《智能信息处理》课程作业

基于形式概念的全职招聘信息分析

刘琦

作业	分数[20]
得分	

2020年11月13日

基于形式概念的全职招聘信息分析*

刘琦

(大连海事大学 信息科学技术学院,辽宁 大连 116026)

摘 要：形式概念为对象集和属性集的集合，形式概念分析的核心数据分析工具是概念格结构模型，可以对集合中具有某种关系或者含有某些共同属性的对象进行分类，进而以数学化方式表达概念和概念层次。本文对全职招聘的岗位进行分析，描述对象与属性集之间的联系，给出从概念到形式概念、从背景到形式背景、从形式背景转化为单值形式背景、进而由单值形式背景构造概念格的整体过程，最后进行了形式概念的识别与推理。

关键词：形式概念分析;形式背景;概念格

中图法分类号：TP301.2 **文献标识码：**A

Full time recruitment information analysis based on Formal Concept

Qi Liu

(Dalian Maritime University, Dalian 116026, China)

Abstract: Formal concept is the set of object set and attribute set. The core data analysis tool of formal concept analysis is concept lattice, which can classify the objects with certain relationship or common attributes in the set, and then express concepts and concept levels in a mathematical way. This paper analyzes the full-time job recruitment, describes the relationship between the object and the attribute set, and gives the whole process from concept to formal concept,background to formal background,formal background to single value formal background.Then it constructs concept lattice from single value formal context. Finally, it identifies and infers formal concepts.

Key words:formal concept analysis;formal concept;concept lattice

0 序言

形式概念分析(FormalConceptAnalysis，FCA)是一种从形式背景出发进行数据分析和规则提取的有效工具。形式概念分析建立在数学基础之上，对组成本体的概念、属性以及关系等用形式化的语境表述出来，然后根据语境，构造出概念格，从而清楚地表达出本体的结构。这种本体构建的过程是半自动化的，在概念的形成阶段，需要领域专家的参与，识别出领域内的对象、属性，构建其间的关系。概念生成后，可以构造语境，然后利用概念格的生成算法 CLCA，自动产生本体。形式概念分析强调以人的认知为中心，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，成为了人工智能学科的重要研究对象，在机器学习、数据挖掘、信息检索等领域得到了广泛的应用[1]。

互联网技术飞速发展，也出现了惠及人们日常生活的各类电商平台，如何为用户呈现出海量信息中有价值的感兴趣的信息是电商平台的关注点。本文以全职招聘信息的分析为例，首先建立背景，进而进行对象/属性约简得到形式背景，然后构建概念格实现概念推理，显现了形式概念分析在实际应用中的重要作用。

1 形式背景

形式概念分析的准备工作就是建立形式背景(formal context)。形式背景被定义为一个三元组，公式为 K=(G,M,I)，其中 G 为对象集合，M为属性集合，I为G和M之间的二元关系。该三元组可以表示为二维表。在下面表 1 所示的形式背景中，招聘职位对象集合G={1,2,3,4,5,6,7,8,9,10}中的对象分别对应销售经理，在线客服，美容助理，经理助理，食品加工，产品经理，诚聘客服，销售人员，后台开发，前台人员10个职位对象。属性集合M={a,b,c,d,e,f,g,h,i,j,k}代表职位相关属性说明，其中 a 表示包住；b 表示餐补；c 表示养老保险；d 表示医疗保险；e 表示失业保险；f 工伤保险；g 表示生育保险；h 表示住房公积金；i 表示双休日，j表示员工体检，k表示交通补助。

表 1 职位信息

对象	a	b	c	d	e	f	g	h	i	j	k
1.销售经理	0	1	1	1	1	1	1	1	1	1	0
2.在线客服	0	0	1	1	1	1	1	1	1	1	0
3.美容助理	1	1	0	0	0	0	0	0	1	0	0
4.经理助理	0	0	1	1	1	1	1	1	1	1	0
5.食品加工	0	0	1	1	1	1	1	1	1	1	0
6.产品经理	1	1	1	1	1	1	1	1	1	1	1
7.诚聘客服	0	0	1	1	1	1	1	1	1	0	0
8.销售人员	0	1	1	1	1	1	1	1	1	0	0
9.后台开发	0	1	1	1	1	1	1	1	0	1	1
10.前台人员	0	1	1	1	1	1	1	1	1	0	0

2 形式概念

概念是对象集合与属性集集合的集合，反映对象的特有属性，是从对象的属性中抽出特有属性概括而成的，表达的语言形式是词或者词组。而在FCA 中形式概念一词可简单的理解为对象集的属性集，它通常用来构建自然概念的层次连通结构。而在形式概念分析中，形式概念被理解为由外延和内涵两部分组成。形式概念的内涵决定外延，外延是指属于这个概念的所有对象的集合，适用对象涵盖的范围，内涵表现了对象的特定结构与具体内容[2]。

2.1 形式概念的获取

形式概念的获取是通过对现有概念中对象和属性的约简。对象的约简是指将具有的属性相同的对象进行合并为一个形式对象， 属性的约简是指对于每个对象若干属性值相同可约简为同一形式属性。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。

2.2 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）。通过表 1 可看出，8、10与 2、4、5 分别为具有相同属性的两组对象，故将其合并；c,d,e,f,g,h 这六个属性可以合并为一个形式属性。最后得到约简后的形式背景如表 2 所示。

2.3 形成单值形式背景

为了便于分析，可以将多值背景转换为单值形式背景。由于表 2 的形式背景的关系为 {0，1} 的二值形式背景，用“×”代替“1”便可得到单值形式背景。由表2得到的单值形式背景如表 3 所示。

2.4 确定父子关系的单值形式背景

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，依照属性集判断父子关系，父概念在上方，子概念排在下方。表 4 所示为最后形成的单值形式背景。令l={c, d, e, f, g, h}。

表 2 约简后的形式背景

对象	a	b	c,d,e,f,g,h	i	j	k
1	0	1	1	1	1	0
2,4,5	0	0	1	1	1	0
3	1	1	0	1	0	0
6	1	1	1	1	1	1
7	0	0	1	1	0	0
8,10	0	1	1	1	0	0
9	0	1	1	0	1	1

表 3 单值形式背景

对象	a	b	c,d,e,f,g,h	i	j	k
1		×	×	×	×	
2,4,5			×	×	×	
3	×	×		×		
6	×	×	×	×	×	×
7			×	×		
8,10		×	×	×		
9		×	×		×	×

表 4 带有父子关系的单值形式背景

对象	a	b	l	i	j	k
7			×	×		
2,4,5			×	×	×	
8,10		×	×	×		
3	×	×		×		
1		×	×	×	×	
9		×	×		×	×
6	×	×	×	×	×	×

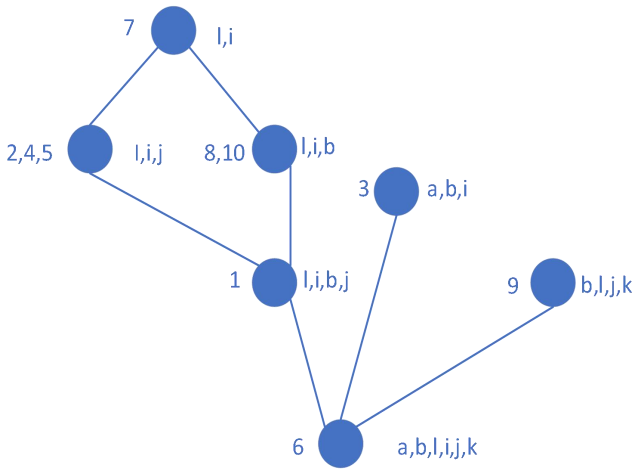

图1 Hasse 图

2.5 绘制 Hasse 图

Hasse 图中的每个结点表示集合 A 中的一个元素，结点的位置按所在偏序中的次序从底向上排列。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。应用 Hasse 图表示各结点所组成的偏序集及节点间的关系，由上到下表示的即为两节点间的父子关系，根据表 4 所绘 Hasse 图如图 1 所示。

3 偏序集和概念格基本概念

3.1 偏序集基本概念

设 H 是一个集合，如果存在 H 上的一个关系 R，对于∀x,y,z∈H，满足如下条件：（1）非完全性：∃ x 与y 不可比较。（2）自反性：xRx。（3）反对称性：xRy,yRx⇒x=y。（4）传递性：xRy,yRz⇒xRz。则称 R 是H 上的一个偏序关系，表示为“≤”，具有这种偏序关系的集合称为偏序集，表示为<H，≤>[3]。

给定集合 H，“<”是 H 上的二元关系，若“<”满足：(1)反自反性：∀a∈H，有 a≮a；(2)非对称性：∀a，b∈H，a<b ⇒ b≮a； (3)传递性：∀a，b，c∈H，a<b 且b<c，则 a<c；则称“<”是H 上的严格偏序或反自反偏序。

给定集合 H，“≤”是 H 上的二元关系，若“≤”满足：(1)自反性：∀a∈H，有 a≤a；(2)反对称性：∀a，b∈H，a≤b 且 b≤a， 则 a=b；(3)传递性：∀a，b，c∈H，a≤b 且b≤c，则 a≤c；则称“≤”是 H 上的非严格偏序或自反偏序。

设<H，≤>为偏序集，对于任意的 B⊆H，如果有 a∈H，并且对 B 的任意元素 x，都满足 x≤a，则称 a 为子集 B 的上界。同理，如果对 B 的任意元素 x，都满足 a≤x，则称 a 为子集 B 的下界。

3.2 概念格的基本概念

偏序集(H，≤)加上它所具有的属性构成一种新的数据结构：概念格。

设<H，≤>为偏序集，B H，a 为 H 任一上界，若对B 的所有上界 y 均有 a≤y，则称 a 为 B 的最小上界，即上确界。同样，若 b 为 B 的任一下界，若对 B 的所有下界 z 均有 z≤b，则称b 为B 的最大下界，即下确界。设<H，≤>为偏序集，如果 H 中任意两个元素都有最小上界和最大下界，则称<H，≤>为格。如果对格的任意非空子集A，A 中元素的上确界和下确界都存在，那么称格是一个完备格[4]。概念格是完备格。

假设给定形式背景(context)为三元组 H=(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<H，≤>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[5]。

3.3 生成概念格

图 1已经给出 Hasse 图，即已得出概念间的偏序关系，只需补出上下确界即可得到概念格。设置（7,3,9）节点，属性为空集。图 2 是产生的概念格。

4 概念识别与概念推理

概念识别，是指从与特定论域对应的概念格中识别其中的形式概念并且识别形式概念之间的关系。如图2，设置({7},{l,i})为形式概念1#，({2,4,5},{l,i,j})为形式概念2#，({8，10}，{l,i,b})为形式概念3#，({1},{l,i,b,j})为形式概念4#，({3},{a,b,i})为形式概念5#，({9},{b,l,j,k})为形式对象6#，({6},{a,b,l,i,j,k})为形式对象7#。

概念推理，是通过在概念格上的结点之间的移动，根据结点所表示的形式概念之间的关系，进行推理的过程。观察图 2 所示的概念格可以识别到，1#是 2#和 3#的共同父概念；4#是 2#和 3#的共同子概念；7#是 4#，5#和6#的共同子概念。可以识别到1#具有的属性7#必然也有，4#中的属性包含了2#和3#的属性。

5 结束语

本文以电商平台给出的全职招聘职位为论域，给出了从概念得到形式概念、从背景转换为形式背景、由形式背景转化为单值形式背景、进而由单值形式背景构造概念格的整体过程。并且通过对概念格中形式概念的识别与推理，显示了形式概念分析，尤其是概念格在知识发现、知识推理中的重要作用，他仍是一个具有巨大潜力可待开发的领域。

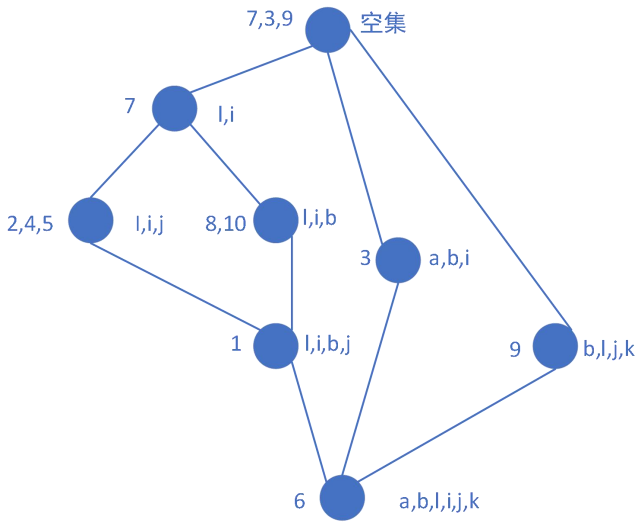

图 2 概念格

参考文献

[1] 刘琳. 基于三支概念格的决策形式背景规则提取[D].

[2] 梁亮. 形式概念分析上概念间的包含度理论研究[D]. 山西大学.

[3] Radek Janostik, Jan Konecny, Petr Krajča. Interface between Logical Analysis of Data and Formal Concept Analysis[J]. European Journal of Operational Research, 2020, 284(2).

[4] 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.

[5] 韩道军, HAN, Dao-jun, 等. 角色工程中一种最小角色集的求解算法[J]. 计算机科学, 2017, 08(v.44): 121-129.

[6] 高俊杰, 邓贵仕. 基于本体的范例推理系统研究综述[J]. 计算机应用研究, 2009, 26(02): 406-410+418.