基于形式概念分析的微博兴趣分析

杨姿
(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。本文对微博发布的内容进行形式概念分析，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。

关键词：形式概念分析；概念格；形式背景
中图法分类号：TP305　　文献标识码：A　　文章编号：1000-7024(2016) 09-1-04

Micro-blog interest analysis based on Formal Concept Analysis

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

Abstract：Formal concept analysis is a powerful tool for data analysis and rule extraction from formal context. Formal concept analysis based on Mathematics and the establishment of the concept, body, attribute and relationship with the formal language, and then according to the context to construct the concept lattice, expressing ontology structure clear. In this paper, micro-blog released the contents of the formal concept analysis, given the formal concept from the concept, from the background into the formal context, reduction of formal context from the formal context, after the formation of single valued context, then we construct the whole process of concept lattice.

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

0 引　言

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，人们不能够从大量信息中得到符合自己兴趣的信息。例如，微博客户端会对用户进行一系列消息的推荐，而用户不一定会对推荐的消息感兴趣，也就是用户的兴趣点不在于此，在微博推荐消息时假如更有针对性，会增加用户对微博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，通过对用户发布的微博内容进行基于形式概念的兴趣分析，能够对用户的兴趣进行了解，一方面，可以对微博用户发布的微博内容分析出用户的喜好；另一方面，在微博向用户推荐消息时更有针对性，使得用户能够能到相对符合自己喜好的信息，从而提高了微博推荐的质量，也使得用户对微博的体验值有所增加[1,2]。

1 形式概念分析简介

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

收稿日期：2016-11-15；修订日期：2016-12-5。
基金项目：国家自然科学基金项目（60972014）。
作者简介：杨姿（1993— ），女，河北石家庄人，硕士研究生，研究方向为智能信息处理。
E-mail：dmuyangz@gmail.com

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 形式背景与形式概念

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。

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 $Y\subseteq M$，$X=\{x\mid x\in G, \forall y\in Y, xIy\}$，$Y=\{y\mid y\in G, \forall x\in x, xIy\}$，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]。

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 中的一个形式对象，m 是 M 中一个形式属性，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（属性间的关系），可以进行知识发现，也就是关联规则的提取。

1.2 概念格

设<H，≤>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≤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。

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<H，≤>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。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 微博内容形式背景

在微博中刻画目标用户的最简单的一种方法是基于用户通常喜欢发布他们感兴趣的信息，因此目标用户的兴趣偏好可以用目标用户本人所发布的微博来刻画。整合微博用户

自身所发布的微博以表示目标用户的兴趣特征。以用户名称作为对象，用户所发布的微博中的词项作为属性，以此建立形式背景，如表 3 所示，其中对象集{ID1,ID2,ID3,…ID10}，属性集为｛时事热点，明星，电影，体育，电视剧，美食，科技｝。其中概念可以形式化为序偶（用户，词项）二元组；形式背景表现为（用户，词项，发布关系）的三元组，用 1 表示此用户发布的微博中包含此词项，而用 0 表示此用户发布的微博中不包含此词项。

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。通过观察将属性电视剧和电影合并，ID2 和 ID6 发布的微博内容的词项相同可以合并，ID3 和 ID8 所发布的微博内容的词项相同可以合并。那么最后得到的形式对象集合为｛ID1，ID2&ID6，ID3&ID8，ID4，ID5，ID7，ID9，ID10｝，最后得到的形式属性集合为{时事热点，明星，影视剧，体育，电视剧，美食，科技｝，约简后得到的形式背景如表 4 所示。

表 3 微博内容形式背景

用户	时事热点	明星	电影	体育	电视剧	美食	科技
ID1	1	0	1	0	1	0	0
ID2	1	1	1	0	1	1	0
ID3	1	0	0	1	0	1	1
ID4	1	0	1	1	1	1	0
ID5	1	0	0	0	0	1	0
ID6	1	1	1	0	1	1	0
ID7	1	0	0	1	0	1	0
ID8	1	0	0	1	0	1	1
ID9	1	0	1	0	1	1	0
ID10	1	0	0	0	0	0	0

表 4 约简后的形式背景

用户	时事热点	明星	影视剧	体育	美食	科技
ID1	1	0	1	0	0	0
ID2,ID6	1	1	1	0	1	0
ID3,ID8	1	0	0	1	1	1
ID4	1	0	1	1	1	0
ID5	1	0	0	0	1	0
ID7	1	0	0	1	1	0
ID9	1	0	1	0	1	0
ID10	1	0	0	0	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 中用 1-10 代表各个用户，用 a，b，c，d，e，f 分别代表各属性，并且将值为“1”的位置用 “×”表示该用户发布的微博中包含此词项，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 微博内容概念格

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≤b 且 a≠b,则 a 排在 b 的下边。如果 a≤b 且 a≠b，且不存在 c∈A 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

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

表 6 带有父子关系的单值形式背景

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

3.2 生成微博内容的概念格

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。

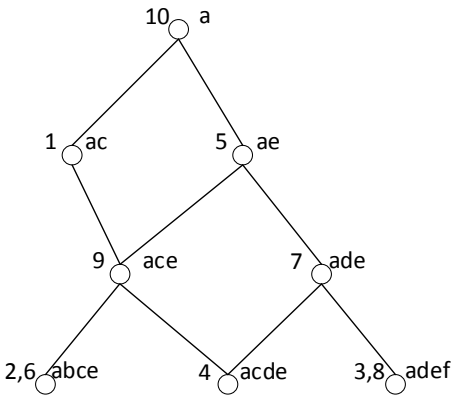

图 1 形式背景的 Hasse 图

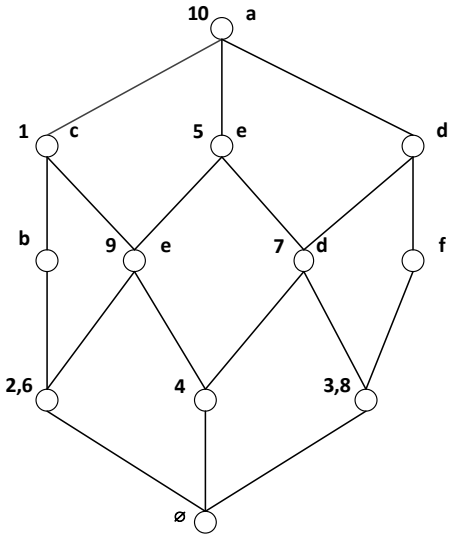

图 2 概念格

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：

(1) 用户9是用户5和用户1共同的子概念，即用户5和用户1所关注的微博内容用户9也都感兴趣；

(2) 用户5是用户9和用户7共同的父概念，即用户5感兴趣的微博，用户9和用户7也同样感兴趣。

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对明星相关的微博感兴趣，那么这个用户也会对影视剧相关的微博感兴趣。另外，对于一个新的微博词项，可以根据早些时段发布微博的用户，根据这些用户的形式概念分析来对其满足同样兴趣的用户进行推送[9]。譬如新词项 D1 被用户 1、2、4、6、9 发布，可得到 If c Then D1 的关联规则[7]，那么就可以将内容 D1 推荐给对影视剧词项感兴趣的用户，如表 7 所示。

表 7 新内容识别

	a	b	c	d	e	f	D1
1	1	0	1	0	0	0	1
2,6	1	1	1	0	1	0	1
3,8	1	0	0	1	1	1	0
4	1	0	1	1	1	0	1
5	1	0	0	0	1	0	0
7	1	0	0	1	1	0	0
9	1	0	1	0	1	0	1
10	1	0	0	0	0	0	0

4 结束语

本文通过基于用户发布的微博内容词项的类别构建形式背景，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消息推荐，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.