

《智能信息处理》课程作业

基于形式概念分析的路径推荐

张新健

作业	分数[20]
得分	

2021 年 11 月 29 日

基于形式概念分析的路径推荐

张新健¹ 李冠宇^{1,2}

¹ (大连海事大学信息科学技术学院 辽宁省大连市 116026)

² (智能信息处理教育部重点实验室(大连海事大学) 辽宁省大连市 116026)

(zxj@dlmu.edu.cn)

User Relationship Mining based on Formal Concept Analysis

Zhang xinjian¹, Li Guanyu^{1,2}

¹ (College of Computer Science and Technology, Dalian Maritime University, Dalian City, Liaoning Province 116026)

² (Key Laboratory of Intelligent Information Processing Ministry of Education (Dalian Maritime University), Liaoning Province 116026)

Abstract Formal concept analysis is a powerful tool for data analysis and rule extraction from the formal context. Formal concept analysis is established on the basis of mathematics. It uses formal language to express the concepts, attributes and interrelationships in the ontology, and then constructs the concept lattice according to the context to clearly express the structure of the ontology. This paper analyzes the formal concept of user activity trajectory, and presents the whole process of obtaining the formal concept from the concept, transforming from the background to the formal background, and obtaining the reduced formal background from the formal background, then forming the single-value formal background, and then constructing the concept lattice. . Use the concept grid for subsequent applications as recommended

Key words formal concept analysis; concept lattice; formal background; trajectory

摘要 形式概念分析是从形式背景进行数据分析和规则提取的强有力的工具。形式概念分析以数学为基础而建立,对本体中的概念、属性以及相互关系等用形式化的语言表述,然后根据语境构造出概念格,清晰的表达出本体的结构。本文对用户活动轨迹进行形式概念分析,给出了从概念得到形式概念,从背景转化为形式背景,从形式背景得到约简形式背景,之后形成单值形式背景,再构造出概念格的全过程。使用概念格进行之后的应用如推荐。

关键词 形式概念分析; 概念格; 形式背景 ; 轨迹

中图法分类号 TP391

收稿日期: 2019-11-08; **修回日期:** 2020-04-16

基金项目: 国家自然科学基金项目(60903098)

This work is supported by the National Natural Science Foundation of China (60903098).

通信作者: 李冠宇 (liguanyu@dlmu.edu.cn)

随着移动互联网技术的飞速发展,信息量呈爆炸式增长,人们获得信息的方式越来越便捷,但也带来了很多的弊端,在路径规划或者地点推荐的应用中,人们不能够从大量信息中得到符合自己的信息。例如,一些旅游 app 会对用户进行一系列地点的推荐,导航 app 会为用户规划路径,而用户不一定会对推荐或者规划的地点或者路径满意。在 app 推荐时假如更有针对性,会增加用户对 app 的好感度以及体验值,使得推荐的消息更加有价值。

形式概念分析(Formal Concept Analysis, 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。原始轨迹数据中隐含着用户关系,通过对用户行动轨迹进行基于形式概念的关系挖掘,能够对用户之间的关系有所了解,一方面,推荐给用户可能存在的朋友;另一方面,存在亲密关系的用户的活动轨迹在很大程度上是相同的,可以根据关系进行更加精准的推荐地点,或者路径规划等等。使得产品对用户更加友好^[1,2]。

1 相关工作

形式概念分析是 20 世纪 90 年代 Wille 提出的一种从形式背景进行数据分析和规则提取的强有力工具^[1],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(concept lattice),从而清楚地表达出概念及概念间关系的结构。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;在概念生成之后,可以构造语境,然后利用概念格的生成算法自动产生概念格。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、本体研究、软件工程、知识发现以及 Web 语义检索等领域得到了广泛的应用^[3,4]。

1.1 形式背景与概念

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,概念反应了对象的特有属性,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。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。形式概念分析是作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。

定义 1.1 形式概念: 设形式对象集 G , 形式属性集 M 二元关系 $I \subseteq G \times M$ 。若 $X \subseteq G$ 并且 $Y \subseteq M$, $X = \{x | x \in G, \forall y \in Y, xIy\}$, $Y = \{y | y \in M, \forall x \in X, xIy\}$, 则二元关系 (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集合; Y 称为形式概念的内涵, 表示属于这个形式概念的额属性的集合^[1]。

定义 1.2 形式背景: 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, $I \subseteq G \times M$ 。若 g 是 G 的一个形式对象, m 是 M 中的一个形式属性, 那么用 $(g, m) \in I$ 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]。

一般而言, 形势背景并不是直接存在的, 需要从数据源中提取, 即对现有概念中对象和属性进行约简。对象的约简是指将具有一样属性的对象进行合并成为一个形式对象, 属性的约简是指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。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 形式背景示例

	A	B	C	D
1	0	0	1	1
2	1	1	0	0
3	1	0	1	1
4	0	0	1	0

通过观察构造好的形式背景的列间关系(即属性之间的关系), 可以进行只是发现, 也就是关联规则的提取。

1.2 概念格

设 $\langle H, \leq \rangle$ 为偏序集, $B \in H$, a 为 H 的任一上界, 若对 B 的所有上界 y 均有 $a \leq y$, 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 $z \leq b$ 则称 b 为 B 的最大下界, 即下确界。

设 $\langle H, \leq \rangle$ 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 $\langle H, \leq \rangle$ 为格。

定义 1.3 概念格: 对于形式背景 $H=(G, M, I)$ 存在唯一的一个偏序集 $\langle H, \leq \rangle$ 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。

概念格的构造通常是首先绘制与形式背景对应的 Hasse 图, 然后通过补齐各形式概念的上下确界, 进而形成概念格。而概念格的构造是形式概念分析应用的前提, 目前构造概念格的算法主要可以分为三大类: 批处理算法、增量算法和并行算法。批处理算法思想^[5]是首先生成所有概念, 然后根据他们之间的直接前驱—后继关系生成边, 完成概念格的构造, 根据其构造格的不同方式, 可以分为 3 类: 即自顶向下算法, 如 Bordat 的算法; 自底而上算法, 如 chein 的算法、枚举算法。增量算法^[5]思想是首先初始化概念格为空, 将当前要插入的对象和现有格中所有的形式概念做交运算, 根据交的结果不同采取不同的行动, 主要的区别在于边的连接, 典型的算法有 Godin、Capineto 和 T.B.Ho 的算法。概念格并行构造思想就是通过形式背景的拆分, 形成分布存储的多个子背景, 然后同时并行构造相应的子概念格, 再由子概念格

的合并得到所需的概念格。如基于分治策略的 Para_Prun 算法, 是基于偏序集上闭包系统分解的思想提出的新的概念格分布处理算法, 使闭包系统的分解既不会有冗余信息的产生, 也不会使信息丢失, 并用于概念格的分布处理^[6]。

1.3 单值背景

为了便于分析, 可以将多值背景转换为单值形式背景。如表 1 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。为了构造形式背景对应的概念格, 需要将单值形式背景转换为带有父子关系的单值形式背景(这里的父子关系是以属性的个数多少进行排序的), 由表 1 生成的带父子关系的单值形式背景如表 2 所示, 然后构造概念格。

表 2 带父子关系的单值形式背景

	A	B	C	d
4	0	0	1	0
1	0	0	1	1
2	1	1	0	0
3	1	0	1	1

2 用户轨迹形式背景

在本节中, 我们主要介绍通过用户活动轨迹的定义以及其形式背景。

2.1 活动轨迹

定义 2.1 轨迹: 用户的轨迹是由一系列有序的时间戳对组成的, $\{(v_1, t_1), (v_2, t_2), \dots, (v_i, t_i), \dots, (v_m, t_m)\}$, 其中 $t_i < t_{i+1}$ 。

本章用元组 (v_i, t_i) 表示一个用户的轨迹表示用户在 t_i 时刻停留在 v_i 位置。其中, 每个时间间隔为 30min。例如 $t_{r1} = \{(2, t_1), (3, t_2), (4, t_3), (6, t_4), (0, t_5), (0, t_6)\}$ 表示用户在 t_1 时刻停留在位置 2, 之后 t_2 时刻去了位置 3, 之后 t_3 时刻去了位置 4, 之后 t_4 时刻去了位置 6, 之后 t_5 时刻去了位置 0。最后停留在了位置 0。

2.2 轨迹形式背景

一张表代表一个时间跨度用户是否到达过地点 i 。以用户名称作为对象,用户在时间跨度 $t_i \rightarrow t_j$ 所到达的地点作为属性,以此建立形式背景。如表 3 所示,其中对象集 $\{ID_1, ID_2, ID_3, \dots, ID_{10}\}$, 属性集为 $\{1, 2, 3, 4, \dots, 6\}$ 。其中概念可以形式化为序偶(用户, 地点)二元组;形式背景表现为(用户, 地点, 到达关系)的三元组,用 1 表示此用户在 $t_i \rightarrow t_j$ 内到达过地点 m , 而用 0 表示此用户未到达过。

表 3 用户轨迹形式背景

用户	1	2	3	4	5	6
ID_1	0	1	0	1	0	0
ID_2	1	1	0	1	1	0
ID_3	0	0	1	0	1	1
ID_4	0	1	1	1	1	0
ID_5	0	0	0	0	1	0
ID_6	1	1	0	1	1	0
ID_7	0	0	1	0	1	0
ID_8	0	0	1	0	1	1
ID_9	0	1	0	1	1	0
ID_{10}	0	0	1	0	0	0

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合并,将相同的属性也进行合并。通过观察将属性电视剧和电影合并, ID_2 和 ID_6 到过的地点相同可以合并, ID_3 和 ID_8 到过的地点相同可以合并。那么最后得到的形式对象集合为 $\{ID_1, ID_2 \& ID_6, ID_3 \& ID_8, ID_4, ID_5, ID_7, ID_9, ID_{10}\}$, 最后得到的形式属性集合为 $\{1, 2 \& 4, 3, 5, 6\}$, 约简后得到的形式背景如表 4 所示。

表 4 约简后的形式背景

用户	1	2&4	3	5	6
ID_1	0	1	0	0	0
$ID_2 \& ID_6$	1	1	0	1	0
$ID_3 \& ID_8$	0	0	1	1	1
ID_4	0	1	1	1	0
ID_5	0	0	0	1	0
ID_7	0	0	1	1	0
ID_9	0	1	0	1	0
ID_{10}	0	0	1	0	0

为了更方便的对微博兴趣进行概念分析, 我们需要将多值形式背景转化为单值形式背景, 在表 4 中将值为“1”的位置用“×”表示该用户到达过次低点, 并将值为“0”的位置空白。由此得到的单值形式背景如表 5 所示。对所生成的单值形式背景按照概念格的构造方法将单值形式背景转化为带有父子关系的单值形式背景, 原则是基于属性个数的数量进行排序。譬如, 在表 5 中用户 10 只有一个属性, 则将它排至第一位, 其它同理。由此得到的带有父子关系的单值形式背景如表 6 所示。

3 活动轨迹概念格

3.1 绘制 Hasse 图

Hasse 图中的每个结点表示集合 A 中的一个元素, 结点的位置按它们在偏序中的次序从底向上排列。即对任意 a, b 属于 A , 若 $a \leq b$ 且 $a \neq b$, 则 a 排在 b 的下边。如果 $a \leq b$ 且 $a \neq b$, 且不存在 $c \in A$ 满足 $a \leq c$ 且 $c \leq b$, 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 $x \leq y$, 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 图的节点即对象, 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。由此而画出的 Hasse 图如图 1 所示。

表 5 单值形式背景

用户	1	2&4	3	5	6
ID_1		×			
$ID_2 \& ID_6$	×	×		×	
$ID_3 \& ID_8$			×	×	×
ID_4		×	×	×	
ID_5				×	
ID_7			×	×	
ID_9		×		×	
ID_{10}			×		

表 6 带父子关系的形式背景

用户	1	2&4	3	5	6
ID_1		×			
ID_5				×	
ID_{10}			×		
ID_7			×	×	
ID_9		×		×	
$ID_2 \& ID_6$	×	×		×	
$ID_3 \& ID_8$			×	×	×
ID_4		×	×	×	

3.2 生成活动轨迹的概念格

概念格是基于父子关系的渐进式构建，概念格需要子集的上下确界都存在，从而需要补充形式概念的上、下确界，即对概念格进行修补^[7]。补全图 1 的 Hasse图，用 1, 2...10 代表形式概念，得到的概念格如图 2 所示。

图 1 形式背景的Hasse图

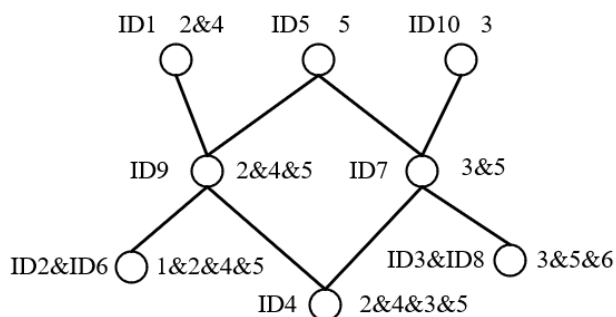
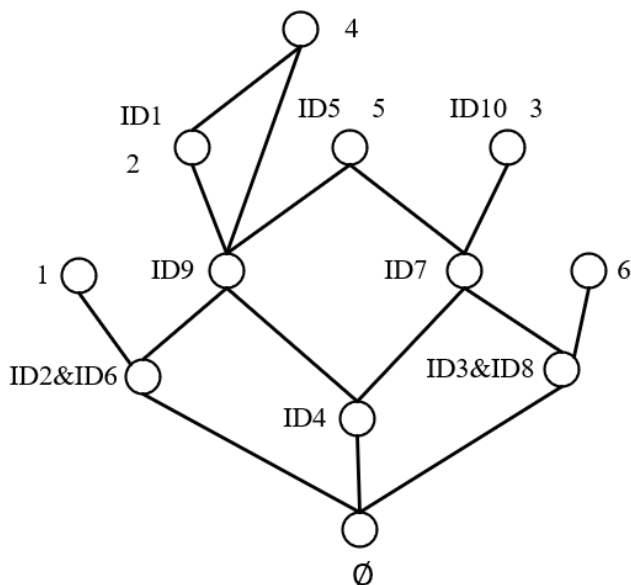


图 2 概念格



3.3 基于概念格的决策

基于概念格的决策包括概念识别和概念推理。

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。观察图 2 所示的概念格，可以识别到：

ID_9 是 ID_2 和 ID_5 的共同子概念，即 ID_2 和 ID_5 所到达的地点 ID_9 也可能会去。

ID_5 是 ID_{10} 和 ID_7 的共同父概念，即 ID_5 所到达的地点 ID_7 和 ID_{10} 也可能会去。这就为路径推荐或者路径规划提供了基础。

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，则可以得到如果用户到达过地点 2，那么这个用户可能也会去 4。另外，对于一个新的地点，可以将其加入地点集合，根据这些用户的形式概念分析来其他用户进行推荐或者路径规划^[9]。譬如用户 ID_1 , ID_2 , ID_4 , ID_6 , ID_9 去过一个新的地点 7，可得到 If ID_1 Then 7 的关联规则^[7]，那么就可以将新地点 7 推荐给去过 1, 2, 4, 5 的用户，如表 7 所示。

表 7 新地点识别

用户	1	2&4	3	5	6	7
ID_1	0	1	0	0	0	1
$ID_2 \& ID_6$	1	1	0	1	0	1
$ID_3 \& ID_8$	0	0	1	1	1	1
ID_4	0	1	1	1	0	1
ID_5	0	0	0	1	0	1
ID_7	0	0	1	1	0	1
ID_9	0	1	0	1	0	1
ID_{10}	0	0	1	0	0	0

4 结论

本文通过基于用户发布的活动轨迹构建形式背景,给出了从概念转化为形式概念、背景转化为形式背景、约简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。最后进行了形式概念的识别与推理,将新的内容通过用户发布确定所属类别实现更准确和更有针对性的地点推荐或者路径规划,提升服务质量,也从中充分体会到形式概念分析以及概念格,在知识发现推理、Web 语义检索和数据挖掘中的重要作用^[4]。

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] 张伟.社交网络中基于形式概念分析的用户推荐.西华大学,2012.
- [3] Baidu.Formal Concept Analysis.<http://baike.baidu.com/view/4660144.htm>[OL],2014,11,1.
- [4] 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.

[5] 杨强,赵明清.概念格研究进展[J].计算机工程与设计,2008,29(20):5293-5296.

[6] 董辉,马垣,宫玺.概念格并行构造算法研究[J].广西师范大学学报:自然科学版,2008,8(3):122-124.

[7] 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.

[8] 茅琴娇,冯博琴,李燕等.一种基于概念格的用户兴趣预测方法[J].山东大学学报:工学版,2010,22(5):159-163.

[9] 刘兆庆,伏玉琛,凌兴宏,熊湘云.基于形式概念分析博客社区发现.苏州大学,2013.

Zhang Xinjian, born in 1998. M.S. His main research interests include Web mining, information retrieval, machine learning.

张新健,1998年生,硕士。主要研究方向为Web挖掘、信息检索和机器学习。



Guanyu Li He is currently a professor in Dalian Maritime University, China. His primary research interests are in Semantic Web, Ontology Engineering, Internet of Things, Knowledge Graph, etc. He has published more than 60 papers in refereed journals and conferences

李冠宇 现任大连海事大学教授。主要研究方向为语义网、本体工程、物联网、知识图谱等,在权威期刊和会议上发表论文60余篇。