《智能信息处理》课程作业

基于形式概念分析的 Web 数据库抽取分析

尹鹏

作业	分数[20]
得分	

2020 年 11 月 22 日

形式概念分析

尹鹏

（大连海事大学 信息科学技术学院, 大连 116026）

摘 要 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。

关键词 形式概念分析；社交软件；概念格；

Review of formal concept analysis

Yin Peng

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

Abstract Formal concept analysis is an effective knowledge acquisition tool proposed by Wille R. It is based on mathematics and expresses the concepts, attributes, and relationships that make up the software ontology in a formal context, and then constructs a concept lattice based on the context, thus clearly expressing the structure of the ontology. At present, it has been widely studied and successfully applied to the fields of machine learning, software engineering, and information acquisition. At present, there are endless types of real-time communication software, and each has its own merits. How to find the right real-time communication software is of utmost importance. If you can discover the unique advantages implicit in the knowledge base of different software product positioning, social methods, social background and other related knowledge bases, it will help users to quickly and accurately select tools and achieve a multiplier result with half the effort. The rule extraction in formal concept analysis can effectively extract the hidden rules in the knowledge base, so it is a very meaningful topic to use formal concept analysis to study several real-time communication software.

Key words Formal concept analysis; Social software; Concept lattice

1 形式概念

1.1 概念的含义及表示

在英文中，概念（concept）一词，其词根词缀为共同（con-）以及拿（-cept-），译为概念、观念、思想，常常指被广泛接受的思想或概念。概念（concept）一词源于拉丁语 conceptus,其过去分词为 concipere。该词的词根词缀意思为共同（com-）以及截取（capere）。而在中文中，概念又由两个字概和念构成。概，是古代一种量具用词，表示用作对古代量具‘斛’的满量状态做出校准。原定义为：量米粟时，使用木板在斗斛上刮平，使其处于一定范围以内，不至于过满，表示对事物做出限定，使其不超出范围。现代字义为：处于一定范围内，如：大概、概念、概括。念，则是令心，心之力，心通

思，自然地思维，常思。

而在现代的定义中，概念具有两个基本特征，也就是概念的内涵和外延。因为概念的外延是从其内涵中演绎出来的，所以外延中的所有元素都具有共同的本质属性。概念的内涵就是指这个概念的含义，即该概念所反映的事物对象所特有的属性。概念的内涵为：任意标识或范畴位置，任意标识在范畴结构中的位置认知。或表达成：概念{S/合{正/反}}：合{正/反}是范畴的结构式，其中的正、反、合是三个固定的范畴结构位置，S 是任意标识，S 可以放到不同的位置上以获得不同的范畴意义而被定义认知为一个有某种抽象意义的概念。概念的外延就是指这个概念所反映的事物对象的范围，即具有概念所反映的属性的事物或对象。概念的外延为：内涵性质和外延种类。概念的外延包括：内涵性质{性质∨范畴}，外延种类{个体∨种类}。概念分为“性质、范畴、个体、种类”共四大基本类型。集合论中某集合的任何元素都可以而且只能属于这四类概念及其他的集合（不包括这个某集合本身）。

2. 形式概念分析

本节主要介绍文中所述的基于形式概念分析数据抽取方法所涉及到的形式化概念分析的内容。这里给出部分概念，对形式化概念分析更为详尽的形式化描述即相关内容可参考文献[6].

定义 1 形式背景(FormalContext)：为一个三元组 K = (O ，A ，I)，其中 O 是对象(实体)集合，A 是描述符(属性)集合，I 是 O 与 A 之间的一个二元关系，即 I ⊆ O × A.

定义 2 Galois 联系，形式背景 K 中，在对象 O 的幂集 X ⊆ O 和属性 A 的幂集 Y ⊆ A 之间可以定义两个映射 f 和 g，定义如下:.

- $f: P(O) \to P(A), f(x) = X' = \{y \in A | xIy, \forall x \in X\}$
- $g: P(A) \to P(O), g(Y) = Y' = \{x \in O | xIy, \forall y \in Y\}$

因为 Galois 联系的两个映射是对偶的，为了行文的方便，以下简写为‘运算.

定义 3 形式概念二元组 c = (X，Y)，其中 X ⊆ O，Y ⊆ A，满足X′=Y 和 X=Y′，则 c 被称为是形式背景 K 的一个形式概念，其中 X 和 Y 分别被称为概念 c 的外延和内涵.形式背景 K 所产生的所有形式概念的集合表示为C_k

假设$(X_1, Y_1)$and(X_2, Y_2)是形式背景 K 的两个形式概念.如果$X_1 \subseteq X_2$或者$Y_1 \supseteq Y_2$，则(X_1, Y_1)称之为概念(X_2, Y_2)的子概念(sub-concept); (X_2, Y_2)称之为概念(X_1, Y_1)的父概念(super-concept).子概念与父概念之间构成了偏序关系，表示为$(X_1, Y_1) \leq (X_2, Y_2)$.

定义 4 父子结点，对于概念格L_k中两个不同结点$c_1 = (X_1, Y_1)$. $c_2 = (X_2, Y_2)$，如果$c_1 \leq c_2$，并且$\nexists c_3 \in L_k$满足$c_1 \leq c_3 \leq c_2$, 则c_1称为c_2的子结点或者直接后继(Immediate successors)，而c_1是c_2的父结点或者直接前驱(ImmediatePredecessors).

定义 5 概念的上/下覆盖(Upper/Lowercover)，由概念 c 的所有直接前驱/直接后驱组成的集合称为概念 c 的上/下覆盖(Upper/LowerCover)，表示为$Cov^{ll}(c)$和$Cov^{l}(c)$.

定义 6 形式概念的势，$\|(X,Y)\| = \|X\|$，其中（X,Y）$\in L_k$.

表 2.1 形式背景 K_1, 其中 O = {1, 2, ..., 7}, A ={a, b, ..., e}.

	a	b	c	d	e
1	X	X			
2		X	X	X	
3				X	X
4				X	X
5		X	X	X	
6		X	X	X	
7	X	X			

例 1 表 2.1 中列出了形式背景K_1，其中 O = 1，2，…，7，A=a，b，…，e;图 2.1 中给出了由形式背景K_1所产生的概念格L_{K_1}.根据定义 2，{12567}′={b} ，{256}′={bcd} ，{b}′={12567}; (12567,b)and(256，bcd)满足伽罗瓦联系，因此它们是形式背景K_1中的形式概念.又因为{256}⊆{12567}，概念(256，bcd)是概念(12567，b)的子概念，表示为(256，bcd) ≥ (12567，b).形式背景K_1中所有的概念集合C_{K_1}，构成概念格L_{K_1}，(图 2.1).从图 2.1 中可以看出，概念(17，ab)和概念(256，bcd)都是概念(12567，b)的子结点.因此，Cov^{l}((12567，b))={(17，ab)，(256，bcd); 同理，Cov^{ll}((256，bcd))={(23456，d)，(12567，b)}.

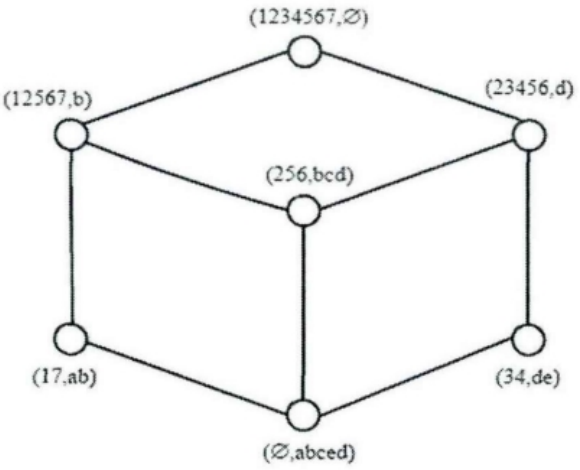

图 2.1　形式背景 K_1 产生的概念格 L_{K_1}

3 概念格

格（lattice）的意义是任两个元素的上确界和下确界都存在的偏序集。完备格为任一子集的上确界和下确界存在的偏序集，其特点是只有一个最高点，且只有一个最低点，且图中任何两点连通。概念格是元素为概念的完备格。概念格，也称为Cralois格，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，每一个形式概念都是由外延和内涵两部分组成。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。从形式背景中生成概念格的过程实质上是一种概念聚类过程。目前，概念格已被广泛应用于机器学习、模式识别、专家系统、数据挖掘、信息检索等领域[2]。

4 Web 数据库

20 世纪 90 年代中期，深网(DeePWeb)的概念由 MatthewKoll 首次提出.相对于用户通过搜索引擎检索到的浅网资源(SurfaceWeb)来说，深网是无法直接通过搜索引擎获取的部分网络资源.深网意味着这些网络资源潜藏在那些静态的、易获取的浅网资源的后面，所示，且其数量巨大.BrightPlanet 公司 2000 年对深网的调查报告中提到几点发现，下面列出其中的一部分:

- 深网资源数量是浅网资源数量的 400-500 倍
- 深网包含 750OTB 的信息，而浅网仅仅包含 19TB 的信息量·
- 深网包含 5500 亿个页面，而浅网则只包含 10 亿个页面;
- 目前(2000 年)存在的深网站点超过 200000 个;
- 60 个最大深网站点包含 750TB 信息，已是浅网的 50 倍二.
- 超过一半的深网内容是关于特定领域的.
- 深网所蕴含的信息质量比浅网高;
- 95%的深网信息都是面向公共服务的，并且免费

5 Web 数据库查询过程

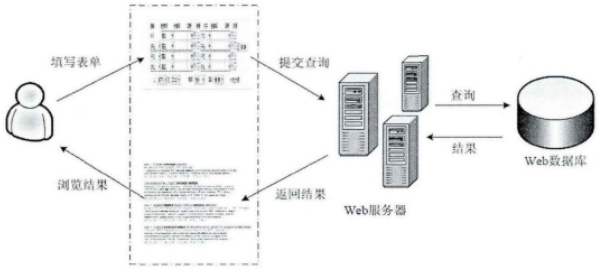

图 2.3　Web 数据库查询过程

1.用户在网站上填写查询表单.一般来说，查询接口分为关键字查询接口和属性值查询接口.如果是关键字查询接口，用户只需要输入待查询的一个或者多个关键字，并且提交查询;而如何是基于属性值的查询接口，用户需要从接口表单中选择合适的属性值或者属性值组合，并且提交查询;
2.浏览器将查询请求提交给网站的 Web 服务器;
3.、Web 服务器根据用户提交的查询表单，动态产生数据库查询请求，并且通过数据库的查询接口将用户的查询请求转交给数据库服务器;
4.数据库服务器执行提交的查询请求
5.数据库服务器将执行后的查询结果返回给 Web 服务器
6.Web 服务器将查询结果动态的包裹到 Web 页而中，并且该页面返回给浏览器;
7.用户通过浏览器查看到查询结果

从以上的 W 已 b 数据库查询过程可以看到，其很难被一般的搜索引擎获取到.由于深网的大规模性、动态性以及异质性等特点，通过手工方式远

远不能在效果和效率上满足用户对信息获取的需要.为了帮助人们快速、准确地利用深网中的海量信息，需要自动化的工具对Web数据库数据进行管理.将 Web 数据库数据抽取到本地是其中一种主要方法.

为了使得 Web 数据抽取过程尽快的结束，查询过程中需要尽可能的以较少的查询次数获取不同查询结果.不同的查询可以获得不同的查询结果，但是也可能获得相同的查询结果.从形式背景 K 的划分格L_H与同形式背景的概念格同态这一点可以看出，多个划分具有相同的概念映射，也就是说不同的查询属性可以导致相同的划分结果.另一方面，依据概念格的性质，不同的概念格具有不同的外延.因此，可以使用概念的内涵表示查询属性;返回的结果就是该概念的外延.这样从 Web 数据库抽取数据的任务就可以转化成在如何在已知的概念空间中，选择合适形式概念作为查询概念的任务;甚至可以使用已知的概念外延在查询过程之前对查询概念的返回结果数量进行预测.这些特性使得概念格这样的数据结构非常适合作为需要限制返回结果数量的情况下的抽取任务.

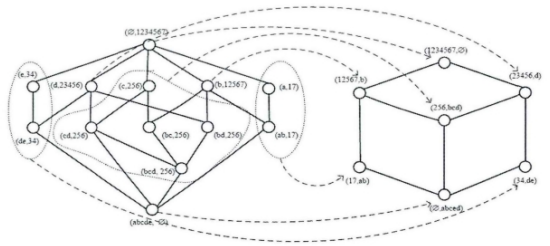

图 2.4　划分格与概念格同态映射（左:划分格 L_H; 右:概念格 L_K）

例 2　图 2.4 显示了由表 2.1 中形式背景K_1所产生的划分格和K_1的概念格之间的同态映射关系.假设表 2.1 中的形式背景K_1是一个深网数据库，用户可以通过发出查询属性 $Y \in P(A)$获取K_1上的信息.为了表示方便使用二元组(Y，Q(Y). O_Y)来代替(Y，Q(Y))表示划分.可以得到如下划分:(a，17)，(b，12567)，(c，256)，(d，23456)，(e，34);(ab，17)，(bc，256)，(bd，256)，(cd，256)，(de，34)，…;(bcd，256)，…，(abcde，∅).因为 $e \subseteq de$，所以(e，34) ≥ (de，34).并且它们属于同一个等价类$[e]_Q$=34，它们映射为概念格L_k中的概念(34，de).根据定义 2.12 划分(b，12567)的势为 5.图 2.4 中详细列出了形式背景K_1的划分格同态于形式背景K_1的概念格的映射情况。

6 总结

由 WilleR 于 1982 年首先提出的形式概念分析提供了一种支持数据分析的有效工具。形式概念包括外延和内涵两部分，它本质上描述了形式对象与形式属性之间的关系。形式背景则以二维表的形式描述了不同形式对象以及他们的形式属性之间的关系。本文借助于形式概念分析的形式化描述能力对受限的 Web 数据库抽取问题进行分析，证明了由属性及属性组合产生的集合划分之间为容差关系，进而又证明其构成一个完全格，并月与概念格同态.因此，使用概念间的偏序关系来刻画属性间的相关性，使用概念内涵为查询属性，概念外延为返回结果的预测。现在，概念格仍是一个高速发展的领域，对于 Web 数据库的抽取问题的研究还需要做很多，比如如何将 Web 数据抽取问题转化为一系列基于形式概念分析的 Web 应用问题等等。

参考文献

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

[2]张卓.基于形式概念分析的 Web 数据库抽取研究[D].武汉大学,武汉，2011.

[3],刘伟,孟小峰,凌妍妍.一种基于图模型的 Web 数据库采样方法[J].软件学报,2008,Vol.19(2):179-93.

[4]李金海,魏玲,张卓,等.概念格理论与方法及其研究展望[J].模式识别与人工智能,2020(7).

[5]毕强，滕广青.国外形式概念分析与概念格理论应用研究的前沿进展及热点分析[J].现代图书情报技术,2010(11)：17-23.

[6]B·GanterandR·Wille·FormalConcePtAnalysis·MathematiealFoundationSIMIBerlin:SPringer 一 Verlag.1999.