《智能信息处理》课程作业

基于实时通信软件的形式概念分析

宋祥鑫

作业	分数[20]
得分	

2020 年 11 月 11 日

基于实时通信软件的形式概念分析

宋祥鑫
（大连海事大学 信息科学与技术学院 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。FCA 自引入以来获得广泛的研究，被成功应用于语言学、软件工程、心理学、人工智能、信息检索等领域。本文简要介绍了形式概念分析的基本概念、过程和基本方法，以若若干实时通信软件信息为形式背景进行形式概念分析。

关键词 形式概念分析，概念格，社交软件

Real-time Communication Software based on the formal concept analysis

Song Xiangxin
(Department of Information Science Technology, Dalian Maritime University, Dalian 116026)

Abstract Formal concept analysis is a branch of applied mathematics, which greatly limited collection has some classify or contain attributes of the elements, attributes and objects found made of concepts relationships, and the mathematical expression of concept and conceptual levels. FCA since its introduction access to extensive research, applied linguistics, software engineering, psychology, artificial intelligence, information retrieval, and other fields. This paper briefly introduces the basic formal concept analysis concepts, procedures and methods, information in the form of a number of mobile social software background of formal concept analysis.

Keywords Formal concept analysis, Concept lattice, Social contact software

1 引 言

形式概念分析是由Wille R提出的一种有效的知识获取工具。它是建立在数学基础之上，对组成软件本体的概念、属性以及关系等用形式化的语境表达出来，然后根据语境，构造出概念格，从而清楚地表达出本体的结构。目前，它已被广泛地研究，并成功应用到机器学习、软件工程和信息获取等领域。

目前实时通信软件的种类层出不穷、各有千秋，如何找到合适的实时通信软件就是至关重要的。如果可以通过不同软件的产品定位、社交方式方法、社交背景等相关知识库的学习发现其中隐含的特有优势将有助于使用者对工具进行快速准确的选择，达到事半功倍的成果。而形式概念分析上的规则提取可以有效的提取知识库中的隐藏规则，因此利用形式概念分析对若干实时通信软件进行研究是一个很有意义的课题。

本文利用若干实时通信软件中的不同特征及其使用过程构造对象与属性之间的二元关系，根据此二元关系构造若干实时通信软件上的形式背景，并对常见方法进行分

析，得出方法的产品定位、社交方式方法与社交背景等之间的决策关系。

2 形式背景

形式概念分析首先要建立形式背景。形式背景被定义为一个三元组，公式为 K=(G,M,I)，其中 G 为所有对象集合，M 为所有属性的集合，$I \subseteq G \times M$ 为 G 和 M 中元素之间的二元关系集合[6]。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 $G=\{g_1,g_2,g_3,g_4\}$，属性集合 M={a,b,c,d}，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。

表 1 形式背景的示例

	a	b	c	d
g_1	1	0	1	1
g_2	1	1	0	0
g_3	1	1	1	0
g_4	0	0	1	0

针对常见的若干实时通信软件，提取相关形式背景，为了便于说明，若干实时通信软件领域形式背景如表 2，其对应关系如下：1 代表微信，2 代表新浪微博，3 代表 Line，4 代表陌陌，5 代表人人网，6 代表知乎，7 代表探探，8 代表 QQ；a 文字为主，b 基于陌生人，c 语音为主，d 可以匿名，e 基于熟人，f 基于地理位置，g 基于热点新闻。

表 2 若干实时通信软件-形式背景

	a	b	c	d	e	f	g
1	1	0	1	0	1	1	0
2	1	0	0	0	1	0	0
3	0	1	1	0	0	0	0
4	0	0	1	0	0	1	0
5	0	1	1	0	1	0	0
6	0	1	0	1	1	0	0
7	0	0	1	0	0	1	0
8	1	0	1	0	0	0	0

3 概念格

建格的过程实际上是概念类聚的过程[1]。因此，在概念格中，建格算法具有很重要的地位对于同一批数据，所生成的格是唯一的，即不受数据或属性排列次序的影响，这也是概念格的优点之一。概念格的建格算法可以分为两类：批处理算法和增量算法。概念格可以添加背景知识，这些知识以 if…then 的规则形式出现。概念格甚至可以只用背景知识建造。

批处理算法根据其构造格的不同方式，可分为 3 类，即从顶向下算法、自底而上算法、枚举算法。从顶向下算法首先构造格的最上层节点，再逐渐往下。自底而上算法则相反，首先构造底部的节点，再向上扩展。枚举算法则是按照一定顺序枚举格的所有节点，然后再生成 Hasse 图，即各节点之间的关系。增量算法和批处理算法不同，增量算法的思想都是大同小异的——基本思想都是将当前要插入的对象和格中所有的概念交，根据交的结果采取不同的行动[2]。主要区别在于连接边的方法。

以下，本文根据已给领域的简单的形式背景产生对应概念格。首先，根据所给形式背景约减生成单值形式背景，再确定单值形式背景中的父子关系，根据父子继承关系绘制 Hasse 图，最后补充各形式概念的上确界和下确界，形成概念格。针对若干实时通信软件领域的形式背景，形式概念分析的具体步骤如下。

3.1 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 2 可看出，4、7 与是一组有相同属性的行，故将其合并。最后得到约减后的形式背景如表 3。

表 3 若干实时通信软件-约减形式背景

	a	b	c	d	e	f	g
1	1	0	0	0	1	0	0
2	1	0	0	0	1	0	1
3	0	1	1	0	0	0	0

4,7	0	0	1	0	0	1	0
5	0	1	1	0	1	0	0
6	0	1	0	1	1	0	0
8	1	0	1	0	0	0	0

3.2 生成单值形式背景

单值的形式背景即根据前一步约减后的形式背景，把值为“1”的位置改为“×”，去掉其他位置的“0”以表示该形式对象有此属性。最后得出结果见表 4。

表 4 若干实时通信软件-单值形式背景

	a	b	c	d	e	f	g
1	×				×		
2	×				×		×
3		×	×				
4,7			×			×	
5		×	×		×		
6		×		×	×		
8	×		×				

3.3 确定父子关系

父子关系也称基于属性个数的排序。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 2 可由对 1 的全部属性继承的基础上添加自身属性 g 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 5 所示为基于属性个数的排序。

表 5 若干实时通信软件-基于属性个数的排序

	a	b	c	d	e	f	g
1	×				×		
3		×	×				
4,7			×			×	
8	×		×				
2	×				×		×
5		×	×		×		
6		×		×	×		

3.4 绘制 Hasse 图

根据偏序关系可生成概念格的 Hasse 图。如果有 $C_1 > C_2$，并且不存在另一个元素 C_3 使得 $C_1 > C_3 > C_2$，则在 C_1 和 C_2 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 5 所绘 Hasse 图如图 1 所示。

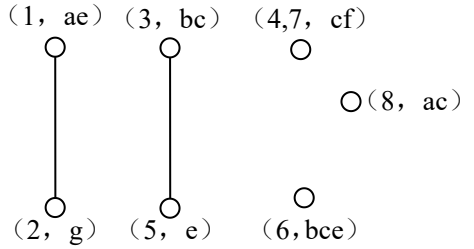

图 1　Hasse 图

3.5 生成概念格

针对表 5 的简单形式背景，采用手工方式生成概念格。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。图 2 是产生的概念格。从({6},{b,d,e})、({5},{b,c,e})和({3,5},{b,c})中抽取属性 b 产生新的形式概念({3,5,6},{b})作为上确界；抽取属性 a 作为({2},{a,e,g})、({8},{a,c})和({1,2},{a,e})的共有属性，产生形式概念({1,2,8},{a})作为二者的上确界；拥有属性 c,f 的形式概念({4,7},{c,f})即为自身的上确界。同理补充上确界({1,2,3,4,5,6,7,8,9,10},φ)和下确界(ϕ,{a,b,c,d,e,f,g,h,i,j})。

4 规则提取及相关系统

4.1 相关系统

从概念格上可以提取规则，目前就此方面已经有了许多研究，如 Godin 的系统：一种增量式概念格构造方法，提出在概念格上提取蕴含规则的算法。LEGAL 的系统：通过通过引进两个参数α和β，改进 Bordat 的算法，将其应用于分类任务。Rulelearne 的系

统：该系统由 Stanford 大学的 Sahami 提出，该算法采用一种“标记法”从格中提取规则。

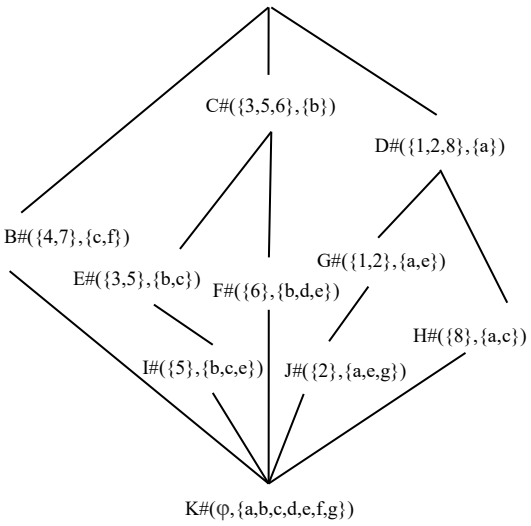

图 2　概念格

4.2 提取规则及相关解释

概念格上提取的规则具有相当好的分类效果，规则提取的定理如下。

定理 1　如果格中结点 H= (X,X′)只有一个双亲结点 M= (Y,Y′)，则 H 所产生的规则前件只能为单个描述符，且存在 p∈{X′-Y′}，都有一条无冗余规则 p→ X′-p。

定理 2　如果格中结点 H=(X,X′)具有 d 个双亲结点 $M_1(Y_1,Y'_1)$，$M_2(Y_2,Y'_2)$，…，$M_d(Y_d,Y'_d)$，则对于任意一个描述符 p∈{X′-($Y'_1 \cup Y'_2 \cup \ldots \cup Y'_d$}，都存在一条规则 p→ X′-p。

定理 3　如果格中结点 H=(X,X′)具有两个双亲结点 $M_1=(Y_1,Y'_1)$和 $M_2=(Y_2,Y'_2)$，则 $p_1 \in \{Y'_1\text{-}Y'1 \cap Y'_2\}$和 $p_2 \in \{Y'_2\text{-}Y'_1 \cap Y'_2\}$，都存在一条规则 $p_1p_2 \rightarrow X'\text{-}p_1p_2$，并且前件为两个描述符的规则总数是 $\| Y'_1\text{-}Y'_1 \cap Y'_2 \| * \| Y'_2\text{-}Y'_1 \cap Y'_2 \|$。

注意到只有当 ‖ Y ‖ > k 时，才可能有前件至多为 k 个描述符的规则，并且规则前件的描述个数至多为其双亲节点的数目。除了前件为单个描述符的规则之外，其他规则的形式与数目仅仅依赖于其双亲节点。表 6 给出根据上述定理所推到出的最简的全局规则示例。

表 6 最简全局规则示例

规则	所属节点范围	相关解释
b→de	C#,F#	基于陌生人可匿名社交，熟人社交
b→c	C#,E#	基于陌生人以语音为主
b→ce	C#,E#,I#	基于陌生人以语音为主，熟人社交
a→e	D#,G#	文字为主且基于熟人社交
a→eg	D#,G#,J#	文字为主，基于熟人社交，围绕热点新闻
a→c	D#,H#	文字与语音并重

5 结 论

在寻找服务类产品设计方法的过程中建造与应用概念层次结构进行方法选取具有很多优势，而概念格的 Hasse 图正好体现了一种概念层次结构，反映了概念之间的共有和私有属性。本文基于若干若干实时通信软件给出该领域的形式背景，根据若干实时通信软件与其属性之间的关系，构造决策背景和概念格，进而利用概念格提取了关联规则，并对这些决策规则进行了解释。很大程度的提高了用户在进行产品选择时的效率和准度，达到事半功倍的成果。关于概念格的的应用还有许多问题有待研究，例如发展高效的构造概念格及剪枝算法；如何从格上产生有意义的规则；如何找到更好的概念之间的计算方法，以更方便的提取规则等。

参考文献：

[1] 左雄辉,糜仲春. 概念格在电子商务中的应用[J].科技进步与对策, 2018, 20(5): 141-142.

[2] 王娜. 基于概念格的知识获取 [J]. 科技创业, 2019, 6(4): 118-120.

[3] 张文修,魏玲,祁建军. 概念格的属性约简理论与方法[J]. 中国科学(E 辑), 2016, 25(5): 490-495.

[4] 宋炜,张铭. 语义网简明教程[M]. 北京:高等教育出版社, 2017.

[5] 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, 2018, 35(6): 628-639.

[6] 杨帆,翟岩慧,曲开社,李德玉. 基于形式概念分析的词义理解研究[J]. 2017, 38(10): 189-191.

[7] 曲开社,翟岩慧.偏序集、包含度与形式概念分析[J].计算机学报,2016,29(2):32- 33.