基于电影类型的形式概念分析

史程

(大连海事大学 信息科学技术学院 大连 中国 116026)

摘 要 形式概念分析是一种在数据分析和规则提取方面非常有效的工具，它可以对集合中具有某种关系或共同属性的元素进行分类，发现概念与概念之间的关系，以数学化的方式表达概念和概念层次关系。本文简要介绍了形式概念分析的基本概念，并以电影类型为例，给出了从概念得到形式概念、背景转换为形式背景、形式背景转化为单值形式背景再构造概念格的整体过程。

关键词 形式概念分析；概念格；形式背景
中图法分类号 TP18 DOI 号：10.3724/SP.J.1016.2014.01229

Movie Type Based On Formal Concept Analysis

SHI Cheng

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

Abstract Formal concept analysis(FCA) is a tool regarding data analysis and rule extraction, which can classify the elements with relationship or Some common attributes in a set, and find relationship between concept and concept constituted by attributes and object, then express it in a mathematical way. This paper briefly introduces the basic concept of formal concept analysis, and take the movie type as an example, giving the whole process of extracting formal concept from concept, converting context to formal context, transforming formal context into a single valued formal context, then forming concept lattice.

Key words formal concept analysis; concept lattice; formal context

1 引言

形式概念分析(Formal Concept Analysis，FCA)是应用数学的一个分支，它是一种知识处理的方法。形式概念分析是由德国的 R.Wille 教授于 1982 年提出的，它是一种从形式背景进行数据分析和规则提取的强有力的工具，它能够帮助我们认识集合中元素之间的关系并构建概念格，用数学的方法表达概念和概念层次。对形式概 念分析的研究主要集 中在对概念格的构造和概念格的应用两个方面。概念格可以通过形式背景进行构造，利用概念格提取关联规则是概念格应用的最成功的一个方面。本文对形式概念、形式背景、关联规则提取、单值形式背景、父子关系的单值形式背景、概念格的基本概念、概念格的构造以及概念推理等方面进行阐述，并以电影类型信息为形式背景进行形式概念分析。

2 形式概念

定义 1 设形式对象集 G，形式属性集 M，二元关系 I⊆G×M。若 X⊆G 并且 Y⊆M，X=｛x | x∈G，∀y∈Y，xIy｝，Y=｛y | y∈G，∀x∈x，xIy｝，

史程，女，1994年生，硕士研究生，E-mail: shicheng2015@126.com，研究方向为智能信息处理.

则二元组(X，Y)称为形式概念其中 X 称为形式概念的外延，表示属于这个形式概念的对象的集合；Y 称为形式概念的内涵，属于这个形式概念的属性的集合[1]。

通过定义 1 可以知道形式概念分为外延和内涵。形式概念的外延是指被表示为属于这个概念的所有的对象集合，形式概念的内涵是指被表示为所有这些对象所共有属性的集合。

3 形式背景

3.1 生成形式背景

定义 2 三元组 K=(*G*，*M*，*I*)被称为形式背景，其中 G 为形式对象的集合，M 为形式属性的集合，I 是 G 和 M 之间的二元关系，I⊆G×M。若 g 是 G 中的一个形式对象，m 是 M 中一个形式属性，那么用（g，m）∈I 表达 g 与 m 之间的关系，读作“形式对象 g 具有形式属性 m”[1]。

实际上，形势背景一般都不是直接存在的，而是通过对数据源进行分析从而提取出来的，也就是对现有概念中的对象和属性进行约简。表 1 为电影类型的形式背景，其中对象集为（*1， 2 ...10*），属性集为（科幻，动作，喜剧，爱情，悬疑）。形式概念表示为（用户，电影类别）二元组；形式背景表现为（用户，电影类别，观看喜好）的三元组，用 1 表示此用户喜爱观看此类电影，而用 0 表示此用户不喜爱观看此类电影。

表 1 电影类型形式背景

用户	科幻	动作	喜剧	爱情	悬疑
1	1	1	1	0	0
2	1	0	1	0	1
3	0	0	1	1	0
4	1	1	0	1	0
5	1	0	1	0	1
6	0	1	0	1	0
7	1	1	0	1	0
8	1	1	1	1	0
9	0	0	1	1	0
10	0	1	1	1	0

3.2 约简形式背景

形式背景的约减包括聚类(行约减)和关联(列约减)，即将属性值相同的对象进行合并和将所有对应于同一个对象集的几个属性合并。通过观察可知，用户 2 和用户 5 观看电影属性相同可以合并，用户 3 和用户 9 观看电影属性相同可以合并，用户 4 和用户 7 观看电影属性相同可以合并。约简后的电影形式背景如表 2 所示。

表 2 约简后的形式背景

用户	科幻	动作	喜剧	爱情	悬疑
1	1	1	1	0	0
2,5	1	0	1	0	1
3,9	0	0	1	1	0
4,7	1	1	0	1	0
6	0	1	0	1	0
8	1	1	1	1	0
10	0	1	1	1	0

3.3 生成单值形式背景

为了便于对电影类型的形式概念分析，需要将表 2 的多值形式背景转换为单值形式背景。将值为“1”的位置改为“×”，表示该用户喜爱观看此类电影，去掉值为“0”的位置，再用 a、b、c、d、e 分别代表各属性，1-10 分别代表各用户，得到单值形式背景如表 3 所示。

表 3 单值形式背景

	a	b	c	d	e
1	×	×	×		
2,5	×		×		×
3,9			×	×	
4,7	×	×		×	
6		×		×	
8	×	×	×	×	
10		×	×	×	

3.4 确定父子关系的单值形式背景

为了方便构造形式背景的概念格，在获取到的单值形式背景的基础上做顺序的调整，找到属性继承的父子关系，通常情况下，为方便查找，按属性由少至多自上而下排列。表 4 为带有父子关系的单值形式背景。

表 4 带有父子关系的单值形式背景

	a	b	c	d	e
6		×		×	
3,9			×	×	
1	×	×	×		
4,7	×	×		×	
2,5	×		×		×
10		×	×	×	
8	×	×	×	×	

3.5 绘制 Hasse 图

Hasse 图中的每个结点表示用户集合中的一个元素，结点的位置按其所在的偏序关系中的次序自下向上排列。即对任意 a、b 属于 A，若 a<b(a≤b∧a≠b)，则 a 排在 b 的下边。如果 a<b，且不存在 c∈A 满足 a<c<b，则在 a 和 b 之间连一条线，这样画出的图叫 Hasse 图。Hasse 图的作图法为：以“圆”表示元素；若 x<y，则 y 在 x 的上层；若 y 覆盖 x，则连线；不可比的元素在同层。通过 Hasse 图可以生动简洁地体现概念格中形式概念之间的泛化和特化关系。表 4 的带有父子关系的单值形式背景得到的 Hasse 图如图 1 所示。

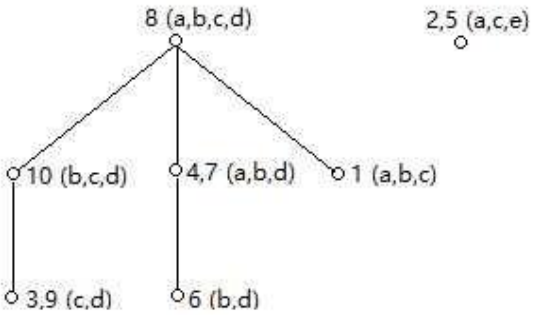

图 1 形式背景的 Hasse 图

4 概念格

4.1 基本概念

定义3 设<K，≤>为偏序集，D∈K，a 为K的任一上界，若对D 的所有上界y 均有a≤y，则称a 为D 的最小上界，即上确界。同样，若d 为D的任一下界，若对D 的所有下界z 均有z≤d，则称d 为D 的最大下界，即下确界[1]。

定义4 设<K，≤>为偏序集，如果K 中任意两个元素都有最小上界和最大下界，则称<K，≤>为格[2]。

定义 5 对于形式背景 K=(O，A，B)存在唯一的一个偏序集<K，≤>与之对应，并且该偏序集的子集的上确界与下确界都存在，这个偏序集产生的格结构称为概念格[3]。

4.2 生成概念格

由于形式背景对应的概念格中形式概念的子集必须上确界与下确界都存在，因此必须对 Hasse 图进行修补，添加形式概念的上下确界，才能形成概念格[4]。图 2 是生成的概念格。

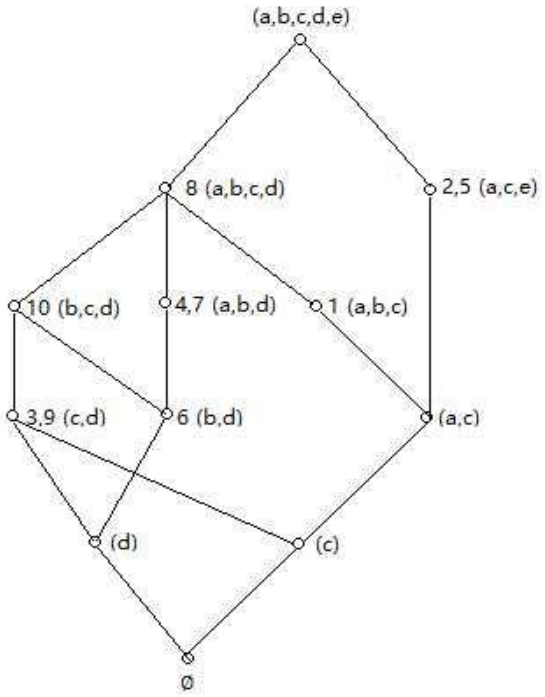

图 2 概念格

4.3 概念格的应用

概念格良好的结构特征使得概念格不仅仅能从数据中分类和定义概念，发现对象之间、属性之间的依赖关系等信息，还能很好地利用概念格中的信息形成知识。因此在智能数据处理、知识发现、数据挖掘方面，概念格的应用很广泛。目前，随着网络技术的普及，概念格在语义 Web 电子商务、Web 服务管理、入侵检测、个性化导航、搜索引擎等方面已经有了广泛的应用。由于格结构面对的都是大型的复杂数据库，因此提高应用系统的数据分析速度和系统的响应速度将是应用系统成功的关键因素。目前由 Google 推出的 BigData、MapReduce 等技术使得大数据的处理迈入了新的发展阶段，因此形式概念分析这一工具将得到更广泛的应用。

5 结束语

本文通过构建基于用户和电影类型的形式背景，给出了从概念得到形式概念、背景转换为形式背景、形式背景转化为单值形式背景再构造概念格的整体过程，全面分析其特征和关系。在这一过程中，充分体会到形式概念分析以及概念格的作用及其应用，在知识发现推理、Web 语义检索和数据挖掘中都起到重要作用[5]。概念格仍是一个年轻并在高速发展的领域。进一步的研究方向包括：高效的建格算法及剪枝算法；从格上产生有用的规则；基于格的数据挖掘等等。

参 考 文 献

[1] [德]B.甘特尔,R.威尔.形式概念分析[M].马垣,张学东等译.北京：科学出版社,2007.
[2] Tim Pattison, Derek Weber, Aaron Ceglar. Enhancing Layout and Interaction in Formal Concept Analysis//Proceedings of the IEEE Pacific Visualization Symposium, Gaithersburg, America, 2014:365-369
[3] K. Sumangali, Ch. Aswani Kumar. Determination of Interesting rules in FCA Using Information gain// Proceedings of the IEEE Pacific Visualization Symposium, Gaithersburg, America, 2014:455-479
[4] 黄映辉.智能信息处理课件:形式概念分析_第 4 章 形式概念分析[R].大连海事大学,2014.
[5] 毕强,滕广青.国外形式概念分析与概念格理论应用研究的前沿进展及热点分析[J].现代图书情报技术,2010,15(11):17-23.