《智能信息处理》课程作业

基于形式概念分析的重复社区发现

赵心田

作业	分数[20]
得分	

2020 年 11 月 29 日

基于形式概念分析的重复社区发现

赵心田
（大连海事大学 信息科学技术学院，辽宁 大连 116026）

摘要 概念格，也称为 Cralois 格，又叫做形式概念分析，由 Wille 于 1982 年首先提出，它提供了一种支持数据分析的有效工具。概念格的每个节点是一个形式概念，由两部分组成:外延，即概念所覆盖的实例;内涵，即概念的描述，该概念覆盖实例的共同特征。另外，概念格通过 Hasse 图生动和简洁地体现了这些概念之间的泛化和特化关系。因此，概念格被认为是进行数据分析的有力工具。本文首先介绍了复杂网络和社区发现的研究现状，并对社区的定义、社 区发现算法特别是重复社区发现算法的优缺点进行了详细的分析和比较 。其次，本文提出了基于形式概念分析的重复社区发现算法，设计概念的相似度模型对核心社区之间的连通性进行度量，在此基础上对概念进行层次聚类，对核心社区进行合并，生成大社区并产生重复节点，达到重复社区发现的目的。

关键词 形式概念分析，概念格，重复社区

中图法分类号 TP311 **文献标识码** A

Duplicate Community Discovery based on Formal Concept Analysis

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

Abstract .Concept lattice, also known as cralois lattice, also known as formal concept analysis, was first proposed by Wille in 1982. It provides an effective tool to support data analysis. Each node of the concept lattice is a formal concept, which consists of two parts: extension, that is, the instance covered by the concept; Connotation, that is, the description of the concept, which covers the common characteristics of the examples. In addition, the concept lattice vividly and concisely reflects the generalization and specialization relationship between these concepts through Hasse diagram. Therefore, concept lattice is considered to be a powerful tool for data analysis. Firstly, this paper introduces the research status of complex networks and community discovery, and analyzes and compares the definition of community, the advantages and disadvantages of community discovery algorithms, especially repeated community discovery algorithms. Secondly, this paper proposes a duplicate community discovery algorithm based on formal concept analysis, designs a concept similarity model to measure the connectivity between core communities, on this basis, hierarchical clustering of concepts, merging of core communities, generating large communities and duplicate nodes, so as to achieve the purpose of duplicate community discovery.

Keywords Formal Concept Analysis,Concept Lattice , Duplicate Community

1 引言

在我们的日常生活中，网络无处不在:经济、政治人际关系等组成了社会网络，汽车、火车等交通工具组成了交通网络，食物链、基因等组成了生物网络，引文网络、科学家合作网络等组成了科研网络，这些网络都具有复杂网络的结构特征。例如，英语词根网就是一个复杂网络的模型，如图 1 所示:

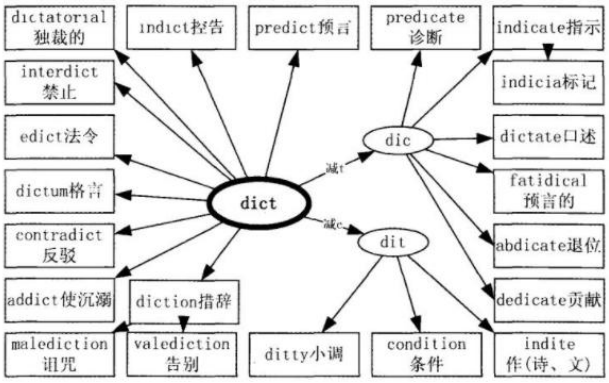

图 1

复杂网络中的实体千差万别，实体的关系也各有不同，通常用图论的方法对网络结构进行描述，将实体抽象成点，实体间的关系抽象成边，这些点和边就是复杂网络的研究对象。本文提出了基于形式概念分析的重复社区发现算法，设计概念的相似度模型对核心社区之间的连通性进行度量，在此基础上对概念进行层次聚类，对核心社区进行合并，生成大社区并产生重复节点，达到重复社区发现的目的。

2 基于形式概念分析的社区定义

2.1 概念格

在概念格的应用过程中，首先要解决的就是格的构建问题[29-32]。概念格的构建过程本质上就是聚类的过程，对于同一个形式背景构建出的概念格是唯一的，这也是概念格算法的优点之一。许多文献都提出了自己的构建算法，这些算法大致可以分为两类:批处理算法和渐进式构造算法。

(1)批处理算法

在生成概念格的过程中，批处理算法先生成形式背景对应的所有形式概念的集合，然后再确定各个概念之间的父子关系。根据任务完成的不同顺序，有两个不同的处理模型:一种模型为任务分割生成模型,先将全部的概念集合生成出来,接着再找到概念之间的父子关系，比如 Nextclourse 算法；另一种模型为任务交叉生成模型，每次只生成少数的概念，将这些概念连接到节点的集合中去，依此类推，直到所有的概念都生成完毕，比如 Bordat 算法。根据构建概念格的不同方式，又可以将其分为自顶向下算法、自底向上算法和枚举算法。自顶向下算法首先构造格的最顶层节点，再逐渐往下，比如 Bordat 算法。自底向上算法和自顶向下算法则正好相反，先构建底部节点，再向上扩展，比如 Chein 算法。枚举算法则根据某种顺序枚举出概念格中的所有节点，然后生成 Hasse 图，比如 Nourine 算法等。

批处理算法的一般步骤是:

1)初始化格 L=(O,D,R);

2)队列 F =(o,D,R);

3)对于队列 F 中的一个概念 C，产生出它的每个子概念 C;

4)如果某个子概念 C;不存在于格 L 中，则将其加入 L;

5)增加概念 C 和其子概念 C 之间的连接关系;

6)反复步骤 3)到步骤 5)，直到队列 F 为空;

7)输出概念格 L。

(2)概念格的应用

形式概念分析以概念格的形式将数据有机组织起来,概念格节点体现了概念内涵和外延的统一，这就使得其在数据挖掘中得到了广泛的应用。其中，基于概念格的关联规则提取是概念格在数据挖掘中应用最广、成果最丰富的领域。Godin 等人在概念格的基础上提取蕴涵规则,首先从形式背景构建概念格,再从概念格中产生连接规则,最后再去除冗余的蕴涵规则。Missaoui 等人对此进行了扩展，提出了提取近似规则的算法。基于概念规则的分类系统也得到了广泛应用和研究，比如 Oosthuizen.G 提出的 GRAND 算法、Liquiere.M 等人提出的 LEGAL 算法、Carpineto.C 等人提出的 GALOIS 算法等，基于概念规则的分类系统在国内的研究起步较晚，主要有谢志鹏等提出的 CLNN&CLNB 算法，这些分类算法使用了不同的策略来进行学习分类，有着不同的特征和适用范围。

随着形式概念分析和概念格研究的深入,概念格的应用领域正在不断扩大。ColeR 等将概念格方法应用于分析和可视化具有 1962 个属性和 4000 个处方摘要的医药数据库;同时，他还提出

利用概念格来建立一个 Email 收集器，并根据从加 Email 文档的头文件中提取接收时间和文档体中的关键词进行分类，并在 CEM 电子邮件管理系统中通过将 Emali 存储在概念格中，而不是常用的树状结构中，使得电子邮件的检索更为灵活。Kent 等提出了建造基于概念格的用于数字图书馆的系统 Nebula 及相应接口。

2.2 基于概念格的社区定义

为了有效的发现 Web 社区，Kumer 等人在文献[1][2]中提出拖网算法，该算法从二分有向图的角度对社区进行了一个明确的定义:对于一个有向的网络结构，将网络 G 中的节点集合划分为两个子集，子集 Fan 和 Center，集合 Fan 中有 i 个节点，集合 Center 中有 j 个节点，如果 Fan 中每个节点到 Center 中的每个节点都存在一条有向边,那么图 G 就是一个完全二分有向图。由此分析,二分有向图中的 Fan 和 Center 分别就是 hub 节点集合和 authority 节点集合。对于整个 Web 网络来说，可以通过寻找二分核心来发现主题的 hub 节点和 authority 节点。如图 2 为一个二分核心的例子。根据随机网络的定理，一个足够大而且稠密的随机二分网络中，将以很高的概率包含一个完全的二分子图。如果在互联网上存在某一个主题的社区，那么这种二分图的核心必将包含其中。在此社区定义的基础上，Kumer 等人通过重复的包含和排除剪枝得到所有的核,然后采用关联规则挖掘算法聚类为较大规模的核的集合来发现 Web 社区。

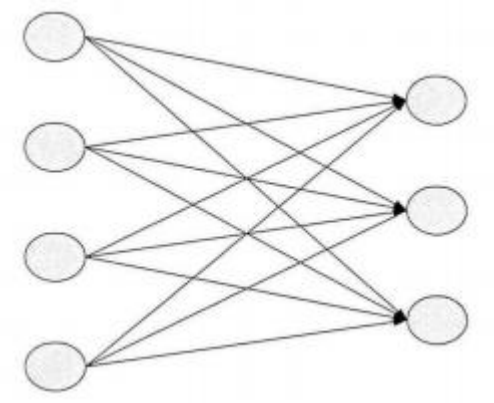

图 2

2005 年，在此基础上，Rome 和 Haralick[3]首次引入了形式概念分析的方法对二分有向图进行描述，并将其应用到数据检索和知识库构建中。Rome 等人先将 Web 图用形式背景(O,D,R)表示出来，O 和 D 都是 Web 页面集合，而 R 是这些页面的链接关系,然后根据形式背景构建概念格,得到的每一个概念就是一个社区核心。和 Trawling 算法类似，对于一个给定的概念 C=(A,B)，它的外延 A 代表了 Fan 集合，而内涵 B 代表了 Center 集合，且集合 A 中的每一个节点都有一条指向集合 B 中节点的边。

类似地，本文也引入形式概念分析的方法，将每一个社区定义为一个最大完全二分图，首先将网络的拓扑结构用形式背景(O,D,R)表示出来，O 和 D 都是节点的集合,而 R 则这些节点的链接关系，用对象 O 表示核心节点，对象 D 表示边缘节点，用概念来表示这种核心社区结构，核心节点共享边缘节点集合，边缘节点共享核心节点集合。在 FCA 的绝大多数应用中，形式背景中对象 O 和属性 D 是没有交集的，但是将其应用到社区发现中时略有不同，对象集和属性集都代表着页面集合，即 O=D。

下图以一个典型的实例来对基于形式概念分析的社区定义进行详细解释。现有一个 Web 有向图如图 3 所示:

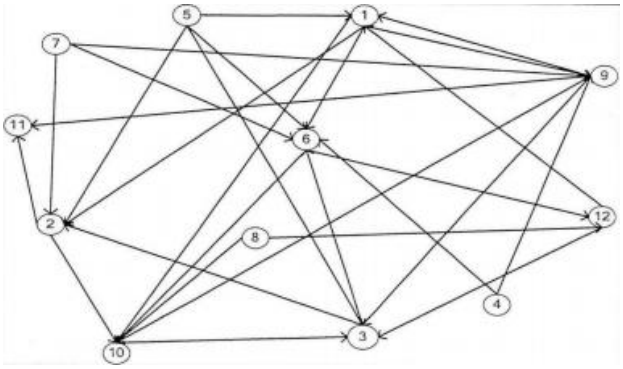

图 3

根据网络中的有向关系，构建形式背景如图 4 所示，构建概念格，得到的概念格如图 5 所示:

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1	0	0	0	1	0	0	1	0	0	0
2	0	0	0	0	0	0	0	0	0	1	1	0
3	0	1	0	0	0	1	0	0	1	0	0	0
4	0	0	0	0	0	1	0	0	1	0	0	0
5	1	1	1	0	0	1	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	1	0	1
7	0	1	0	0	0	1	0	0	1	0	0	0
8	0	0	0	0	0	0	0	0	0	1	0	1
9	1	0	1	0	0	0	0	0	0	1	1	1
10	1	0	1	0	0	0	0	0	0	0	0	0
11	1	0	1	0	0	0	0	0	0	0	0	0
12	1	0	1	0	0	0	0	0	0	0	0	0

图 4

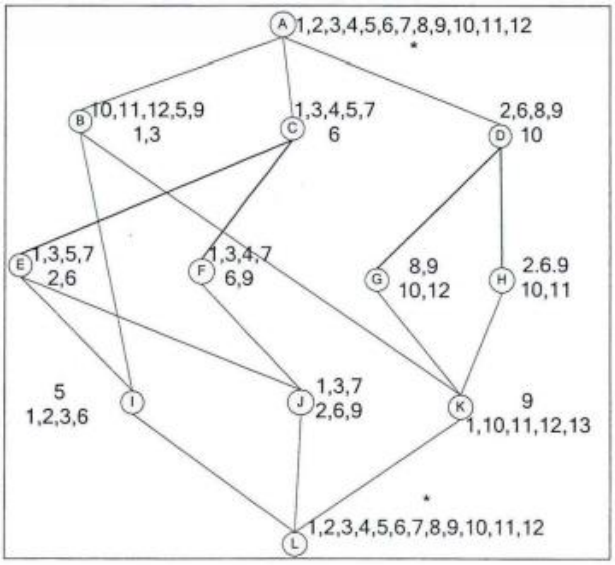

图 5

如图 5 所列，去除包含空集的概念，得到的概念集共有 10 个，每个概念都是一个完全二分子图,也是一个核心社区,它包含两个不相交的子集 V(fans)和 V(center)。以概念 J(137,269)和 C(13457,6)为例对社区定义进行说明，J(137,269)所表示的社区结构是一个(3,3)二分核心，C(13457,6)表达的核心边缘结构中，6 处于社区的核心，而其它节点 1,3,4,5,7 处于社区的边缘，如图 6 所示。

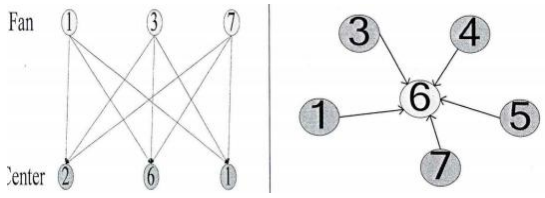

图 6

另外，Hasse 图形象直观的表示出概念之间的关系，如图 3-4 所示，由 Hasse 树结构可知，概念 E(1357,26)和概念 F(1347,69)是兄弟概念，他们有共同的父节点 C(13457,6)和子节点 J(137,269)，他们具有较好的紧密性。基于 Hasse 树结构，研究核心社区之间的关系，为下一步核心社区之间关系的挖掘以及重叠社区发现提供基础。

在 Trawling 算法中，首先需要设置二分核心社区的大小，即 i 和 j 的值是固定的，再通过循环迭代计算出所有的二分核心社区。但是，本文运用 FCA 的方法来发现网络中的核心社区就具有很好的优越性:无需指定二分核心社区规模的大小，一次执行就能全部提取出 Web 图中不同粒度的核心社区，并可视化社区之间的层次交互关系,确保了信息的完整性。

3 基于 FCA 的 CCD 算法研究

为了解决核心社区发现的问题，并解决概念格构建中算法复杂度过高的问题，根据复杂网络中形式背景的特点,提出了一种基于 FCA 的核心社区发现算法(An CenterCommunity Detection Algorithm Based On Formal Concept Analysis)。本节主要介绍如何根据网络结构获取形式背景，将形式背景转换成 0-1 矩阵，并进行矩阵运算以获取概念，生成 Hasse 树，从而达到核心社区发现的目的。取形式背景，将形式背景转换成 0-1 矩阵，并进行矩阵运算以获取概念，生成 Hasse 树，从而达到核心社区发现的目的.

3.1 形式背景构建

应用 FCA 发现核心社区，首先要解决的问题是如何根据网络的拓扑结构构建形式背景。由于概念格运用到社区发现中的特殊性，对象 O 和属性 D 都是网络中节点的集合,即 O=D;当网络中存在 n 个节点时,即表示存在 n 个 O 和 n 个 D,即 |O|=|D|=n 。

假设存在形式背景 T=(O,D,R)，其中，包含 m 个对象和 n 个属性，O={o:1...m}，D={d:1...n}，表示为|O|=m，|D|=n。从形式背景中构建 m*n 矩阵 A。

假设存在形式背景 T=(O,D,R)， o_i是对象集 O 中的第 i 个对象，d_j表示属性集中的第 j 个属性，a_{ij}表示矩阵 A 中的第 i 行、第 j 列元素。当 a_{ij}=1 时，$(o_i, d_j)\in$R,表示为o_iRd_j。

在社区发现中，利用网络拓扑结构构建形式背景，对象 O={o:1...n}，属性 D={d:1...n}，对于形式背景 T=(O,D,R)，关系 R 表示 O 和 D 之间的连接关系，当存在这种 O 指向 D 的连接关系则用 1 表示它的值;反之，当不存在这种关系则用 O 表示，由此得到一个单值背景，形式背景就是一个 0-1 矩阵。图 4 就是对应的形式背景。

对于有向图，节点之间有很明确的有向关系，连接的起点表示为对象，终点表示为属性;对于无向图，节点之间没有很明确的有向关系，因此关系形式背景就是一个对称的 0-1 矩阵。由于在本文的后续算法中，利用了概念格的对偶性，由于对称矩阵中的元素关于主对角线对称,因此,本文只需要存储矩阵中上三角或下三角中的元素，取其对角矩阵作为形式背景，降低矩阵

运算的复杂度。

3.2 核心社区的发现

(1)基础概念的获取

在形式背景确立之后，下一步就是要根据形式背景生成概念格，得到核心社区。本文提出了一种基于矩阵运算的概念格构建算法,它是一种自底向上的任务分割生成算法，通过矩阵的运算得到概念格。此算法中形式概念被分为基础概念和增加概念两类，下面通过定义、定理、推论和伪代码，并结合实例对概念格构建过程即核心社区发现过程进行说明。

假设矩阵 A(m*n)对应一个形式背景，其中 A’为 A 的转置矩阵，令 C= A’⊗A，则 C = $\{d_k \in D | a_{ik}=1, a_{kj} = 1, k = 1...n\}$(i=1...m, j =1..m)，其中$C_{ij}$表示矩阵 C 的第 i 行、第 j 列元素，$a_{kj}$,表示矩阵 A’的第 i 行、第 j 列元素。$C_{ij}$表示第 i 个对象和第 j 个对象具有的共同属性集。

以上可得矩阵 A 和转置 A’分别为（图 7）：

	A	B	C	D	E	F
1	1	1	0	1	1	1
2	1	0	1	1	1	1
3	0	1	1	1	1	1

图 7

$$A=\begin{pmatrix}1&1&0&1&1&1\\1&0&1&1&1&1\\0&1&1&1&1&1\end{pmatrix} A'=\begin{pmatrix}1&1&0\\1&0&1\\0&1&1\\1&1&1\\1&1&1\\1&1&1\end{pmatrix}$$

（2）算法伪代码（图 8）

```
BEGIN
    Conceptset=∅;
    FOR i← 1 to |o|DO    //|o|是对象的总个数
        FOR j← 1 to |o|DO    //双重 for 循环获取基本概念
            IF(c_ij*,c_ij) ∉ Conceptset THEN//避免概念的重复而进行判断
                Conceptset ← Conceptset∪(c_ij*,c_ij);
    RETURN Conceptset;
END
```

3.3 小结

在本小节中，形式概念分析的方法被应用到社区发现领域中。在详细分析介绍了形式概念分析的基本知识及应用的基础上，结合社区的核心边缘结构，提出了基于形式概念分析的社区定义，用概念来表示核心社区，并给出了全文研究的步骤;然后重点提出基于形式概念分析的核心社区发现算法,提出了一种简单快捷的建格方法来发现核心社区，首先提出了拓扑结构的形式背景转换方法，得到形式背景 0-1 矩阵，然后给出了基于矩阵运算的核心社区发现算法，该算法将形式概念发展成基础概念提出了一系列的假设用于解释怎么从形式背景中获得这两种概念，并给出了概念生成的伪代码，发现所有的核心社区。

4 本文总结

社区发现是复杂网络研究中的一个热点，然而，对于重叠社区发现的研究相对较少且不成熟，且对社区发现结果的实际意义分析较少。如何从现实的网络中挖掘出具有实际意义的社区结构成为一个难题，本文的研究正是在此基础上展开的。

社区发现是复杂网络研究中的一个热点，然而，对于重叠社区发现的研究相对较少且不成熟，且对社区发现结果的实际意义分析较少.如何从现实的网络中挖掘出具有实际意义的社区结构成为一个难题，本文的研究正是在此基础上展开的。

本文介绍了形式概念分析方法的理论知识及其应用，并讨论了以关系作为属性构建形式背景的方法，在此基础上，提出了基于矩阵运算的算法来生成概念，降低核心社区发现的算法复杂度；又提出了生成 Hasse 图的方法来得到核心社区和核心社区之间的关系.在此，算法得到的社区的空间结构不仅是节点的集合，还是社区内集合间的层次关系。

5 工作展望

本文提出的基于形式概念分析的重叠社区发现方法可以较准确的识别网络中的重叠社区结构，然而在对该方法实验及其应用的研究过程中发现，还存在一些问题有待进一步的研究。

本文提出的基于形式概念分析的重叠社区发现方法可以较准确的识别网络中的重叠社区结构，然而在对该方法实验及其应用的研究过程中发现，还存在一些问题有待进一步的研究.

(1) 在本文所提算法的适用能力有限，在运用形式概念分析发现核心社区时，本文只是对单值背景的网络结构进行探讨,当运用到加权网络时,需要运用到多值背景，针对多值背景的核心社区发现算法需要进--步的研究和探索。

参考文献

[1] Kumar R, Raghavan P, Rajagopalan s. Trawling the web for emergingcyber-communities[C]. The 8h Int'l www Conf, Toronto, Canada, 1999:403-415.

[2] Kumar R, Raghavan P, Rajagopalan S. Extracting large-scale knowledge base from theweb[C]. The 25th Int'l Conf. on Very Large Data Bases(VLDB99)，Edinburgh,Scotland, 1999: 639-650.

[3]形式概念分析中的概念约简与概念特征[J]. 魏玲,曹丽,祁建军,张文修. 中国科学:信息科学. 2020(12).

[4]朱新华,马润聪,孙柳,等.基于知网与词林的词语语义相似度计算 [J]. 中 文 信 息 学 报 , 2016, 30(4): 29-36.

[5]Diversified top- k maximal clique detection in Social Internet of Things[J] . Fei Hao,Zheng Pei,Laurence T. Yang. Future Generation Computer Systems . 2020 (C).

[6]模糊三支形式概念分析与概念认知学习[J]. 徐伟华,杨蕾,张晓燕. 西北大学学报(自然科学版). 2020(04)

[7]Jain Sanjay, Stephan Frank, Zeugmann Thomas On the amount of nonconstructivity in learning formal languages from text[J] Information and Computation, 2020(prepublish).

[8] 池哲洁,张全.基于概念基元的词语相似度计算研究[J].电子与信息学报,2017,39(1):150-158.