y \in Y$，xRy}

2）Y= $\beta(X)$={$y \in D|\forall x \in X, xRy$}

在概念格节点间能够建立起一种偏序关系。给定 H_1= (X_1，Y_1)和 H_2= (X_2，Y_2),则，领先次序意味着 H_1 是 H_2 的父节点或称直接泛化。根据偏序关系可生成格的 Hasse 图：如果H_1<H_2,并且不存在另一个元素H_3 使得 H_1<$H_3 < H_2$，则从 H_1，到 H_2就存在一条边。

表 2 表示出了一个形式背景。其中 O={1, 2, 3}，D={a_1，a_2，a_3，b_1，b_2，b_3，c_1，c_2，c_3，d_1，d_2，d_3 }，R 描述了 O 中元素拥有的 D 中的属性值集。

表 2 概念格范例

	A	B	C	D
1	a_1	b_1	c_1	d_1
2	a_2	b_2	c_2	d_2
3	a_3	b_3	c_3	d_3

2.4 单值形式背景

为了方便查看和分析，可以将多值形式背景，例如{0，1}二值形势背景，转换为单值形式背景，用“•”来代替“1”即可。而为了构造对应的概念格，需要把单值形式背景转换为带有父子继承关系的单值形式背景，以属性的个数多少来排序，个数少的为父节点，相对的属性个数多的为子节点。由表 1 生成的单值形式背景如下：

表 3 单指形式背景范例

	1	2	3
a	0	1	0
b	0	0	1
c	1	0	1

3 阅读推荐形式背景

根据用户通常阅读的书籍类型，来刻画感知用户的阅读偏好，从而为用户推荐按更符合用户口味的书籍。以用户名称作为对象，以用户所选择的书籍中的标签作为属性集，以此建立形式背景，如表 3 所示，其中对象集{id1,id2,id3,id4,id5,….,id9}，属性集为{现代，古代，甜文，虐文，校园，都市，穿越，重生}。其中概念可以形式化为序偶（用户，标签）二元组，形式背景表现为（用户，标签，选择关系）的三元组，用 1 表示该用户选择阅读的书籍中有该标签，而 0 表示该用户选择阅读的书籍中没有该标签，形式背景如表 4 所示：

表 4 阅读标签形式背景

用户名	现代	古代	甜文	虐文	校园	都市	穿越	重生
id1	1	0	1	0	1	0	0	0
id2	1	0	0	0	0	0	0	0
id3	1	0	0	1	1	0	0	0
id4	1	0	0	1	1	1	0	0
id5	1	0	1	0	0	0	0	0
id6	1	0	0	1	0	0	0	0
id7	0	1	1	1	0	0	0	0
id8	0	1	1	1	0	0	0	0
id9	1	0	0	0	0	0	0	0

对所示的形式背景对所示的形式背景进行按照概念格的生成步骤对其进行形式背景约简。把属性值相同的对象进行约简，将相同的属性进行约简。通过观察将属性穿越和重生合并，id2 和 id9 选择阅读的书籍中包含的标签一致，可以合并；id7 和 id8 选择阅读的书籍中包含的标签一致，也可以合并。所以最后的形式对象集合为{id1，id2&id9，id3，id4，id5，id6，id7&id8}，而最后得到的形式属性集合为{现代，古代，甜文，虐文，校园，都市，穿越重生}，约简后得到的形式背景为表 5 所示：

表 5 约简后的阅读标签形式背景

用户名	现代	古代	甜文	虐文	校园	都市	穿越重生
id1	1	0	1	0	1	0	0
id2&id9	1	0	0	0	0	0	0
id3	1	0	0	1	1	0	0
id4	1	0	0	1	1	1	0
id5	1	0	1	0	0	0	0
id6	1	0	0	1	0	0	0
id7&id8	0	1	1	1	0	0	0

4 阅读推荐概念格

4.1 Hasse 图

Hasse 图中的每个结点表示集合 U 中的一个元素，结点的位置按它们在偏序中的次序从底向上排列。即对任意 a,b∈U，若 a≤b 且 a≠b,则 a 排在 b 的下边。如果 a≤b 且 a≠b，且不存在 c∈U 满足 a≤c 且 c≤b，则在 a 和 b 之间连一条线。这样画出的图叫哈斯图，又称偏序集合。Hasse 图的作图法是以“圆圈”表示元素；若 x≤y,则 y 画在 x 的上层；若 y 覆盖 x,则连线；不可比的元素可画在同一层。Hasse 图的节点就是对象，省略了自反，省略了箭头，指向朝下，由上到下表示的即为两节点间的父子关系：

表 6 单值形式背景

	a	b	c	d	e	f	g
1	•		•		•		
2&9	•						
3	•			•	•		
4	•			•	•	•	
5	•		•				
6	•			•			
7&8		•	•	•			

表 7 带有父子关系的单值形式背景

	a	b	c	d	e	f	g
2&9	•						
5	•		•				
6	•			•			
3	•			•	•		
4	•			•		•	
7&8	•	•	•				
1	•		•		•		

4.2 阅读推荐概念格

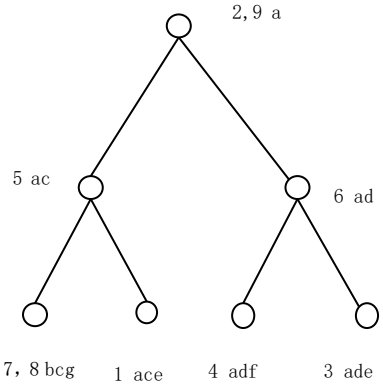

图 1 单值形式背景的 Hasse 图

概念格就是基于父子关系的进一步构建，也就是需要补全形式概念的上下确界，使得每一个子集的上下确界都存在。

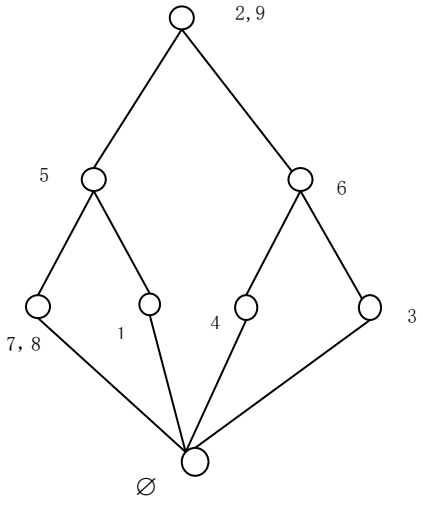

图 2 概念格

5 小结

文章介绍了形式概念分析和概念格的基本概念并对概念格在阅读推荐的应用进行了研究，对用户阅读偏好进行形式概念分析，给出了从概念到形式概念，从背景到形式背景，从形式背景到约简形式背景，再然后形成单值形式背景，绘制 Hasse 图，最后构造出概念格的全过程。

参考文献

[1]徐海玲,张海涛,张枭慧,魏明珠.基于概念格的高校图书馆群体用户兴趣画像研究[J].情报科学,2019,37(09):153-158+176.

[2]基于形式概念分析的微博兴趣分析

[3]张伟. 社交网络中基于形式概念分析的用户推荐[D].西华大学,2015.

[4]姜传菊.概念格在数字图书馆中的应用研究[J].情报科学,2010,28(12):1908-1911.

[5]沈夏炯,叶曼曼,甘甜,韩道军.基于概念格的信息检索及其树形可视化[J].计算机工程与应用,2017,53(03):95-99.