《智能信息处理》课程作业

基于形式概念分析的语义角色挖掘算法

曹福笑

作业	分数[20]
得分	

2021 年 11 月 28 日

基于形式概念分析的语义角色挖掘算法

曹福笑

（大连海事大学 信息科学技术学院，辽宁 大连 116026）

摘 要 形式概念分析是一种对形式背景中的数据进行分析和规则提取的强有力工具,形式概念分析的核心数据结构是概念格。形式背景是由对象集合、属性集合、以及对象集合与属性集合之间的二元关系组成的三元组。语义角色的概念是美国语言学家 Charles J. Fillmore 于 1968 年提出的，他巧妙地沿用语法层面“格”的概念，但是把“格”分成两类，浅层格呢就是主格宾格属格与格这些语法上的格，可以看作是名词/名词组及代词在句子里承担的“语法角色”，而深层格就是语义的事了，也就是这些名词/名词组及代词在句子里还承担着“语义角色”。所以“语义角色”可以说是一个将语义和语法融合到一起的概念。基于形式概念分析生成用户权限概念格及用户属性概念格，将用户权限概念格翻转后映射为初始候选角色状态使得生成的角色有效地减轻了管理员授权的负担并且增强了角色的可解释性。

关键词 形式概念分析，概念格，角色挖掘

中图法分类号 TP311 **文献标识码** A

Semantic role mining algorithm based on formal concept analysis

Cao Fu Xiao

(School of Information and Communication Engineering, Dalian Maritime University, Dalian 116026，China)

Abstract .Formal concept analysis is a powerful tool for data analysis and rule extraction in the formal context. The core data structure of formal concept analysis is concept lattice. The formal background is a triplet consisting of object collections, attribute collections, and binary relations between object collections and attribute collections. The concept of semantic role was proposed by American linguist Charles J. Fillmore in 1968. He cleverly followed the concept of "case" at the grammatical level, but divided the "case" into two categories. The shallow case is the nominative, accusative, and genitive. The grammatical cases of dative case can be regarded as the "grammatical role" of nouns/noun groups and pronouns in the sentence, and the deep case is a matter of semantics, that is, these nouns/noun groups and pronouns are still in the sentence. To assume a "semantic role". So "semantic role" can be said to be a concept that combines semantics and grammar. Based on the formal concept analysis, the user authority concept lattice and the user attribute concept lattice are generated, and the user authority concept lattice is flipped and mapped to the initial candidate role state so that the generated role effectively reduces the burden of administrator authorization and enhances the Interpretability of the role.

Keywords Formal concept analysis, concept lattice, role mining

1 引言

形式概念分析理论是上世纪八十年代由德国数学家 Wille 提出的一种数据分析理论。该理论以形式背景为基础，从中提炼出形式概念，从而对哲学中的概念这一人类思维的基本单元进行了形式化表达，进一步又形成概念格这一核心数据结构，以进行知识的可视化表示。形式概念分析是一种强有力的数据分析和规则提取工具，在机器学习、数据挖掘、信息检索、软件工程等领域得到了广泛的应用。形式概念分析不仅能够实现同时对用户和权 限进行分组，而且形式概念分析中定义的形式概念及概念格与 RBAC 模型中定义的角色与角色层次具有十分完美的对应关系。

基于角色的访问控制是当前应用得最为广泛的访问控制模型之一，它将角色作为用户-权限分配的桥梁，极大程度地简化了用户的授权操作，实现了灵活方便安全的授权管理。角色是 RBAC 模型的基本特征也是其实现的基础。现有的角色挖掘算法挖掘出的角色集大部分都是扁平化的结构，在大型系统中不具有实际意义，本文主要关注能够挖掘角色层次的角色挖掘算法。

本文提出的算法是在文献［6］的基础上，基于形式概念分析生成用户权限概念格，将用户权限概念格翻转后映射为初 始候选角色状态，通过约简、精简操作最终得到带层次结构的角色集以及用户角色与角色权限的分配关系，之后基于形式 概念分析生成用户属性概念集，为角色定义最近似表达式，最大程度地为每个角色赋予现实生活中的语义信息，最终得到具有语义意义及层次结构的角色状态，增强了角色的可解释性，减轻了管理员的授权负担。

2 相关理论基础

2.1 形式背景

形式背景 K 是一个三元组（O,A,I），是形式概念分析的输入。其中 O 是对象集，A 是属性集，$I \subseteq 0 * A$ 是 O 与 A 之间的二元关系，对一个对象 o $\subseteq O$，属性 a $\in A$ 那么 oIa 就表示对象 o 具有属性 a：

$\forall o \in O$，$a \in A$有oIa $\rightarrow (o,a) \in I$

设 $X \subseteq O$和$Y \subseteq A$，定义如下两个映射：

1）$f(X) = \{a \in A | (\forall o \in X) oIA\}$

2）$g(Y) = \{o \in O | (\forall a \in Y) oIA\}$

2.2 概念格

概念格理论，也称形式概念分析理论，首先由德国的数学家于 1982 年提出。概念格作为形式概念分析中核心的数据结析的两个基本柱石。概念格理论的主要思想是在形式背景中寻找所有的概念并构造出格结构以此刻画出数据集中对象与属性之间的关系。我们定义概念背景 K 上所有形式概念 C 机器关系<构成的偏序集，记为 L（C，<）。设$(X_j, Y_j)(j \in J)$是概念格中的一个非空有限子集，其上确界记为$\vee j \in J(X_j, Y_j)$其下确界记为$\wedge j \in J\ (X_j, Y_j)$

$\vee (X_j, Y_j) = (g(\cap Y_j), \cap Y_j), \cup X_j \subseteq g(\cap Y_j)$

$\vee (X_j,\ Y_j) = (\cap X_j, f(\cap X_j)), \cup Y_j \subseteq f(\cap X_j)$

2.3 形式概念的关系：

对于同一形式背景中两个不同的形式概念（X1，Y1）和（X2，Y2），定义关系<, 若（X1,Y1）<(X2,Y2)

$\longleftrightarrow X1 \subset X2 \longleftrightarrow Y2 \subset Y1$ 则称概念（X1，Y1）是（X2，Y2）的亚概念，概念（X2，Y2）是（X1，Y1）的超概念。若

$\longrightarrow \exists$（$X3$，$Y3$）（$X1$，$Y1$）<（$X3$，$Y3$）<（$X2$，$Y2$）则称概念（X1，Y1）是（X2，Y2）的子概念，概念（X2，Y2）是（X1，Y1）的父概念，并记为（X1，Y1）<（X2，Y2）

2.4 角色挖掘

给定访问控制配置 $\rho = \langle U, P, UP \rangle$ 其中 U 是所有用户的集合，P 是所有权限的集合，$U, P \subseteq U \times P$ 是用户-权限关系。找出一个与 ρ 一致的 RBAC 状态<R,UA,PA,RH,UPDA>，其中 R 是角色集，$UA \subseteq U \times R$ 是用户-角色分配关系，$PA \subseteq P \times R$ 是权限-角色分配关系，$RH \subseteq R \times R$ 是角色层次及角色之间的偏序关系，$DUPA \subseteq U \times P$ 是直接的用户-权限分配关系。在 RBAC 状态中，若 U 中的每一个用户拥有的权限集与 UP 相同，则称 RBAC 状态与 ρ 一致。

3 语义角色挖掘算法

3.1 基于形式概念分析的角色挖掘

若 $\bigcup_{j \in J} X_j = g(\bigcap_{j \in J} Y_j)$，则概念 $\bigvee_{j \in J} (X_j, Y_j)$ 可由概念集$(X_j, Y_j)(j \in J)$表示。因此直接由概念格表示角色状态会出现冗余角色以及冗余的分配关系。本文方法将翻转的用户权限概念格映射为初始候选角色状态（角色继承关系的定义与本文概念关系的定义是相反的），其中候选角色与概

念对应，用户角色分配关系与概念的外延对应，权限角色分配关系与概念的内涵对应，角色层次关系与概念关系对应，在此基础上进行角色挖掘。约简概念的外延去除了冗余的用户角色分配关系，其内涵去除了冗余的权限角色分配关系。消除冗余后的分配关系后最多存在 4 类候选角色。

1）$\bigcup_{j\in J} X_j = X, \bigcup_{k\in K} Y_k = Y$,既没有用户也没有权限

2）$\bigcup_{j\in J} X_j \subset X, \bigcup_{k\in K} Y_k = Y$,有用户没有权限

3）$\bigcup_{j\in J} X_j = X, \bigcup_{k\in K} Y_k \subset Y$,没有用户有权限

4）$\bigcup_{j\in J} X_j \subset X, \bigcup_{k\in K} Y_k \subset Y$,既有用户也有权限

其中第四类候选角色是必须保留的，但前三类候选角色并不是必须消除，需要考虑消除候选角色之后是否使得 RBAC 状态更优。这里我们为前三类候选角色分别制定了以下三条规则。

规则 1：对于既没有用户也没有权限的候选角色 r，候选角色仅仅作为其他候选角色的一个连接点，移除该候选角色能够减少创建该候选角色以及与该候选角色相关的边带来的花费，但是为了保持继承关系的正确性，需要增加一些边。当满足以下条件时，移除候选角色 r。

规则 2：对于有用户但没有权限的候选角色 r，如果候选角色 r 被移除，则需要将 r 中的所有用户分配给 r 所有的直接下级候选角色 Jun，同时为了保持继承关系的正确性，需要增加一些边。

规则 3：对于有权限但没有用户的候选角色 r，如果候选角色 r 被移除，则需要将 r 中的所有权限分配给 r 所有的直接上级候选角色 Sen，同时为了保持继承关系的正确性，需要增加一些边。

由于概念格自身的特点，其映射的候选角色状态中满足规则 2 的候选角色往往集中于上层且向下精简，而满足规则 3 的候选角色往往集中于下层且向上精简，因此，采用自上而下的顺序消除满足规则 2 的候选角色，采用自下而上的顺序消除满足规则 3 的候选角色，使得最终得到的角色状态的带权结构复杂度相较而言会更小，即角色状态更优。

3.2 基于形式概念分析的语义赋予

语义可以简单地看作是数据所对应的现实世界中的事物所代表的概念的含义，以及这些含义之间的关系，是数据在某个领域上的解释和逻辑表示。对于角色来说，拥有语义意义的角色应该能够用用户或是客体属性来表示，本文仅以用户属性为例对角色进行语义赋予。本文我们为定理定义了如下几种表达式。

定义 1（属性表达式）：一个属性表达式能够表述成以下两种形式：

1）$e(A) = \Delta, A = \phi$：任何用户都满足属性表达式

2）$e(A) = a1 \wedge a2 \wedge a3 ... ak, A = \{a1, a2, a3, ..., ak\}$ 一个用户 u 满足配置 σ 下的属性表达式 e（A），当且仅当 $\forall i \in [1,k]$，有（u，ai）$\in UAT$ 。

定义 2（一致表达式）：给定一组用户属性配置 $\sigma =< U, A, UAT >$ 和与访问控制配置 $\rho =< U, P, UP >$ 一致的 PBAC 状态 γ，当且仅当 $U_{\gamma(r)} = U[e(A)]$ 时，称属性表达式 e（A）是角色 r 的一致表达式，U_γ（r）表示被分配角色 r 及其父角色的用户集。

定义 3（近似表达式）：给定属性配置 $\sigma =< U, A, UAT >$ 和访问控制配置 $\rho =< U, P, UP >$ 一致的 RBAC 状态 γ，当且仅当 U_γ（r）$\subseteq U_\sigma[e(A)]$ 时，称属性表达式 e（A）是角色 r 的近似表达式。特别的每个角色至少有一个近似表达式；如果角色 r 拥有一致表达式，那么这个一直表达式一定是其近似表达式。

定义 4（最近似表达式）：当且仅当属性表达式 e（A）是角色 r 的近似表达式，并且不存在一个 $A' \supset A$ 使得 e（A'）也是角色 r 的近似表达式时，称属性表达式 e（A）是角色 r 的最近似表达式。特别的对于给定的 RBAC 状态和用户属性配置，角色 r 有且仅有一个最近似表达式；如果角色 r 拥有一致表达式，那么这个一致表达式是其唯一的最近似表达式。

3.3 语义角色挖掘算法描述;

语义角色挖掘算法分为两部分：基于形式概念分析的角色挖掘算法和基于形式概念分析的语义赋予算法。其中算法 1 是基于形式概念分析的角色挖掘算法，算法 2 是基于形式概念分析的语义赋予算法。

算法 1 角色挖掘算法 Mine Role

Function MineRole （U，P，UP）

输入：用户 U,权限 P，用户权限分配关系 UP

输出：RBAC 状态<R,UA,PA,RH>

Begin

1. L=BuildLattice(U,P,UP);//构造概念格

2. <R,UA,PA,RH>=Reverse(R,UA,PA,RH);//翻转概念格

3. <R,UA,PA,RH>=L;//将翻转后的概念格映射为初始候选角色状态

4. <R,UA,PA,RH>=Reduce(R,UA,PA,RH);//使用约简概念计算公式约简初始候选角色状态。

5. <R,UA,PA,RH>=Prune(R,UA,PA,RH);//根据指定的 3 条消除规则执行精简操作，从而得到最终结果。

算法 2 语义赋予算法 Endow Meaning

Function Endow Meaning(r,U,A,UAT)

输入：角色 R，用户 U，属性 A，用户属性分配关系 UAT

输出：角色 R 的最近似属性表达式 e

Begin

1. L=BuildLattice(U,A,UAT);//构造概念格

2. $C_{UAT} = L.C$ //得到用户属性概念集

3. $C_{UAT} = Sort(C_{UAT})$ //将用户属性概念集排序

4. For r in R

5. $A_r = Search(r, C_{UAT})$ //搜索匹配的概念外延

6. $e_r = Conjunection(A_r)$ //合取式，即为近似表达式

7. End For

假设用户数为 m，权限数 n，候选角色数为 r，候选角色层次边数为 e，则构造概念格算法的时间复杂度为O（（$(m+n)^2+r^2$），约简初始候选角色 O（r）精简约简勾选角色状态算法的时间复杂度为 O(re), 最终算法 1 的时间复杂度为O（（$\mathrm{m}+\mathrm{n})^2+r(r+\mathrm{e}$））。虽然本文的算法时间复杂度略大于之前提出的 ORCA 算法与 GO 相当，但依然在可接受范围内且得到的角色状态更优。

4.实例分析：

4.1 实例描述

为验证算法的有效性，本文选用文献{12}中提到的电子病历系统作为背景实例，利用基于形式概念分析的语义角色挖掘算法进行角色挖掘和语义赋予，从而产生具有语义意义和层次结构的角色状态。在该实例中，用户岗位分为普通岗位和管理岗位两大类。普通岗位包括挂号员（1）、外科医生（2）、内科医生（3）、妇科医生（4）、护士（5）和药剂师（6）。管理岗位包括外科主任（7）、内科主任（8）、妇科主任（9）、医务科长（10）、总护士长（11）、药房主任（12）以及院长（13）。根据各个场景信息的读写以及各个职能的授权操作，枚举系统中所使用到的权限如下：读病人基本信息（a）、写病人基本信息（b）、读住院信息（c）、写住院信息（d）、读历史记录（e）、读诊断信息（f）、读药方（g）、读护士报告（h）、写内科历史记录（i）、写外科历史记录（j）、写妇科历史记录（k）、写内科诊断信息（l）、写外科诊断信息（m）、写妇科诊断信息（n）、写内科药方（o）、写外科药方（p）、写妇科药方（q）、写护士报告（r）、内科医生授权（s）、外科医生授权（t）、妇科医生授权（u）、药剂师授权（v）、护士授权（w）。科室与职能信息枚举系统中所使用到的属性如下：内科（A）、外科（B）、妇科（C）、配药（D）、挂号（E）、诊断（F）、护理（G）、主任（H）。最终拥有 13 类用户、23 类权限以及 8 类属性。

4.2 算法实现

步骤 1：根据提供的用户权限关系使用 Godin 算法构造用户权限概念格，将其翻转后映射为候选角色状态，如图一所示。

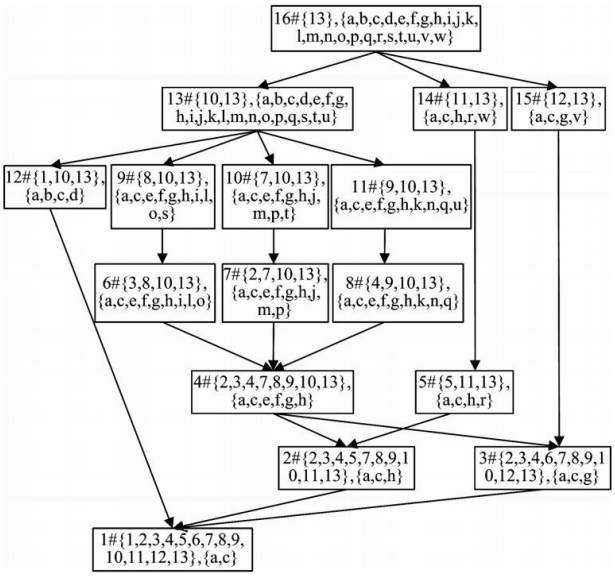

图 1 初始候选角色状态

步骤 2 查询每一个概念的子概念集和父概念集，根据约简概念计算公式计算其对应的约简概念，建立概念与约简概念的对应关系，从而得到简约的候候选角色状态，如图二所示。概念 4#{2，3，4，7，8，9，10，13}，{a,c,e,f,g,h}的子概念集为{6#，7#，8#}其外延的并集为{2，3，4，7，8，9，10，13}父概念集为{2#，3#}，其内

涵的并集为{a,c,g,h}, 计算得到概念 4#对应的约简概念为{}

步骤 3 设置$W=<1,1,1,1\infty>$，根据消除规则，自上而下地消除满足消除规则 2 的候选角色{16#}，自下而上地消除满足消除规则 3 的候选角色{2#}，得到用户权限精简格，即为最终的角色状态，如图 3 所示。

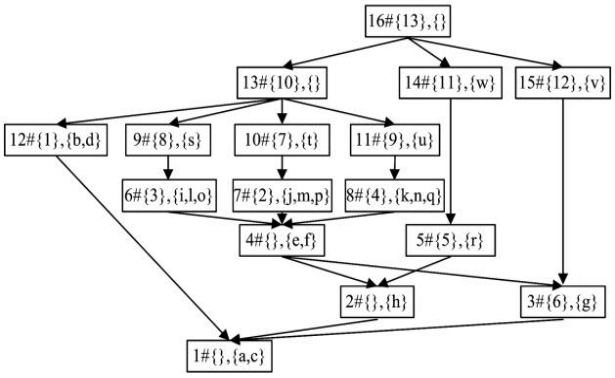

图 2 约简候选角色状态

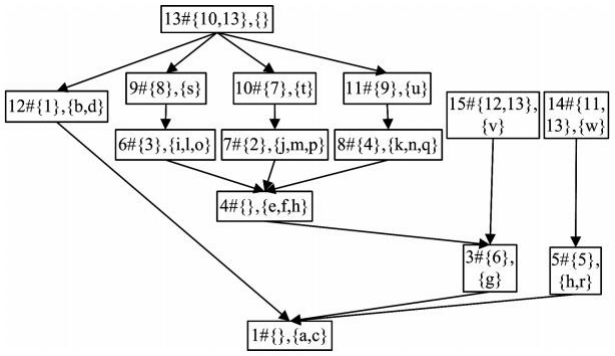

图 3 最终角色状态

步骤 4 根据上述提供的用户属性关系，使用 Ganter 提出的 NextClousure 算法生成用户属性概念集，并依据用户数目和权限数目对生成的概念集进行排序，从而得到一个有序的用户属性概念集，如图 4 所示。

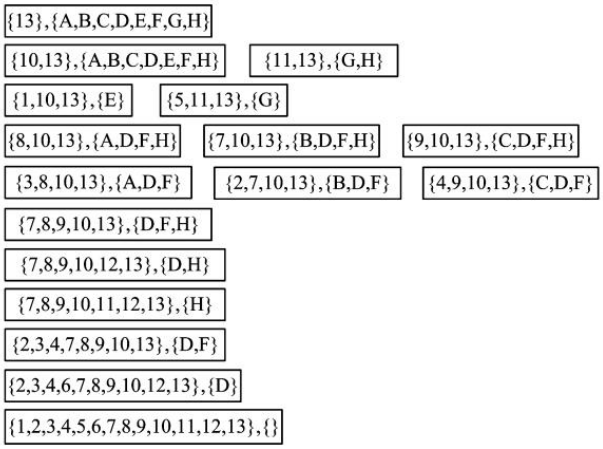

图 4 用户属性概念集

步骤 5 使用步骤 2 中建立的约简概念与概念的对应关系，将图 3 中的每个角色恢复成图 1 中的概念。

步骤 6 在图 5 所示的概念集中，对于每个角色对应的概念的外延，自上而下按序搜索其最近似表达式，为每个角色赋予语义意义。

4.3 算法比较

将本文算法挖掘到的角色结构（图 3）、文献{12}中挖掘到的角色结构与原始结构进行对比可以看出，仅仅挖掘最小角色集不足以表示所有的层次结构，虽然最小角色集看起来比较简洁，但是扩展起来却比较复杂。本文算法挖掘的角色结构扩展起来更加方便，需要添加的分配关系更少。同时，本文算法使用用户属性最近似表达式为角色赋予语义意义，相较于文献{6}中根据角色的权限和用户在系统中的功能和实际岗位为角色赋予语义意义更加精确。

结束语

基于形式概念分析的方法，本文提出了一种新的语义角色挖掘算法，该算法不仅能得到具有层次结构的角色集，而且能得到用户−角色以及角色−权限的分配关系，同时引入用户属性最近似表达式赋予角色现实生活中的概念作为语义，实现了授 权的半自动化，增强了角色的可解释性，能够有效地指导管理员进行授权管理，提高授权效率，减轻授权负担。虽然基于形式概念分析进行语义角色挖掘获得了十分优秀的角色结构，但是概念格构造算法本身的时间复杂度较高，从数据规模非常大的访问控制背景中挖掘角色结构十分费时。然而，随着计算机的发展和大数据时代的到来，概念格的并行构造算法也成为研究重点，基于形式概念分析的语义角色挖掘算法也同样具有很好的应用前景。

参 考 文 献

[1] SANDHURS，COYNEEJ，FEINSTEINHL，et al．Role-based-access control models［J］．Computer，1996，29（2）：38-47.

[2] SCHLEGELMILCH J，S T EFFENS U．Role mining with ORCA［C］//Proceedings of the tenth ACM symposium on Access control Models and Technologies ACM,2005:168-176

[3] SARM A H A K，HAZARIKA S M，SIN HA S K．Formal concept analysis：current trends and directions［J］．Artificial Intelligence Review,2015,44(1):47-86.

[4]GANT ER B，WILLE R．Formal concept analysis：mathematical foundations［M］．New York：Springer Science & Business Media，2012

[5]GANT ER B．Two Basic Algorithms in Concept Analysis［C］//International Conference on Formal Concept Analysis．Springer-Verlag,2010:312-340

[6]张磊，张宏莉，韩道军，等．基于概念格的 RBAC 模型中角色最小化问题的理论与算法［J］．电子学 报，2014,42(12):2371-2378

[7]孙钦东，张德运，高鹏．基于时间序列分析的分布式拒绝服务攻击检测［J］．计算机学报，2005，28（5）：767-773