DOI:10.11992/tis.201606006
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160808.0830.014.html

基于权值最大圈的概念格构造算法

毛华,刘祎超
(河北大学 数学与信息科学学院,河北 保定 071002)

摘 要:概念格作为一种有效的知识发现与数据处理的工具,在许多领域得到了广泛应用。寻找形式背景下的所有概念是概念格理论研究的一个基本问题。对于一个给定的形式背景,在属性拓扑图的基础上,结合图论的思想,给出了一种概念格的构造算法。算法过程如下:首先,构造弱化的属性拓扑图;其次,通过寻找弱化的属性拓扑图中的每个权值最大圈方法来生成概念,形式背景的所有概念被生成;最后,构造出概念格。通过分析说明此算法复杂度比以往的一些算法复杂度低。此外,用一个实例验证了这一算法的有效性与正确性。为知识获取提供了有益的思路与方法。

关键词:形式背景;概念格;概念;权值;最大圈;属性拓扑;数据处理
中图分类号: TP18 **文献标志码**:A **文章编号**:1673-4785(2016)04-0519-07

中文引用格式:毛华,刘祎超. 基于权值最大圈的概念格构造算法[J]. 智能系统学报, 2016, 11(4): 519-525.
英文引用格式:MAO Hua, LIU Yichao. An algorithm for concept lattice construction based on maximum cycles of weight values [J]. CAAI Transactions on Intelligent Systems, 2016, 11(4): 519-525.

An algorithm for concept lattice construction based on maximum cycles of weight values

MAO Hua, LIU Yichao
(School of Mathematics and Information Science, Hebei University, Baoding 071002, China)

Abstract: As an effective tool for knowledge discovery and data processing, the concept lattice has been widely applied in many fields. Searching all concepts in a formal context is a basic problem for research into concept lattice theory. On the basis of attribute topology and combined with the idea of graph theory, an algorithm to construct a concept lattice in a fixed formal context is given. The process is as follows: firstly, a weakened attribute topology was built up; then, by applying the method of searching the maximum cycle with a weight in the above weakened attribute topology, all of the formal context concepts were obtained; finally the concept lattice was established. Subsequent analysis illustrated that the algorithm can reduce complexity compared with some existing algorithms. In addition, using an example, the accuracy and validity of the algorithm was verified. The result presents a useful idea and method for knowledge acquisition.

Keywords: formal context; concept lattice; concept; weight value; maximum cycle; attributes topology; data processing

概念格[1]是对背景中属性、对象及其关系进行分析研究的理论。它提供了一种支持数据分析和知识处理的数学工具[2-3]。目前,概念格已经广泛应用于数据挖掘[4]、信息处理[5]、软件工程[6]和其他方面[7-8]。概念格理论的研究不仅能用于解决知识发现领域中所涉及的关联规则、蕴含规则、分类规则的提取,还能够实现对信息的有机组织、减少冗余度、简化信息表,所以对概念格理论及其算法的研究

收稿日期:2016-06-02. 网络出版日期:2016-08-08.
基金项目:国家自然科学基金项目(61572011);河北省自然科学基金项目(A2013201119).
通信作者:刘祎超.E-mail:1026074348@qq.com.

具有重要的意义。

概念是人类进行知识表达的一种手段,数据库知识发现的过程就是将数据库中蕴含的知识形式化成有用的概念的过程。对形式背景的表示及寻找背景下的所有概念是概念格理论研究的基本问题。近年来许多学者从图论的方面对概念格进行研究,例如,张涛等[9]提出用属性拓扑图来表示形式背景,并在此属性拓扑图的基础上进行概念计算;A.Berry 等[10]将形式背景构造成二部图,利用团的思想生成概念;此外,李立峰等[11]利用弦二部图对概念格进行表示,其中判断弦二部图中是否有圈,是判断弦二部图的关键。这也证实图论特别是图中的圈,在概念格的研究中之重要。

本文结合图论的知识,将形式背景以属性拓扑图表示出来,通过构造弱化的属性拓扑图,然后寻找弱化的属性拓扑图中的权值 w 之最大圈,用以生成概念,从而构造出概念格,并结合实例分析了这一算法的有效性。

1 基本概念

本节将回顾概念格与图论的一些性质和定义,关于概念格的更多内容参见文献[12],有关图论详细内容参见文献[13],并且简单描述形式背景的属性拓扑图表示方法,更多详细内容参见文献[9,14]。

1.1 概念格

定义 1

1)形式背景(O,M,I)是一个三元组,其中 O 是对象集,M 是属性集,$I\subseteq O\times M$。O 和 M 中的元素分别称为对象和属性。

2)设 $A\subseteq O$ 且 $B\subseteq M$, 定义

$$A'=\{m\in M\mid(\forall g\in A),gIm\}$$

$$B'=\{g\in O\mid(\forall m\in B),gIm\};$$

若 $A'=B$ 且 $B'=A$,则元素对(A,B)是一个概念。A 为概念(A,B)的外延,B 为概念(A,B)的内涵。形式背景(O,M,I)的所有概念的集合用$\beta(O,M,I)$表示,称$\beta(O,M,I)$为(O,M,I)的概念格。

3)对于$\beta(O,M,I)$中的概念(A_1,B_1)和(A_2,B_2),如果 $A_1\subseteq A_2$,我们写作$(A_1,B_1)\leqslant(A_2,B_2)$。很容易看到$(\beta(O,M,I);\leqslant)$是一个完备格。

例 1 形式背景(O,M,I),其中 $O=\{1,2,3,4,5,6\}$,$M=\{a,b,c,d,e,f,g\}$,关系 I 如表 1 所示。

表 1 形式背景(O,M,I)

Table 1 Formal context(O,M,I)

对像	a	b	c	d	e	f	g
1	×	×	×				
2	×	×		×			×
3		×	×		×		
4	×			×	×		×
5			×			×	
6				×			

表 1 对应的形式背景的概念格见图 1。

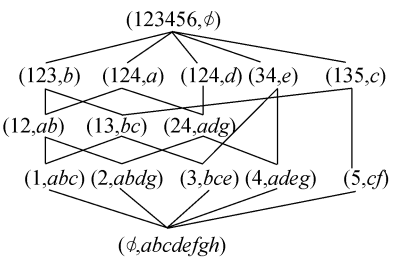

图 1 $\beta(O,M,I)$

Fig.1 $\beta(O,M,I)$

说明 1 本文所讨论的形式背景中不含有满足以下条件的属性和对象,$m\in M$, $m'=O$ 或 $m'=\varnothing$;$g\in O$, $g'=M$ 或 $g'=\varnothing$。

1.2 图论

定义 2 1)称数学结构 $G=\{V(G),E(G),\psi_G\}$为一个图,其中 $V(G)$为非空集合,ψ_G是从集合 $E(G)$到 $V(G)\times V(G)$的一个映射, 则称 G 是一个以 $V(G)$为顶集合,以 $E(G)$为边集合的有向图,$V(G)$中的元素称为 G 的顶点。$E(G)$中的元素称为 G 的边,ψ_G 称为 G 的关联函数。若 $\psi_G(e)=(u,v)$,$e\in E(G)$,$(u,v)\in V(G)\times V(G)$,简写成 $e=uv$,称 u 是有向边 e 的尾,v 是有向边 e 的头。擦掉有向图中的箭头,则得到无向图。

2)在顶边交错链 $P=v_0e_1v_1e_2\cdots v_ke_k$ 中,$e_i\in E(G)$,$i=1,2,\cdots,k$,$v_j\in V(G)$,$j=1,2,\cdots,k$,且 $e_i=v_{i-1}v_i$,则称 P 是 G 的一条道路,其中允许 $v_i=v_j$或 $e_i=e_j$,$i\neq j$。称 v_0是 p 的起点,v_k是 p 的终点。各顶相异的道路称为轨道;起点与终点重合的轨道称为圈。

3)在一个无向图中,只有一个顶的圈叫做自环;$\psi_G(e_1)=\psi_G(e_2)=(u,v)$,则称 e_1与 e_2是重边。

说明 2 由上述定义可知自环、重边均为圈。

1.3 属性拓扑图

定义 3 设(O,M,I)是一个形式背景。按如下规则构造属性拓扑图$(A(O,M,I),w)$:

1) 设 $m_1,m_2\in M$ 且 $m_1\neq m_2$。

①若 $m'_1\not\subset m'_2$,$m'_2\not\subset m'_1$且 $m'_1\cap m'_2\neq\varnothing$,则用“↔”连接 m_1和 m_2;

②若 $m'_1 \subset m'_2$ 且 $m'_1 \cap m'_2 \neq \varnothing$,则用“→”连接 m_1 和 m_2 表示为 $m_2 \rightarrow m_1$;

③若 $m'_1 \cap m'_2 = \varnothing$, 则 m_1 和 m_2 没有边连接。

2) 设 $(A(O,M,I),w)$ 为 (O,M,I) 的属性拓扑图,$e(m_i,m_j) \in E(A(O,M,I),w))$,$E(A(O,M,I),w)$ 为 $(A(O,M,I),w)$ 的边集,$e(m_i,m_j)$ 上的权值用 $w(m_i,m_j)$ 表示,$w(m_i,m_j)$ 为属性 m_i 和 m_j 之间的公共对象 $\{g_1, g_2, \cdots, g_n\}$ 的集合,称 $w(m_i,m_j)$ 为 m_i 和 m_j 之间的权值。

3) 设 $m \in M, b \in M$,若与 m 连接的边均为非单向入边,即与 m 连接的边均为 $m \rightarrow b$ 或 $m \leftrightarrow b$,则称 m 为顶层属性,顶层属性的集合用 T 表示。

例 2　图 2 为表 1 形式背景 (O,M,I) 对应的属性拓扑图。

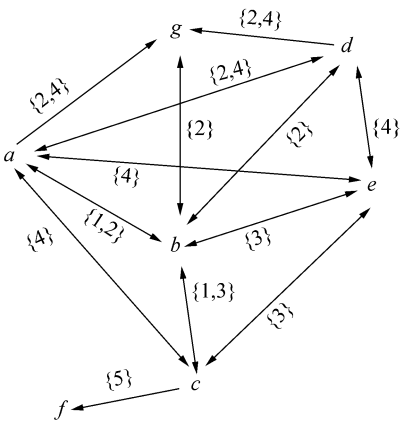

图 2　$(A(O,M,I),w)$

Fig.2　$(A(O,M,I),w)$

引理 1　设 (O,M,I) 是一个形式背景,$(A(O,M,I),w)$ 为 (O,M,I) 的属性拓扑图,若 $m \in T$,则 $(m', \mathrm{m}) \in \beta(O,M,I)$。

2　概念格的构造

在搜索概念的过程中,为了不受方向的限制,首先进行属性拓扑图弱化,将有向图变为无向图,实际上目前结合图论生成概念格的算法,都是在无向图的基础上进行的。其次,给出弱化后属性拓扑图关于某个权值的最大圈的定义。最后,给出利用权值的最大圈构造概念算法,并进行算法分析。

2.1　弱化的属性拓扑图

设 (O,M,I) 是一个形式背景,按照如下规则对属性拓扑图进行弱化:

1) 去掉属性拓扑图中的方向,得一无向图。

2) 若 m 在 $(A(O,M,I),w)$ 中为顶层属性,则在 1) 中的无向图中,加一个以 m 为顶点的自环。

3) 若在 1) 中的无向图中,包含权值 $w(u,v)$ 的只有一条边 e,其中,u、v 为 e 的两个端点,则在 u 与 v 之间再添加一条边 e_1(图中用虚线表示),并且令 e_1 的权值也为 $w(u,v)$。

完成 1)~3) 后得到的加权无向图称为弱化的属性拓扑图,用 $(R(O,M,I),w)$ 表示。

此外,显然,在上述 3) 中的 e 与 e_1 是重边。

例 3　下面图 3 为图 2 所对应的 $(A(O,M,I),w)$ 之弱化的属性拓扑图。

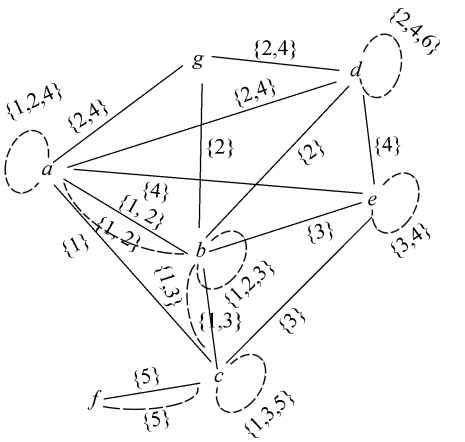

图 3　$(R(O,M,I),w)$

Fig.3　$(R(O,M,I),w)$

定义 4　设 (O,M,I) 是一个形式背景,$(R(O,M,I),w)$ 是 (O,M,I) 对应的弱化的属性拓扑图,$\{m_1,m_2,\cdots,m_h\} \subseteq M$ 且 $\{m_1,m_2,\cdots,m_h\}' \neq \varnothing$,若不存在任意 $m_a \in M$, $m_a \notin \{m_1,m_2,\cdots,m_h\}$, 使得 $\{m_1,m_2,\cdots,m_h\}' = \{m_1,m_2,\cdots,m_h, m_a\}'$,则称 $Y = \{m_1,m_2,\cdots,m_h\}$ 为权值 $w(\{m_1,m_2,\cdots,m_h\}')$ 的最大圈。

例如图 3 中,$Y = \{bdga\}$ 为权值 $\{2\}$ 的最大圈。

说明 3　为了描述方便,有时将 $w(m_i,m_j)$ 简写为 w。

2.2　算法过程

对于给定的形式背景 (O,M,I),构造概念格的过程如下:

输入　形式背景 (O,M,I) 以及 $W(R(O,M,I),w) = \{w_1, w_2, \cdots, w_n\}$, $w_r \neq w_s$, $r, s, n = 1, 2, \cdots, \frac{|M^2|}{2}$。

输出　所有概念 $\beta(O,M,I) \backslash \{(O,\varnothing),(\varnothing,M)\}$。

1) 对于 (O,M,I),绘制属性拓扑图,根据属性拓扑图中的箭头指向,确定顶层属性集合 T。

2) 将属性拓扑图转化为弱化的属性拓扑图。

3) ①初始将 $W(R(O,M,I),w)$ 赋值给 W_r,对任意 $w_i, w_j \in W_r$,求 $w_i \cap w_j$, $i,j = 1,2,\cdots, \frac{|M^2|}{2}$。若 $w_i \cap w_j = \varnothing$ 或 $w_i \cap w_j = w_t$,此处 w_i, w_j, $w_t \in W(R(O,$

$M,I),w)$, $i,j,t=1,2,\cdots,\frac{|M^2|}{2}$。则进行 4)。

②若 $w_i\cap w_j=w_t, w_t\notin W(R(O,M,I),w), i,j,t=1,2,\cdots,\frac{|M^2|}{2}$,则将 $W_{rr}=\{w_t;w_i\cap w_j=w_t,w_i,\ w_j,\in W(R(O,M,I),w),w_t\notin W(R(O,M,I),w)\}$ 添加到 $W(R(O,M,I),w)$ 中。将 W_{rr} 赋值给 W_t,执行①。

4)取 $\max|w_s|$,开始寻找边上权值包含 w_s 的最大圈,记录权值 w_s 最大圈的顶点为 Y,对应概念为 $C=\{(A,B):A=w_s,\ B=Y\}$。

5)①根据 4)的原则,若 W 中存在 $w_{s+1}, s=1,2,\cdots,|M|^2$,则选定 w_{s+1},重复 4)。

②若 W 中不存在 w_{s+1},则停止。

若 $W(R(O,M,I),w)$ 中的元素满足 3)中的①,若 $w_i\cap w_j=\varnothing$,或 $w_i\cap w_j=w_t$,此处 $w_i,\ w_j,w_t\in W(R(O,M,I),w)$, $i,j,t=1,2,\cdots,\frac{|M^2|}{2}$,则能够进行 4)、5),又因为 $|W(R(O,M,I),w)|$ 是有限的,因此有限步后算法可以停止。

若 $W(R(O,M,I),w)$ 中的元素满足 $w_i\cap w_j=w_t, w_t\notin W(R(O,M,I),w), i,j,t=1,2,\cdots,\frac{|M^2|}{2}$,此时会将新生成的 w_t 添加到 $W(R(O,M,I),w)$ 中,由于 $w_r\neq w_s, w_t\subseteq O$,$|O|$ 为有限的,因此经过有限步后一定可以进行 3)中①,因此有限步后算法可停止。

2.3 算法分析

根据文献[9],可以看出 1)的复杂度为 $O\left(\frac{|M|^2}{2}\right)$;步骤 2 将属性拓扑图弱化,首先判断每个属性是否为顶层属性,其复杂度为 $O(|M|)$,其次需要判断是否为不能构成权值 w 的最大圈,其复杂度为 $O\left(\frac{|M|^2}{2}\right)$,所以 2)的复杂度为 $O\left(\frac{|M|^2}{2}\right)$;若是 3)中①,首先对 W 中任意两元素取交,有 $O(|W|)$ 个元素,进行 $O\left(\frac{|W|^2}{2}\right)$ 次,若是 3)中②,对新生成的集合 W_{rr} 重复次 3),最多重复 $|O|$ 次,所以 3)的复杂度为 $O\left(\frac{|O||W_{rr}|^2}{2}\right)$,其中 $|W_{rr}|$ 为元素最多的集合;4)中,每到一个属性节点最多需要判断 $|M|-1$ 次该节点是否在当前权值 w 的最大圈中,最多判断 $|M|$ 次,所以 4)的复杂度为 $O(|M|^2)$;5)的复杂度为 $O(|O||W_{rr}|)$。

因此,整个算法的复杂度为 $O(2^{|M|\times|O|})$。

引理 2 圈有且仅有以下 3 种情况:

1)由一条边构成,也即自环;

2)由两条边构成,也即重边;

3)由 3 条或 3 条以上的边构成,也即非自环非重边的圈。

证明 由定义 2 中 3)可知,自环是只有一个顶点的圈;重边是由两个顶点的圈;由定义 2 中 3)可知,非自环非重边的圈之顶点个数大于 2。

当圈只有一个顶点时,根据定义 2 中 3)可知,此时的圈为一个自环;

当圈有两个顶点时,根据定义 2 中 3)可知,此时的圈为重边;

当圈的顶点个数大于 2 时,符合定义 2 中 2)。

因此,圈有且仅有自环、重边、非自环非重边的圈 3 种情况。

定理 1 设 (O,M,I) 是一个形式背景,$(R(O,M,I),w)$ 是 (O,M,I) 对应的弱化的属性拓扑图,权值最大圈一定能够生成一个概念。

证明 由引理 2 可知,弱化的属性拓扑图的权值 w 最大圈有且仅有 3 种情况,下面关于这 3 种情况分别说明。

1)当圈为自环时

$m\in M$,圈 $Y=\{m\}$,由弱化的属性拓扑图的构造 1)可知,有 $m\in T$。根据引理 1,$(m',\ m)\in\beta(O,M,I)$。而 $m'=w(m,m)$,因此,$(w(m,m),\ m)\in\beta(O,M,I)$。

2)当圈为重边时

$m_1,m_2\in M$,圈 $Y=\{m_1,m_2\}$,由弱化的属性拓扑图的构造 2)可知,不存在其他权值为 $w(m_1,m_2)$ 的边,即不存在其他顶 $m_i,\ m_j\in M$,使 $w(m_1,m_2)\subseteq w(m_i,\ m_j)$, $i\neq j$, $i\geqslant 3$, $j\geqslant 1$。这就是说,除 $m_1,\ m_2$ 外,不存在其他属性所拥有的对象集包含 $w(m_1,m_2)$,所以 $(w(m_1,m_2),Y)\in\beta(O,M,I)$。

3)= 当圈的顶点个数大于等于 3 时

$m_1,m_2,\cdots,m_i\in M, i\geqslant 3$,若圈 $Y=\{m_1,m_2,\cdots m_i\}$,证明过程与第 2 种情况类似,易证 $(w(m_1,m_2),Y)\in\beta(O,M,I)$。

引理 3 设 $W(R(O,M,I),w)$ 是一个集族,则任意的其中 $w_i,\ w_j,\ w_t\in W(R(O,M,I),w), w_u\notin W(R(O,M,I),w), i,j,t,u=1,2,\cdots,\frac{|M^2|}{2}$,它们之间的关系有且仅有以下 3 种情况之一发生:

1)$w_i\cap w_j=\varnothing$;

2)$w_i\cap w_j=w_t$;

3)$w_i\cap w_j=w_u$。

证明 根据文献[15],可得若 $W(R(O,M,I),$

w)是一个集族,则对于任意的 w_i, w_j, $w_t \in W(R(O,M,I),w)$, $w_u \notin W(R(O,M,I),w)$, $i,j,t,u=1,2,\cdots,\frac{|M^2|}{2}$ 有且仅有 $w_i \cap w_j = \varnothing$、$w_i \cap w_j = w_t$ 或 $w_i \cap w_j = w_u$,3 种情况之一发生:

定理 2　设(O,M,I)是一个形式背景,$(R(O,M,I),w)$是(O,M,I)对应的弱化的属性拓扑图,通过权值 w 最大圈算法一定能够得到 $\beta(O,M,I)\setminus\{(O,\varnothing),(\varnothing,M)\}$。

证明　由引理 3 可知集族 $W(R(O,M,I),w)$中的权值之间有且仅有 3 种情况,对于任意的 w_i, w_j, $w_t \in W(R(O,M,I),w)$, $w_u \notin W(R(O,M,I),w)$, $i,j,t,u=1,2,\cdots,\frac{|M^2|}{2}$ 以下对引理 3 中的 3 种情况分别说明。

1) $w_i \cap w_j = \varnothing$

说明任意 3 个属性之间没有公共对象,根据属性拓扑图的构造过程,其弱化的属性拓扑图为定理 1 的第 2 种情况,得到的 $W(R(O,M,I),w)$能够包括所有概念的外延,因此,依次搜索 $W(R(O,M,I),w)$中的每一个权值 w 最大圈,即可得到$\beta(O,M,I)\setminus\{(O,\varnothing),(\varnothing,M)\}$。

2) $w_i \cap w_j = w_t$, $w_t \in W(R(O,M,I),w)$, $i,j,t=1,2,\cdots,|M|^2$,说明得到的 $W(R(O,M,I),w)$能够包括所有概念的外延,符合定理的第 2、3 种情况。因此,依次搜索 $W(R(O,M,I),w)$中的每一个权值 w 最大圈,即可得到$\beta(O,M,I)\setminus\{(O,\varnothing),(\varnothing,M)\}$。

3) 若 $w_i \cap w_j = w_t$, $w_t \notin W(R(O,M,I),w)$, $i,j,t=1,2,\cdots,|M|^2$,则说明当前 $W(R(O,M,I),w)$中不能包含所有概念的外延,将 $W_r=\{w_t : w_i \cap w_j = w_t, w_t \notin W(R(O,M,I),w)\}$ 添加到 W 中,只需对 W_r中的任意两个元素取交集即可,由于$|O|$是有限的,因此一定会有 $w_i \cap w_j = w_t$, $w_t \in W(R(O,M,I),w)$, $i,j,t=1,2,\cdots,|M|^2$,此时说明得到的 $W(R(O,M,I),w)$能够包括所有概念的外延,并且 $W(R(O,M,I),w)$中的元素对应的最大圈符合定理 1 的第 2、3 种情况。因此,依次搜索 $W(R(O,M,I),w)$中的每一个权值 w 最大圈,即可得到$\beta(O,M,I)\setminus\{(O,\varnothing),(\varnothing,M)\}$。

以上说明了本算法的正确性。下面将通过与已有的图论方法构造概念格的相关著名算法或方法的比较,分析得出本算法的优势。

1)张涛等[14]在属性拓扑的基础上给出概念计算方法,实际上是将图论中已有的深度优先算法应用于概念的寻找,如此可能导致产生冗余概念。冗余概念的产生必然导致算法的储存空间的增加,引发空间复杂度的加大。

本文算法用图论中的权值与最大圈结合来寻找概念,由于不会重复对同一权值寻找其相应的最大圈,因此不会有冗余概念的产生。从而,必然在概念寻找中,降低数据的储存空间,空间复杂度较张涛等的方法减少成为显然之事。

2)Berry 等[10]将形式背景构造成二部图,利用团的思想生成概念,其计算每个概念的复杂度为 $O(|M|^2)$。

对于弱化的属性拓扑图产生概念有:

1)对于只含有一个顶点属性的情况,由引理 1 可知,属性 m 为某个概念的内涵。在弱化的属性拓扑图中属性 m 构成一个自环时,本文计算每个概念的复杂度 $O\left(\frac{|M|^2}{2}\right)$。

2)对于至少含有两个顶点属性的情况,本文计算每个概念的算法复杂度 $O(|M|^2)$。

以上两种情形说明,当情形 1)时,本文算法的复杂度小于 Berry 等的;当情形 2)时,本文算法的复杂度与 Berry 等的相同。这说明,对于情形 1),本文的算法优于 Berry 等的,虽然在其他情况(也就是情形 2)),本文的算法与 Berry 等的具有相同的时间复杂度。

再有,由图论知识可知,每一个团必包含至少一个圈,所以在判断团的过程中必然存在对圈的判断过程,当一个团中含有两个以上圈时,此时对团的判定过程会重复圈的判断过程。因此,Berry 等的方法会造成数据存储量过大。而本文算法,不会对相同权值的圈进行重复判断与存储,因此,降低了数据存储空间复杂度。

3)李立峰等[11]仅是从理论方面指出弦二部图的概念格表示,并没有给出算法过程。所以他们的方法只是理论过程,而无法直接实现。

本文中不仅给出了理论分析,并且将理论的内容通过一个可行的算法加以实现,故此,本文的思想和方法可操作性强,易于直接理解与实现。

由以上 1)~3)的分析可以看出,本文给出的算法与已有算法相比,计算出全部概念的时间复杂度并不低于以往的算法,基本相同。在数据存储空间方面,本文给出的算法与已有算法相比,空间复杂度降低。这样必然使得本算法在整个计算过程能在占用更小内存的情况下完成,同时也就对计算机器系统的运行空间降低了要求。因此本文算法要优于其他一些已有算法或方法。

3 实例

结合实例,说明第 2.2 节中的算法有效性。以表 2 为形式背景(O_1,M_1,I_1),进行概念的搜索,该背景从 UCI 机器学习数据库中,随机选取 BLOGGER 数据集的前 40 个对象进行试验,整理后得到如表 2 所示的形式背景。

表 2 (O_1,M_1,I_1)

Table 2 Formal context (O_1,M_1,I_1)

对像	a_1	a_2	a_3	a_4	a_5	a_6	a_7
1	×			×			×
2		×			×		×
3	×					×	×
4		×				×	×
5		×		×			×
6			×			×	
7	×					×	
8			×	×			×
9	×			×			
10		×			×		

其中,a_1代表博主高学历,a_2代表博主中等学历,a_3代表博主低学历,a_4代表政治立场为左派,a_5代表政治立场中立,a_6代表政治立场为右派,a_7代表博客被当地媒体转载。

根据 1),按照定义 3 中 1)得到以表 2 为形式背景的属性拓扑图,如图 4。根据 2),按照定义 4,构造出弱化的属性拓扑图($R(O_1,M_1,I_1),w$),如图 5。

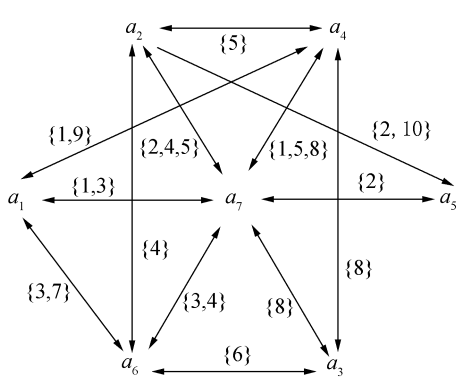

图 4 ($A(O_1,M_1,I_1),w$)

Fig.4 ($A(O_1,M_1,I_1),w$)

根据 3),$W=\{w(a_1,a_1),\ w(a_2,a_2),\ w(a_3,a_3),\ w(a_4,a_4),\ w(a_6,a_6),\ w(a_7,a_7),w(a_1,a_4),\ w(a_1,a_6),\ w(a_1,a_7),\ w(a_2,a_4),\ w(a_2,a_5),\ w(a_2,a_6),\ w(a_2,a_7),\ w(a_3,a_4),\ w(a_3,a_6),\ w(a_4,a_7),\ w(a_5,a_7),\ w(a_6,a_7)\}$。

对于任意两个 $w(a_i,a_j)$求交集,$i,j=1,2,\cdots,7$,根据 3)中②,可以发现存在 $w(a_1,a_1)\cap w(a_4,a_7)=w(a_1,a_4)\cap w(a_4,a_7)=\cdots=\{1\}$及 $w(a_1,a_1)\cap w(a_6,a_7)=w(a_1,a_7)\cap w(a_6,a_7)=\cdots=\{3\}$,$\{1\},\{3\}\notin W$,将$\{1\}$与$\{3\}$添加到 W,重复 3)。

对$\{\{1\},\{3\}\}$进行步骤 3)中②,$\{1\}\cap\{3\}=\varnothing$,进行 4)。

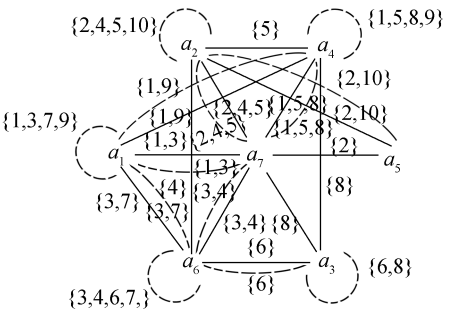

图 5 ($R(O_1,M_1,I_1),w$)

Fig.5 ($R(O_1,M_1,I_1),w$)

根据 4),$W=\{w(a_1,a_1),\ w(a_2,a_2),\ w(a_3,a_3),\ w(a_4,a_4),\ w(a_6,a_6),\ w(a_7,a_7),w(a_1,a_4),\ w(a_1,a_6),\ w(a_1,a_7),\ w(a_2,a_4),\ w(a_2,a_5),\ w(a_2,a_6),\ w(a_2,a_7),\ w(a_3,a_4),\ w(a_3,a_6),\ w(a_4,a_7),\ w(a_5,a_7),\ w(a_6,a_7),\ \{1\},\{3\}\}$。

因为 $6=|w(a_7,a_7)|\geqslant 4=|w(a_1,a_1)|=|w(a_2,a_2)|=|w(a_4,a_4)|=|w(a_6,a_6)|\geqslant 3=|w(a_2,a_7)|=|w(a_4,a_7)|\geqslant 2\ \ |w(a_1,a_4)|=|w(a_1,a_6)|=|w(a_1,a_7)|=|w(a_2,a_5)|=|w(a_3,a_3)|=|w(a_6,a_7)|\geqslant 1=|w(a_2,a_6)|=|w(a_2,a_4)|=|w(a_3,a_4)|=|w(a_3,a_6)|=|w(a_5,a_7)|=|\{1\}|=|\{3\}|$,所首先寻找包含 $w(a_7,a_7)=\{1,2,3,4,5,8\}$的最大圈,$Y_1=\{a_7\}$,对应的概念为(1 2 3 4 5 8,$a_7$)。

根据 5)中①,依次选择 W 中的其他元素重复 4),概念分别为(1 3 7 9,a_1),(2 4 5 10,a_2),(1 5 8 9,a_4),(3 4 6 7,a_6),(2 4 5,a_2 a_7),(1 5 8,a_4 a_7),(1 9,a_1 a_4),(3 7,a_1 a_6),(1 3,a_1 a_7),(2 10,a_2 a_5),(6 8,a_3),(3 4,a_6 a_7),(4,a_2 a_6 a_7),(5,a_2 a_4 a_7),(8, a_3 a_4 a_7),(6,a_3 a_6),(2, a_2a_5 a_7),(1, a_1 a_4 a_7),(3, a_1a_6 a_7)。

根据 5)中②,停止。

最后添加(1 2 3 4 5 6 7 8 9 10,∅)和(∅,a_1a_2 a_3 a_4 a_5 a_6 a_7)后,得到概念格$\beta(O_1,M_1,I_1)=\{$(1 2 3 4 5 6 7 8 9 10,∅),(1 2 3 4 5 8,a_7),(1 3 7 9,a_1),(2 4 5 10,a_2),(1 5 8 9,a_4),(3 4 6 7,a_6),(2 4 5,a_2 a_7),(1 5 8,a_4 a_7),(1 9,a_1a_4),(3 7,a_1 a_6),(1 3,a_1 a_7),(2 10,a_2 a_5),(6 8,a_3),(3 4,a_6 a_7),(4,a_2 a_6 a_7),(5,a_2 a_4 a_7),(8, a_3 a_4 a_7),(6,a_3 a_6),(2, a_2a_5 a_7),(1, a_1a_4 a_7),(3, a_1a_6 a_7) ,(∅,

$a_1a_2\ a_3\ a_4\ a_5\ a_6\ a_7)\}$。

在本实例中,步骤 2),弱化属性拓扑图,以 $w(a_1,a_1)$为例进行复杂度计算,判断是否有 $w(a_1,a_1)\subseteq w(a_i,a_j)$,其中 $i,j=1,2,\cdots,16$,对于 $w(a_1,a_1)$进行 16 次比较可得 a_1为顶层属性。对每个 $w\in W$ 重复上述比较过程,可得到弱化的属性拓扑图。

3)对任意两个元素 $w_i,w_j\in W,i,j=1,2,\cdots,16$,取交集,此过程需要进行 $\frac{18\times 17}{2}$ 次,得到$W\cup\{1\}\cup\{3\}$,对$\{\{1\},\{3\}\}$执行步骤 3)中②,此过程进行 1 次,可以看出符合 3)中①,可转 4)。

4)取 W 中的元素 $w(a_7,a_7)$,判断 $w(a_7,a_7)\subseteq w(a_i,a_j)$ 其中 $i,j=1,2,\cdots,20$,每个元素比较 20 次,寻找最大圈,得到概念(1 2 3 4 5 8,a_7)。

5)依次取 W 中的其他元素,重复 4),在此例 W 中的 18 个元素,4)需要重复 18 次。

在复杂度上,本文算法与张涛等的算法相同。并且利用张涛等的算法对(O_1,M_1,I_1)进行概念的计算,得到的概念格与本文算法的结果相同。从而说明了本文算法的有效性与正确性。

4　结束语

本文结合图论的知识,将形式背景对应的属性拓扑图弱化,提出了一种利用权值最大圈寻找概念的算法。与现有的算法比较,本文提出一种新的思路来搜索概念,此外通过弱化的属性拓扑图,对于概念的可视化也得到了很好的体现;通过实例可知,该方法能够有效地构造概念格,为知识获取和数据处理提供了一种有益的思想。通过分析可知,虽然本文提出的算法产生全部概念的空间复杂度降低,但由于其时间复杂度仍为指数级,因此对于数据量较大的情况,计算时间方面需要进一步研究,以便提高应用其进行数据分析的效率。

参考文献:

[1]WILLE R. Restructuring lattice theory: an approach based on hierarchies of concepts[M]//RIVAL I. Ordered Sets. Dordrecht: Springer, 1982.

[2]BELOHLAVEK R, SIGMUND E, ZACPAL J. Evaluation of IPAQ questionnaires supported by formal concept analysis [J]. Information sciences, 2011, 181(10): 1774-1786.

[3]NGUYEN T T, HUI S C, CHANG Kuiyu. A lattice-based approach for mathematical search using formal concept analysis[J]. Expert systems with applications, 2012, 39(5): 5820-5828.

[4]王旭杨, 李明. 基于概念格的数据挖掘方法研究[J]. 计算机应用, 2005, 25(4): 827-829.
WANG Xuyang, LI Ming. Method of data mining based on concept lattice[J]. Computer applications, 2005, 25(4): 827-829.

[5]SIFF M, REPS T. Identifying modules via concept analysis [C]//Proceedings of International Conference on Software Maintenance. Washington, DC, USA: IEEE Computer Society, 1997: 170-179.

[6]FERJANI F, ELLOUMI S, JAOUA A, et al. Formal context coverage based on isolated labels: an efficient solution for text feature extraction [J]. Information sciences, 2012, 188: 198-214.

[7]邓君, 马晓君, 张巨峰, 等. 基于概念格的实体档案馆用户行为研究[J]. 图书情报工作, 2014, 58(12): 109-117.
DENG Jun, MA Xiaojun, ZHANG Jufeng, et al. Study on entity archives' user behavior based on concept lattice[J]. Library and information service, 2014, 58(12): 109-117.

[8]张晓, 龙伟, 卢斌. 基于概念格的关联规则在排产管理的应用[J]. 计算机工程与应用, 2014, 50(9): 264-270. v ZHANG Xiao, LONG Wei, LU Bin. Application of association rule based on concept lattice for scheduling management[J]. Computer engineering and applications, 2014, 50(9): 264-270.

[9]张涛, 任宏雷. 形式背景的属性拓扑表示[J]. 小型微型计算机系统, 2014, 35(3): 590-593.
ZHANG Tao, REN Honglei. Attribute topology of formal context[J]. Journal of Chinese computer systems, 2014, 35(3): 590-593.

[10]BERRY A, SIGAYRET A. Representing a concept lattice by a graph[J]. Discrete applied mathematics, 2004, 144(1/2): 27-42.

[11]李立峰, 刘三阳, 罗清君. 弦二部图的概念格表示[J]. 电子学报, 2013, 41(7): 1384-1388.
LI Lifeng, LIU Sanyang, LUO Qingjun. Representing chordal bipartite graph using concept lattice theory[J]. Acta electronica sinica, 2013, 41(7): 1384-1388.

[12]DAVEY B A, PRIESTLEY H A. Introduction to lattices and order[M]. 2nd ed. New York: Cambridge University Press, 2002: 66-69.

[13]王树禾. 图论[M]. 北京: 科学出版社, 2009.

[14]ZHANG Tao, REN Honglei, WANG Xiaomin. A calculation of formal concept by attribute topology[J]. ICIC express letters part B: applications, 2013, 4(3): 793-800.

[15]方嘉琳. 集合论[M]. 长春: 吉林人民出版社, 1982.

作者简介:

毛华,女,1963 年生,教授,博士,主要研究方向为计算机数学及其应用、拟阵理论、离散数学。发表学术论文 90 余篇,其中被 SCI、EI 检索 20 余篇。