# 基于若干饮料信息的形式概念分析*

朱宗梅

(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（Formal Concept Analysis）是应用数学的一个分支，它可以极大限度地对集合中具有某种关系或者含有某些共同属性的元素进行分类，发现由属性和对象构成的概念和概念之间的关系，进而以数学化方式表达概念和概念层次。本文基于几种常见的饮料，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，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**：TP18 **文献标识码**：A

## Formal Concept Analysis base on some drinks information

ZHU Zongmei

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

**Abstract**: Formal Concept Analysis is a branch of applied mathematics，which can classify the elements with relationship or Some common attributes in a set to the great limit, and find relationship between concept and concept constituted by attributes and object, then express it in a mathematical way.Based on several common drinks, this paper gives the whole process of extracting formal concept from concept，converting context to formal context, transforming formal context into a single valued formal context, and by the single valued formal context forming concept lattice. And in the end, the recognition and reasoning of formal concept is conducted.

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

形式概念分析(Formal Concept Analysis，FCA)是由德国达姆斯塔特工业大学的 R.Wille 教授提出的一种形式化描述概念的方法，它是格论与序论相结合而产生的理论，能将概念和概念层次以数学形式清楚的表示出来。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 lat-tice),即本体，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，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 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。

本文从几种常见饮料出发，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，关于对象集合 G={1,2,3,4,5,6,7,8,9,10}，其中 a 表示 $CO^2$；b 表示乳酸菌；c 表示原果汁；d 表示酸味剂；e 表示果香型食用香精；f 表示鲜乳；g 表示乳制品；h 表示新鲜\冷藏水果；i 表示新鲜\冷冻蔬菜。属性集合 M={a,b,c,d,e,f,g,h,i}，二元关系 I 为确定性关系。实际上，形式背景一般都不是直接存在的，需要从数据源中提取，从而就需要对数据源进行分析，采取不同的策略和算法来提取形式背景。

* Supported by the National Natural Science Foundation of China under Grant No.60902090, 61073057.

表 1 饮料信息

| | a | b | c | d | e | f | g | h | i |
|---|---|---|---|---|---|---|---|---|---|
| 1.碳酸型饮料 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.果汁型 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 3.果味型 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 4.果汁饮料 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 5.蔬菜汁饮料 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 6.含乳饮料 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 7.配制型含乳饮料 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| 8.发酵型含乳饮料 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 9.桔汁汽水 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10.桔汁饮料 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |

## 2 形式概念

就“概念”一词而言，“概”本身可指“刮平斗斛的木板”，其延伸意为“标准化、精炼”，而“念”表示“常思”，因此，概念一词本意为被精炼的常思。实际上，概念是一种反映对象的特有属性的思维方式，它主要是从对象的属性中抽出特有属性概括而成的。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。形式概念的外延是指被表示为属于这个概念的所有对象的集合，形式概念的内涵是指被表示为所有这些对象所共有的特征（或属性）集合。

### 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全一样的对象进行合并为一个形式对象，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。

例如：({碳酸型饮料}，{$CO^2$，原果汁})和({果汁饮料}，{原果汁，新鲜\冷藏水果})是两个概念，对其共有的属性取其共有的属性可以得到一个形式概念({碳酸型饮料，果汁饮料}，{原果汁})。

### 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，1、9 与 2、4、10 是两组有相同属性的行，故将其合并；f 和 g 这两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。

表 2 约简后的形式背景

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

### 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。

表 3 单值形式背景

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

### 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，例如 7 可由对 6 的全部属性继承的基础上添加自身属性 d 得到；8 可由对 6 的全部属性继承的基础上添加自身属性 b 得到。通常情况下，为方便查找，从上倒下按属性的多少进行排列。表 4 所示为最后形成的单值形式背景。

表 4 带有父子关系的单值形式背景

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

### 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 $a<b$($a\leqslant b\wedge a\neq b$)，则 a 排在 b 的下边。如果 $a<b$，且不存在 $c\in A$ 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。

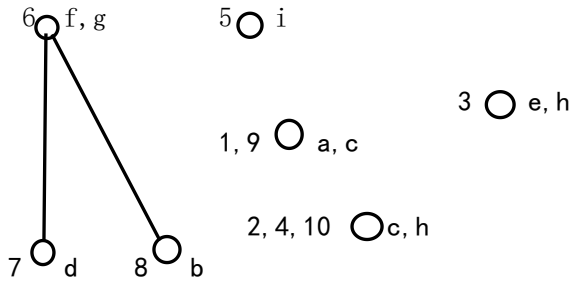


图 1 Hasse 图

## 3 概念格

### 3.1 偏序集和概念格基本概念

3.1.1 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$\forall$ x,y,z$\in$H，满足如下条件:（1）非完全性：$\exists$ x 与 y 不可比较。（2）自反性：xRx。（3）反对称性：xRy,yRx$\Rightarrow$x=y。（4）传递性：xRy,yRz$\Rightarrow$xRz。则称 R 是 H 上的一个偏序关系，表示为“$\leqslant$”，具有这种偏序关系的集合称为偏序集，表示为<H，$\leqslant$>。

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。

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。

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。

3.1.2 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 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。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。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。

### 3.2 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。用 1#，2#...11#代表形式概念，图 2 是产生的概念格。从({1,9},{a,c})和({2,4,10},{c,h})中抽取属性 a 产生新的形式概念({1,2,4,9,10},{c})作为上确界；从({2,4,10},{c,h})和({3},{e,h})中抽取属性 h 产生新的形式概念({2,3,4,10},{h})作为上确界；

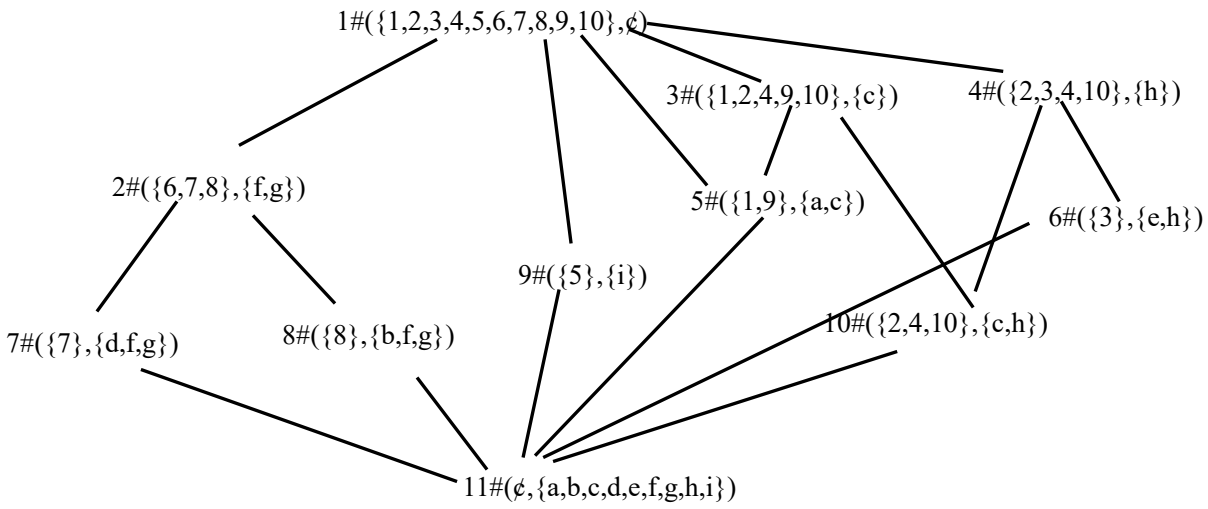


图 2 概念格

## 4 概念识别与概念推理

### 4.1 概念识别

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，10#是 3#的子概念，同时 10#是 3#和 4#的共同子概念；4#是 6#和 10#的共同父概念；2#是 7#和 8#的共同父概念；5#是 3#的子概念。可以识别到 10#({2,4,10},{c,h}),可以识别到形式对象{2,4,10}都有属性 c 和 h,也就是{果汁型，果汁饮料，桔汁饮料}都会有{原果汁，新鲜\冷冻水果}的特性；可以识别到 8#({8},{b,f,g}),可以识别到形式对象{8}具有属性 b,f 和 g,也就是{发酵型含乳饮料}会有{乳酸菌，鲜乳，乳制品}的特性，其它的结点也可识别出类似的结论。

### 4.2 概念推理

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 的概念格，可以得出以下的推理：(1)对于结点 3#，5#和 10#，由于 5#和 10#是 3#的子概念，所以可以得到 If a($CO^2$) then c(原果汁),并且 If h(新鲜\冷藏水果) then a($CO^2$)；(2)对于结点 9#，10#和 11#，11#是 9#和 10#的共同子概念，所以可以得到 If g(乳制品) then

hi(新鲜\冷藏水果，新鲜冷冻蔬菜)；(3)对于结点 9#和 10#，由于二者无直接或者间接继承关系，故可得 If c(原果汁) Then ~i(不是新鲜\冷冻蔬菜)。

## 5 结束语

本文以几种常见的饮料为论域，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。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] 李冠宇. 智能信息处理课件. 大连海事大学, 2014.

[2] 干特尔，威尔，马垣.形式概念分析.科学出版社，2007.

[3] 曲开社.偏序集、包含度与形式概念分析.计算机学报，2006，29(2):219-226.