# 约束概念格及其构造方法

张继福[1,2] , 张素兰[1] ,胡立华[1]
(1. 太原科技大学 计算机科学与技术学院 ,山西 太原 030024; 2. 中国科学院自动化所 模式识别国家重点实验室 ,北京 100080)

**摘 要**:概念格是一种有效的数据分析和知识提取的形式化工具. 然而 ,随着要处理的数据量的剧增 ,基于原始形式背景构造出的概念格结点数目庞大 ,占用大的存储空间 ,同时概念格结点中一些属性集形成的内涵 ,用户并不都感兴趣 ,因而从中提取用户需求知识费时. 为了降低概念格构造的时空复杂性 ,增强实用性和针对性 ,首先采用谓词逻辑描述用户感兴趣的背景知识 ,并将背景知识引入到概念格结构中 ,提出了一种新的概念格:约束概念格. 在此基础上 ,提出了基于背景知识的约束概念格构造算法 CCLA. 理论分析表明 ,该算法能有效地减少概念格的存储空间和建格时间. 最后 , 采用恒星天体光谱数据作为形式背景 ,实验验证了该算法的有效性.

**关键词**:数据挖掘;约束概念格;谓词逻辑;背景知识;恒星光谱数据
**中图分类号**: TP311 **文献标识码**: A **文章编号**:1673-4785(2006)02-0031-08

# Constrained concept lattice and its construction method

ZHANG Ji-fu[1,2] , ZHANG Su-lan[1] , HU Li-hua[1]
(1. School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China; 2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China)

**Abstract:** Concept lattice is an effective formal tool for data analysis and knowledge mining. However, with the increase of data volume, the node number of the constructed concept lattice from the original formal context usually increases enormously, and large storage is required accordingly. Meantime, users are not interested in all intensions of attributes set, and more computational time is unnecessarily consumed as a result. In order to reduce time and storage complexity and improve the utility and pertinence to the concept lattice construction, predicate logic is used to describe the user interested background knowledge, and a new concept lattice structure-constrained concept lattice is presented. Then based on the background knowledge, a construction algorithm (CCLA) is also provided. Through some theoretical analysis, it is shown that the proposed algorithm can reduce the storage and time complexity of concept lattice construction process. Finally, the experiments with celestial body spectra as the formal context validate the proposed algorithm.

**Keywords:** data mining; constrained concept lattice; predicate logic; background knowledge; star spectra data

概念格是一种有效的形式化数据分析工具 ,由德国的 R. Wille 教授在 20 世纪 80 年代初提出[1]. 概念格的每个结点是一个形式概念 ,由内涵(属性集)和外延(拥有该属性集的实体集)两部分组成. 这种格的结构及其相应的哈希图形式 ,反映了一种概念层次结构 ,本质上体现了实体(对象、记录、交易)和属性(特征、项目)之间的关系. 概念内涵和外延的统一 ,生动而简洁地表明了概念之间的泛化和特化关系 ,成为一种很有用的数据分析和知识提取工具. 这种形式概念分析工具已经被成功地用于数字图书馆、文献检索、软件工程、基于案例数据分析、知识发现等领域[2-4].

目前 ,国内外学者对概念格进行了多方面深入研究:概念格的构造算法研究;基于概念格的知识提取(数据分类、聚类及关联规则提取);概念格与其他

收稿日期:2006-02-15.
基金项目:国家自然科学基金资助项目(60573075).

理论的融合(粗集、遗传算法和模糊理论)等[5-9];其中,概念格的结构和构造效率始终是研究的重点,先后提出了许多种格结构及其构造算法[10-12].

为了提高概念格的构造效率,减少时空复杂性,增强实用性和针对性,首先采用谓词逻辑作为用户感兴趣的背景知识,将背景知识引入到概念格结构中,提出了一种新的概念格:约束概念格. 在此基础上,提出了一种基于背景知识的约束概念格构造算法 CCLA,理论证明了该算法能有效地节省概念格的存储空间和建格时间. 最后采用天体专家知识作为背景知识,恒星天体光谱数据作为形式背景,构造了约束概念格,从而验证了约束概念格构造算法的有效性.

## 1 问题的提出

概念格是一种有效的数据挖掘和知识提取的形式化分析工具,数据挖掘是在积累了巨量数据集后,从中挖掘出有效的、新颖的、潜在有用的、最终可理解并加以有目的的利用知识的过程,是从宏观角度利用积累的巨量数据进行知识抽象的高级阶段. 可以看出数据挖掘是一项高级的智能活动,因此数据挖掘的过程离不开背景知识的支持. 目前将背景知识融合在数据挖掘过程中的研究还处于初始阶段,因而使得数据挖掘技术在实际应用中受到了一定的限制[12-13]. 以用户提供的背景知识(感兴趣、不感兴趣)为指导形成概念格,不仅有利于挖掘出用户感兴趣的知识,而且也可以减少概念格构造的时空复杂性.

谓词逻辑是一种形式语言系统,它用逻辑方法研究推理的规律,适合于表示事物的状态、属性、概念等事实性的知识,也可以用来表示事物之间确定的因果关系,即规则. 因此具有自然性、精确性、严密性和容易实现等优点,是一种广泛使用的知识表示技术. 采用谓词逻辑作为表示指导概念格构造的用户感兴趣的背景知识是可行的.

然而,一般概念格都是基于形式背景进行构造的,一些属性组合成的概念格内涵,用户并不都感兴趣,例如利用概念格从海量天体数据中挖掘分类知识时,从原始的形式背景(光度、温度)中由属性光度、温度组合形成的概念格的内涵对7类恒星光谱数据的分类就无任何指导意义,因此,在概念格的构造过程中,用户对含有这些属性组成的内涵是不感兴趣的. 同时,基于形式背景构造出含有所有属性组合成内涵的结点明显存在以下不足:构造的结点数目庞大,占用大的存储空间,基于这些概念格提取有用知识(关联规则、分类规则、聚类规则等)费时,特别随着要处理的数据量的激增,这些不足日益明显. 所以,基于背景知识的概念格构造研究无论在理论上还是在实际应用上都具有重要的意义.

## 2 一般概念格

**定义1** 给定一个形式背景为三元组 $T=(U,I,R)$,其中 $U$ 为对象集,$I$ 为属性集,$R$ 是 $U$ 与 $I$ 之间存在的一个二元偏序关系. 由这个二元偏序关系可以形成一个概念格 $L$.

**定义2** 概念格的每一个结点为一个形式概念 $h=(O,D)$,其中,$O\in\rho(U)$ 称为概念的外延,$D\in\rho(I)$ 称为概念的内涵,$D$ 是由 $O$ 中对象(记录、交易)的共同特征(属性、项目)所组成的集合. 具有这种结构的格称为一般概念格(General concept lattice).

**定义3** $(O,D)$ 关于 $R$ 满足完备性 $\Leftrightarrow\forall O\subseteq U: f(O)=\{d\in I\mid\forall x\in O: xRd\}$ 和 $\forall D\subseteq I: g(D)=\{x\in U\mid\forall d\in I: xRd\}$ 同时成立.

**定义4** 设 $h_1=(O_1,D_1)$ 和 $h_2=(O_2,D_2)$ 是2个不同的结点,则 $h_1<h_2\Leftrightarrow D_2\subset D_1\Leftrightarrow O_1\subset O_2$,如果不存在 $h_3=(O_3,D_3)$ 有 $h_1<h_3<h_2$ 成立,则 $h_2$ 称为 $h_1$ 的父结点(父概念,直接前趋),$h_1$ 称为 $h_2$ 的子结点(子概念,直接后继).

表1是一个形式背景,其中对象集 $U=\{1,2,3,4,5\}$,属性集 $I=\{A,B,C,D,E\}$,$R$ 描述了 $U$ 中所具有的 $I$ 中的属性值集,该形式背景所构成的一般概念格如图1所示.

**表1 形式背景**

**Table 1 Formal context**

| $U$ | $I$ | $A$ | $B$ | $C$ | $D$ | $E$ |
|---|---|---|---|---|---|---|
| 1 | | √ | √ | | | |
| 2 | | | √ | | √ | √ |
| 3 | | | √ | √ | | |
| 4 | | √ | | | √ | |
| 5 | | | | √ | | |

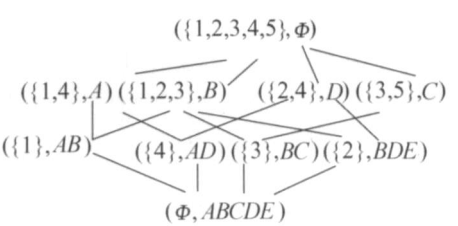


**图1 一般概念格**

Fig. 1 General concept lattice

## 3　面向概念格构造的背景知识

在概念格的构造过程中,由所有属性组成的内涵并非都是用户感兴趣的,同时一些属性组成的内涵在实际应用中并无意义.因此,可以根据用户对数据集的兴趣、了解、认识等作背景知识为指导来构成概念格,从而使概念格的结构更具有针对性和实用性.采用谓词逻辑表示知识时,首先定义描述背景知识的谓词,并指出每个谓词的确切含义,然后再用连接词( ∧(与)、∨(或)、¬(非)、→(蕴含)、∀(全称量词)、∃(存在量词))把有关的谓词连接起来,形成一个谓词公式以表达一条完整的背景知识.

一个二维表可表示为一个 $n$ 元有序组的集合,一个集合可用一个特性谓词刻画,故一个 $n$ 元有序组的集合可用一个 $n$ 元特性谓词刻画.基于一阶谓词逻辑的概念格构造中所涉及到的背景知识描述如下:

**定义** 5　一个格结点集合 $G(z)$ 为一个一元谓词,表示 $z$ 是一个格结点.

**定义** 6　Concept$(z,x,y)$ 为一个三元谓词,表示格结点 $z$ 具有内涵 $x$,外延 $y$.

**定义** 7　Include$(x,y)$ 表示由某属性集 $y$ 组成的内涵 $x$.

**定义** 8　Interest$(z)$ 为一个一元谓词,表示 $z$ 是一个关心结点.

概念格结点的内涵由属性组成,在实际的概念格构造中,用户往往对含有某些属性组合的内涵和不含有某些属性组合的内涵感兴趣,因此,可将背景知识分为 2 类知识:其中内涵由用户关心的含有某些属性集合组成的知识定义为第 I类背景知识,内涵由用户关心的不含有某些属性集合组成的知识定义为第 II类背景知识.

**定义** 9　设 $P_1(Z) = \forall z((G(z) \wedge \text{concept}(z,x,y) \wedge \text{include}(x,y_0)) \rightarrow \text{interest}(z))$ 为一谓词公式,$P_1(Z)$ 表示一个格结点,如果其内涵 $x$ 是由用户关心的属性子集 $y_0$ 组成,则该结点为关心结点.

**定义** 10　设 $P_2(Z) = \forall z((G(z) \wedge \text{concept}(z,x,y) \wedge (\text{include}(x,y_0) \wedge \text{include}(x,y_1))) \rightarrow \text{interest}(z))$ 为一谓词公式,$P_2(Z)$ 表示一个格结点,如果其内涵 $x$ 是由用户关心的属性子集 $y_0$、$y_1$ 组成,则该结点为关心结点.

**定义** 11　$P_3(Z) = \forall z((G(z) \wedge \text{concept}(z,x,y) \wedge (\text{include}(x,y_0) \vee \text{include}(x,y_1))) \rightarrow \text{interest}(z))$ 为一谓词公式,$P_3(Z)$ 表示一个格结点,如果其内涵 $x$ 是由用户关心的属性子集 $y_0$ 或 $y_1$ 组成,则该结点为关心结点.

**性质** 1　谓词公式 $P_1$、$P_2$ 和 $P_3$ 描述了第 I类背景知识.

**证明**　通过 ∧(与)、∨(或)运算所形成的谓词公式 $P_1$、$P_2$ 和 $P_3$ 描述了这样一类格结点,其内涵是由用户关心的含有某属性集合组成,因此描述了第 I类背景知识.

以表 1 的形式背景为例,若用户关心的概念格结点的内涵含有属性 $A$、属性 $A$ 且 $B$、属性 $A$ 或 $B$,则形成的概念格如图 2～4 所示.

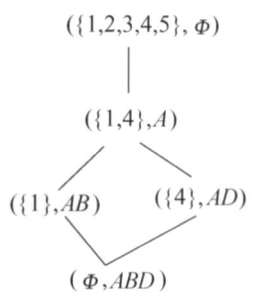


图 2　属性 $A$ 的概念格

Fig. 2　Concept lattice on attribute $A$

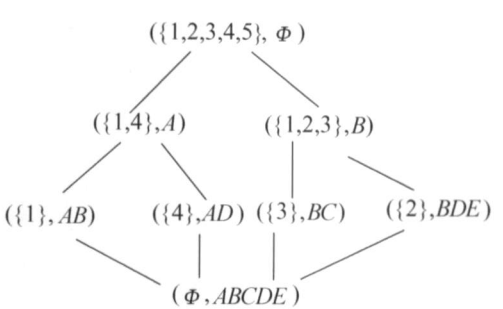


图 3　属性 $A \vee B$ 的概念格

Fig. 3　Concept lattice on attribute $A$ or $B$

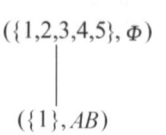


图 4　属性 $A \wedge B$ 的概念格

Fig. 4　Concept lattice on attribute $A$ and $B$

**定义** 12　设 $P_4(Z) = \forall z((G(z) \wedge \text{concept}(z,x,y) \wedge \neg\text{include}(x,y_0)) \rightarrow \text{interest}(z))$ 为一谓词公式,$P_4(Z)$ 表示一个格结点,如果其内涵是由用户关心的不含属性子集 $y_0$ 组成,则该结点为关心结

点.

**定义** 13 设 $P_5(Z) = \forall z((G(z) \wedge \text{concept}(z,x,y)) \wedge (\neg(\text{include}(x,y_0)) \wedge \neg(\text{include}(x,y_1))) \rightarrow \text{interest}(z))$ 为一谓词公式,$P_5(Z)$ 表示一个格结点,如果其内涵是由用户关心的不含属性子集 $y_0$ 且不含属性子集 $y_1$ 组成,则该结点为关心结点.

**定义** 14 设 $P_6(Z) = \forall z((G(z) \wedge \text{concept}(z,x,y)) \wedge (\neg(\text{include}(x,y_0)) \vee \neg\text{include}(x,y_1))) \rightarrow \text{interest}(z))$ 为一谓词公式,$P_6(Z)$ 表示一个格结点,如果其内涵是由用户关心的不含属性子集 $y_0$ 或者不含属性子集 $y_1$ 组成,则该结点为关心结点.

**性质** 2 谓词公式 $P_4$、$P_5$ 和 $P_6$ 描述了第 II类背景知识.

**证明**:通过 $\wedge$(与)、$\vee$(或)、$\neg$(非)运算所形成的谓词公式 $P_4$、$P_5$ 和 $P_6$ 描述了这样一类格结点,其内涵是由用户关心的不含有某属性集合组成,因此描述了第 II类背景知识.

以表 1 形式背景为例,图 5 为用户关心的概念格结点的内涵不含有属性 $A(\neg A)$ 的格结构,图 6 为用户关心的概念格结点的内涵不含有子集 $A$ 且不含有子集 $B(\neg A \wedge \neg B)$ 的格结构,图 7 为用户关心的概念格结点的内涵不含有属性子集 $A$ 或者不含有属性子集 $B(\neg A \vee \neg B)$ 的格结构.

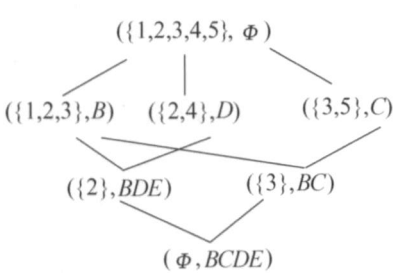


图 5 $\neg A$ 概念格

Fig. 5 Concept lattice on attribute $\neg A$

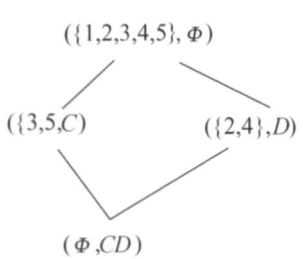


图 6 $(\neg A \wedge \neg B)$ 概念格

Fig. 6 Concept lattice on attribute $\neg A$ and $\neg B$

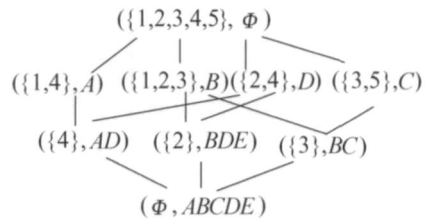


图 7 $(\neg A \vee \neg B)$ 概念格

Fig. 7 Concept lattice on attribute $\neg A$ or $\neg B$

**性质** 3 由 $P_1(Z)$、…、$P6(Z)$ 所构成的合式逻辑公式,描述了 2 类背景知识.

**证明**:由性质 1 和性质 2 可容易得证.

## 4 约束概念格及其构造

### 4.1 约束概念格

**定义** 15 概念格的每一个结点为一个形式概念 $h = ((O,D),P)$,其中:$P$ 是由 $P_1$、$P_2$、…、$P_6$ 所构成的合式逻辑公式,且 $P((O,D)) = .T.$(逻辑值为真),$O \in \rho(U)$ 称为概念的外延,$D \in \rho(I)$ 称为概念的内涵,$D$ 是由 $O$ 中满足 $P$ 的所有对象(记录、交易)的共同特征(属性、项目)组成的集合,具有这种结构的格称为约束概念格(constrained concept lattice).

**定义** 16 设 $h_1 = ((O_1,D_1),p)$ 和 $h_2 = ((O_2,D_2),P)$ 是约束概念格中的 2 个不同的结点,则 $h_1 < h_2 \Leftrightarrow D_2 \subset D_1 \Leftrightarrow O_1 \subset O_2$,如果不存在 $h_3 = ((O_3,D_3),P)$ 有 $h_1 < h_3 < h_2$ 成立,则 $h_2$ 称为 $h_1$ 的父结点(父概念,直接前趋),$h_1$ 称为 $h_2$ 的子结点(子概念,直接后继).

**定义** 17 在同一形式背景下,设 $h_1 = ((O_1,D_1),p)$ 是约束概念格中的一个结点,$h_2 = (O_2,D_2)$ 是一般概念格中的一个结点,如果 $D_1 \subset D_2$,$O_1 = O_2 = \Phi$,则称 $h_1$、$h_2$ 为 2 个等价结点.

**定义** 18 在同一形式背景下,如果约束概念格中的任一结点都为一般概念格中的一个结点,任一条边也都为一般概念格中的一条边,则称约束概念格为一般概念格的子格.

**引理** 1 当背景知识为第 I类时,约束概念格为一般概念格的子格.

**证明** 当背景知识为第 I类时,约束概念格中的任一结点内涵都是由用户关心的属性子集组成,且为一般概念格中的一个结点,由定义 4、定义 16

和定义 17 可知,约束概念格中的任一边也都为一般概念格中的一条边,由定义 18 可知,背景知识为第一类时,约束概念格为一般概念格的子格.

**引理 2**　当背景知识为第 II类时,约束概念格减少了一般概念格的中的节点数和边数.

**证明**　当背景知识为第 II类时,约束概念格中的任一结点内涵都是由用户关心的不含有某属性子集的集合组成,而一般概念格中的节点是所有原始形式背景中的属性集合,所以,约束概念格中的节点比一般概念格的节点少,同样由定义 4、定义 16 和定义 17 可知,约束概念格中的边也比一般概念格中的边少. 因此,当背景知识为第 II类时, 约束概念格减少了一般概念格中的节点数和边数.

**定理 1**　约束概念格比一般概念格的格结构的复杂性要低.

**证明**:由引理 1 和 2 容易得证.

**定理 2**　当 $P((O,D))$ 为空时,约束概念格退化为一般概念格.

**证明**:给定一个形式背景 $T=(U,I,R)$,一般概念格的任意结点 $h=(O,D)$,由于 $P$ 为空,则 $P(h)$ 为真,因而 $((O,D),P)$ 也是约束概念格的一个结点. 另外设 $h_1=(O_1,D_1)$ 和 $h_2=(O_2,D_2)$ 是一般概念格 2 个不同的结点,而且 $h_1<h_2$,由于 $P$ 为空,$((O_1,D_1),P)$ 和 $((O_2,D_2),P)$ 也是约束概念格中的两个不同的结点,由定义 4 和 16 可得 $((O_1,D_1),P)<((O_2,D_2),P)$. 因此,约束概念格退化为一般概念格.

**定理 3**　约束概念格是完备的.

**证明**:对一般概念格中的一个点,凡是满足 $P$ 的点都在约束概念格中,那么 $h=((O,D),P)$ 关于 $R$ 满足完备性 $\Leftrightarrow O\subseteq U: f(O)=\{d\in I \mid \forall x\in O: xRd\}$ 且 $P((O,D))=.T.$ 和 $\forall D\subseteq I: g(D)=\{x\in U \mid \forall d\in I: xRd\}$ 且 $P((O,D))=.T.$ 同时成立,显然约束概念格是完备的.

### 4.2　基于背景知识的约束概念格构造

当为第 I类背景知识时,即用户关心的属性集为 $X$、$X\vee Y$($X$ 或者 $Y$)或者 $X\wedge Y$($X$ 且 $Y$),则概念格结点的内涵只能为含有 $X$、$Y$、或者 $XY$ 的属性集. 由引理 1 可知,由用户关心的属性集形成的概念格为原概念格的子格,那么在概念格构造过程中,只需要对含有关心属性的对象进行概念格的渐进式构造.

当为第 II类背景知识时,即用户不关心的属性集为 $X$、$X\vee Y$($X$ 或者 $Y$)或者 $X\wedge Y$($X$ 且 $Y$),则概念格结点的内涵为不含有 $X$、$Y$、或者 $XY$ 的属性集,可用如下方法构造:

1) 将含有不关心属性的对象做上删除标志,不关心的属性值记为空值;

2) 渐进式生成概念格时,如果结点为空时,先生成不含有删除标志的对象的格结点;

3) 然后求解其与前面有删除标志对象结点的关系(交集). 由交的结果决定是否生成新结点. 若交集不为空时,则生成新结点,否则不做任何处理.

4) 再求解带删除标志对象结点间的关系(交集). 由交集的结果决定是否生成新结点. 若交集不为空时,则生成新结点,否则不做任何处理.

由上述算法的构造思想及相关定理,可给出如下约束概念格构造算法 CCLA.

**算法** CCLA (constrained concept lattice algorithm)

**输入**:原约束概念格 $L_w$,用户关心的属性子集 $X$、$X\vee Y$($X$ 或者 $Y$)、$X\wedge Y$($X$ 且 $Y$),用户不关心的属性子集 $\neg X$、$\neg X\vee\neg Y$(非 $X$ 或者非 $Y$)、$\neg X\wedge\neg Y$(非 $X$ 且非 $Y$).

**输出**: 更新后的约束概念格 $L_r$.

1) Mark: = $\Phi$;/ * Mark 约束概念格结点集合 */

2) 对渐进式追加的每个对象 $x$,

3) If 输入用户关心的属性子集 $X$、$X\wedge Y$ 或者 $X\vee Y$ Then

4) If 关心的属性子集为 $X$、或者 $X\wedge Y$ Then

5) If $X\in f(x)$ or $XY\in f(x)$ Then

6) Generate()

7) Else

8) If $X\in f(x)$ or $Y\in f(x)$ Then

9) Generate()

10) End if

11) End if

12) End if

13) Else

14) If 不关心的属性子集为 $\neg X$、或者 $\neg X\vee\neg Y$ Then

15) If $X \in f(x)$ or $XY \in f(x)$ Then

16) $f(x)$ 的属性值 $X$ 或者 XY 记为空值,原 $f(x) \rightarrow f(x^*)$

17) Generate()

18) End if

19) Else

20) If $X \in f(x)$、$Y \in f(x)$ Then

21) $f(x)$ 的属性值 $X$、$Y$ 记为空值,原 $f(x) \rightarrow f(x^*)$

22) Generate()

23) End if

24) End if

25) End if

26) End CCLA

27) (Generate()

28) For $L_r$ 中的每个格结点 $h_r = ((O, D), P)$ 按| $D$| 升序排列 $Do$/ * 更新约束概念 */

29) If $D \subseteq f(x)$ Then

30) $O := O \cup \{x\}$;

31) Mark := Mark $\cup \{h_r\}$;

32) If $D = f(x)$ Then 退出 For 循环; Else

33) 调用过程 Gennew();

34) End if

35) End for

36) End Generate()

37) Gennew()

38) If 输入用户关心的属性子集 $X$、$X \wedge Y$ 或者 $X \vee Y$ Then

39) inter := $D \cap f(x)$;

40) Else

41) inter := $D \cap f(x^*)$;

42) Endif

43) If inter $\neq \Phi$

44) If 用户输入的子集为 $X \vee Y$, Then

45) If $X \in$inter or $Y \in$inter Then/ *约束新增概念 */

46) If 不存在 $h_k \in$Mark 使得 $D_k$ := inter Then

47) $N_r = (O \cup \{x\}$, inter); / * $N_r$ 新增结点 */

48) End if

49) Mark := Mark $\cup N_r$;

50) 增加边 $h_x \leftarrow N_r$;

51) End if

52) For 对于 Mark 中的每个格结点 $h_m = (O', D', w')$ 按| $D'$| 降序排列 $Do$

53) If 存在 $h_m \in$Mark 使得 $D' \subset$inter Then 增加边 $N_w \leftarrow h_m$;

54) If 存在 $h_m$ 是 $h_x$ 的双亲 Then 删去边 $h_x \leftarrow h_m$;

55) End if

56) End for

57) End if

58) End Gennew()

**算法分析**:

在上述 CCLA 算法中,不但只根据新追加对象内涵与原格内涵的交集结果,而且还要根据用户关心和不关心的属性子集的组合决定格节点的生成,用户不感兴趣的属性子集的内涵的格结点将不被生成.

对于一个新对象 $h_x = \{x, f(x), w\}$,最多可能存在 $2^{f(x)}$ 个内涵包含于 $f(x)$ 的概念. 因此,当所有原始形式背景的行对象的属性都包含用户关心的属性集时,无节点被删除. 根据文献[8]建格算法分析可知,若设| $f(x)$| = $k$,则算法的复杂度为 $O(2^{2k}|U|)$. 而在实际应用中,对象 $x$ 具有的属性内涵并不都是用户感兴趣的,不包含用户关心的属性集的对象不进行渐进式构造,在实际概念格的构造中,随着要处理数据量的增大,| $U$| 的量减少,生成的节点数将明显减少,算法的复杂度要小于 $O(2^{2k}|U|)$. 同样地,在实际应用中,当生成用户定义的不含有某属性子集的内涵的格结点时,含有某属性子集的内涵的格结点将不被生成,随着要处理数据量的增大,$f(x)$ 即 $k$ 的量减少,因此,算法的复杂度远小于 $O(2^{2k}|U|)$. 所以,该算法能有效地节省概念格的存储空间和建格时间.

## 5 实验分析

当前我国正在建造一台大天区面积多目标光纤光谱望远镜(简称 LAMOST),它是国家"九五"计划重大工程项目,总投资达 2.35 亿人民币. 由于 LAMOST 具有以较高效率大规模测量天体光谱的能力,可提供的研究课题将遍及天文学多个层次,从恒星、银河系、星系、星系团、活动星系核,直到宇宙

大尺度结构. LAMOST 计划的主要目标是用来进行大规模光谱巡天,预计从 2006 年底起,每个观测夜晚将收集 2～4 万条光谱的数据,LAMOST 所观测到的光谱数据容量可达 4TG[15]. 如何利用数据挖掘技术从海量天体光谱数据中发现未知的、特殊的天体和天体规律是值得研究和探索的新应用领域.

在 PentiumIII-1. 0G CPU,256MB 内存,Windows2000 操作系统,DBMS 为 ORACLE9i,用 Visual Basic6. 0 实现了 CCLA 算法. 选用 2 500 条 M、K、G、F、A、B、O 等类恒星光谱数据为数据集,经过以下预处理后构成该实验中的形式背景:1) 选定间隔为 40 的 100 个波长 3 510,3 550, …, 8 330 Å,依据流量、峰宽和形状,将每个波长离散化为 13 种值;2) 恒星的任意类型温度等间隔离散化为 3 种,7 类恒星温度被离散化为 21 种值;3) 根据恒星的光度、化学丰度、微湍流、其他参数等间隔离散化为 3 种值;4) 根据恒星的物理参数,等间隔离散化为 5 种值.

如果用户对恒星温度和化学分度组成的内涵感兴趣,对物理 4 和微湍流 1 组成的内涵不感兴趣,其约束概念格的结点个数和建格时间如表 2、3 所示.

表 2　CCLA 算法实验结果 1

Table 2　The experiment result 1 of CCLA algorithm

| 背景知识 | 结点数 | 建格时间/ s |
|---|---|---|
| Φ(一般概念格) | 6 713 | 2 823 |
| 温度 $a$ ∨化学丰度 2 | 3 860 | 949 |
| 温度 $a$ | 1 002 | 121 |
| 化学丰度 2 | 3 351 | 851 |
| 温度 $a$ ∧化学丰度 2 | 473 | 71 |

表 3　CCLA 算法实验结果 2

Table 3　The experiment result 2 of CCLA algorithm

| 背景知识 | 结点数 | 建格时间/ s |
|---|---|---|
| Φ(一般概念格) | 6 713 | 2 823 |
| ¬物理 4 ∨¬微湍流 | 16 319 | 2 760 |
| ¬微湍流 1 | 6 090 | 2 479 |
| ¬物理 4 | 4 122 | 879 |
| ¬物理 4 ∧¬微湍流 1 | 3 665 | 732 |

由表 2、3 中的实验结果可以看出,1) 在相同形式背景下,约束概念格比一般概念格的结点数少,建格时间短. 所以,利用背景知识,算法 CCLA 可以有效地减少概念格结点数,节省概念格构造所用的时间. 因此,约束概念格利用用户提供的背景知识,可以降低概念格的时空复杂性. 2) 随着背景知识对概念格结点约束程度的加大,其满足约束的格结点个数和相应的建格时间将明显减低. 3) 由于背景知识描述了用户感兴趣或不感兴趣的内涵,因此约束概念格提高了概念格的针对性和实用性.

## 6　结束语

在应用概念格进行知识提取时,一些概念格内涵的属性组合,并非用户都感兴趣. 同时,随着要处理的数据量的激增,概念格的构造时间长,占用大的存储空间. 因此,为了提高概念格的构造效率,减少建格的时间复杂度和空间复杂度,提出了一种全新的概念格结构:基于背景知识的约束概念格. 利用背景知识来指导概念格的构造,使构造出的概念格更具有针对性和实用性. 进一步工作是基于约束概念格的知识提取及在天体光谱数据知识发现中的应用.

## 参考文献:

[1] WILLE R. Restructuring lattice theory: an approaches based on hierarchies of concepts[A]. In: Rival I ed. Ordered Sets[M]. Dordrecht: Reideal, 1982.

[2] GODIN R, MISSAOUI R. An Incremental concept formation approach for learning from databases [J]. Theoretical Computer Science, 1994, 133(2): 387 - 419.

[3] BELEN D A, PEDRO A. Formal concept analysis as a support technique for CBR[J]. Knowledge-based Systems, 2001, 14(3): 163 - 171.

[4] YOUNG P. Software retrieval by samples using concept analysis [J]. The Journal of Systems and Computer, 2000, 54(3): 179 - 183.

[5] CHU S, CESNIK B. Knowledge representation and retrieval using conceptual graphs and free document self-organisation techniques[J]. International Journal of Medical Informatics, 2001, 62(2/3): 121 - 33.

[6] KENT R E. Rough concept analysis[A]. In: Ziarko W P ed. Rough Sets, Fuzzy Sets and Knowledge Discovery

(RSKD,93)[C]. London : Springer-Verlag , 1994.

[7] CARPINETO C, ROMANO G. A lattice conceptual clustering system and its application to browsing retrieval [J]. Maching Learning, 1996,24(2):95 - 122.

[8]王志海,胡可云,胡学钢,等. 概念格上规则提取的一般算法与渐进式算法[J]. 计算机学报,1999,22(1):66 - 70.
WANG Zhihai, HU Keyun, HU Xuegang, et al. General and incremental algorithms of rule extraction based on concept lattice[J]. Chinese Journal of Computers, 1999, 22(1):66 - 70.

[9]陈世权,程里春. 模糊概念格[J]. 模糊系统与数学, 2002, 16(4):12 - 18.
CHEN Shiquan, CHENG Lichun. Fuzzy concept lattice [J]. Fuzzy Systems and Mathematics, 2002, 16(4): 12 - 18.

[10] NOURINE L, RAYNAUD O. A fast algorithm for building lattices [J]. Information Processing Letters, 1999,71(5 - 6):199 - 204.

[11]谢志鹏,刘宗田. 概念格的快速渐进式构造算法[J]. 计算机学报, 2002, 25(5):490 - 495.
XIE Zhipeng, LIU Zongtian. A fast incremental algorithm for building concept lattice[J]. Chinese Journal of Computers, 2002, 25(5):490 - 495.

[12] HAN J, KAMBR M. Data mining concepts and techniques [M]. San Fransisco: Morgan Kaufmann Publishers,2000.

[13] HAN J, LAKS V S, Raymond T. Constraint-Based Multidimensional data Mining [J]. Compuer, 1999, 32 (8):46 - 50.

[14]张 凯,胡运发,王 瑜. 基于互关联后继树的概念格构造算法[J]. 计算机研究与发展, 2004,41(9):1493 - 1499.
ZHANG Kai, HU Yunfa, WANG Yu. An IRST-based algorithm for construction of concept lattices[J]. Journal of Computer Research and Development, 2004, 41 (9):1493 - 1499.

[15]覃冬梅. 天体光谱信号的自动识别方法研究[D]. 北京:中国科学院自动化研究所,2003.
QIN Dongmei. The research on automatic recognition of astronomical spectra data [D]. Beijing: Institute of Automation Chinese Academy of Sciences,2003.

**作者简介**:

张继福,男,教授,2005年毕业于北京理工大学,获工学博士学位,CCF高级会员,主要研究方向:数据仓库与数据挖掘、人工智能及应用. 已发表学术论文50余篇,其中被SCI、EI收录20余篇. E-mail:jifuzh@sina.com.

张素兰,女,副教授,2003年毕业于太原科技大学,获工学硕士学位,主要研究方向:概念格与数据挖掘. 已发表学术论文10余篇.

胡立华,女,助教,2006年毕业于太原科技大学,获工学硕士学位,主要研究方向:概念格与数据挖掘. 已发表学术论文3篇.