《智能信息处理》课程作业

基于形式概念的英语语句结构简析

马盼盼

作业	分数[20]
得分	

2020 年 11 月 23 日

基于形式概念分析的英语语句结构简析

马盼盼

(大连海事大学信息科学技术学院，辽宁大连 110291)

摘 要：英语是当今世界上的国际通用语言之一，也是世界上使用最多的语言。如果把单词比作是英语的血肉，那么语法一定是英语的骨骼。对英语语句结构的分析是形成语法的第一步。而形式概念分析是从形式背景进行数据分析和规则提取的有力工具。因此本文用形式概念分析的方法来对英语语句结构进行初步的分析。

关键词：形式概念分析；概念格；形式背景

Brief analysis of English sentence structure based on formal concept analysis

MA PANPAN

(College of Information Science and Technology，Dalian Maritime University，Dalian 110291，China)

Abstract：English is one of the universal languages in the world today, and it is also the most used language in the world. If you compare words to the flesh and blood of English, then grammar must be the bones of English. The analysis of English sentence structure is the first step in forming grammar. The formal concept analysis is a powerful tool for data analysis and rule extraction from the formal background. Therefore, this paper uses the method of formal concept analysis to make a preliminary analysis of the structure of English sentences.

Key words：formal concept analysis；concept lattice；formal contexts

0 引言

英语是当今世界上的国际通用语言之一,也是世界上使用最广泛的语言。根据以英语作为母语的国家及人数的统计，英语可能是世界上的第三或第四大语言，但它是世界上使用最广泛的第二语言。然而，英语的学习并没有特别容易。如果说单词量是基础，那么语法就是框架。希望学好英语，这两样缺一不可。

形式概念分析是 20 世纪 90 年代 Wille 提出的一种从形式背景进行数据分析和规则提取的强有力工具[1]。概念是人类进行知识表达的手段。知识发现的过程就是将数据中蕴含的知识形式化成有用概念的过程。形式概念分析中的概念在哲学上理解为由外延和内涵两部分组成的思想单元。在形式概念分析中，概念的外延被理解为属于这个概念的所有对象的集合，而内涵则被认为是所有这些对象所共有的特征或属性集。所有的概念连同他们之间的泛化与例化关系可以构成一个概念格。基于概念的这一哲学理解对概念进行了形式化描述，提出形式概念分析可以用于概念的发现、排序和显示。

本文在形式概念分析的理论基础上，对英语语句结构进行初步研究，提高学习英语的效率。

1 形式概念分析相关概念

形式概念分析（Formal Concept Analysis）是应用数学的一个分支,它建立在概念和概念层次的数学化基础之上,根据用二元关系表达的形式背景，从中提取概

念层次结构,即概念格[2]。概念格的每个节点就是一个概念,由两部分组成：外延（Extension）, 即概念所覆盖的实例 ；内涵（Intension）,即概念的描述,该概念覆盖实例的共同特征。

定义 1[3]：一个形式背景（formal context）是一个三元组 K=(G,M,I), 其中 G 是对象的集合，M 是属性的集合，I 是 G 和 M 之间的二元关系。对于 g∈G，m∈M 若（g，m）∈I 表示对象 g 与属性 m 的关系，读作“对象 g 具有属性 m”。记做 gIm。

一般而言，形势背景并不是直接存在的，需要从数据源中提取，即对现有概念中对象和属性进行约简。对象的约简是指将具有一样属性的对象进行合并成为一个形式对象，属性的约简是指将所有对应于同一个对象集的几个属性合并为一个形式属性。不能约简的对象和属性会转换为相应的形式对象和形式属性。将经过对象和属性约简后得到的形式概念与形式属性以形式对象-形式属性集的形式置于二维表中得到的就是形式背景。通过观察构造好的形式背景的列间关系（属性间的关系），可以进行知识发现，也就是关联规则的提取。

定义 2：对于一个对象集 A，定义 A'={m∈M|gIm, 对所有 g∈A}(A 中所有对象共有的属性集合)。相应地，对于一个属性集 B，定义 B'={g∈G|gIm,对所有的 m∈B}(即包含所有 B 中属性的对象集合)。

定义 3：语境的形式概念（formal concept）是一个集合对（A,B)。其中：A⊆G,B⊆M 并且 A'=B,B'=A。A、B 分别称做概念的外延和内涵。β(G,M,I)表示语境 k=(G,M,I)中的所有概念集合。

定义 4[4]：概念格（concept lattice）对于形式概念(A1 ,B1),(A2 ,B2)均是语境中的概念，并且 A1⊆A2 ，那么(A1,B1) 被称做(A2 ,B2)的子概念，(A2 ,B2) 则是(A1 ,B1)的超概念，记为(A1 ,B1)≤(A2 ,B2)，“≤”反映了概念间的层次关系。这种形式背景中所有形成概念的子概念－ 超概念的偏序关系所诱导出的格就是概念格（Concept lattice)。

概念格的构造通常是首先绘制与形式背景对应的 Hasse 图，然后通过补齐各形式概念的上下确界，进而形成概念格。而概念格的构造是形式概念分析应用的前提，目前构造概念格的算法主要可以分为三大类：批处理算法、增量算法和并行算法。

2 分析英语语句结构

2.1 形式背景

以英语语句作为对象，语句中的成分作为属性，以此建立形式背景。如表 1 所示，a 表示主语，b 表示谓语，c 表示宾语，d 表示表语，e 表示定语，f 表示状语，g 表示补语。其中对象集为{S1,S2,S3.....S10}，属性集为{a，b，c，d，e，f，g}。其中概念可以形式化为序偶（语句，成分）二元组；形式背景表现为（语句，成分，包含关系）的三元组，用 1 表示此语句中包含此成分，而用 0 表示此语句中不包含此成分。

表 1 英语语句形式背景

语句	主语	谓语	宾语	表语	定语	状语	补语
S1	1	1	1	0	0	0	0
S2	1	1	1	0	1	1	1
S3	1	1	0	0	0	0	0
S4	1	1	0	1	0	0	0
S5	1	1	1	0	1	0	0
S6	1	1	1	0	0	0	0
S7	1	1	1	0	0	0	1
S8	1	1	1	0	0	1	0
S9	1	1	0	0	1	0	0
S10	1	1	0	0	0	0	0

对所示的形式背景进行按照概念格的生成步骤对其进行形式背景约简。把属性值相同的对象进行约简，将相同的属性进行约简。S1 和 S6 语句的成分相同可以合并，S3 和 S10 语句的成分相同可以合并。那么最后得到的形式对象集合为{S1&S6,S2, S3&S10, S4, S5，S7，S8，S9}，最后得到的形式属性集合为{主语，谓语，宾语，表语，定语，状语，补语}，约简后得到的形式背景如表 2 所示。

表 2 约简后的形式背景

语句	主语	谓语	宾语	表语	定语	状语	补语
S1，S6	1	1	1	0	0	0	0
S2	1	1	1	0	1	1	1
S3，S10	1	1	0	0	0	0	0
S4	1	1	0	1	0	0	0
S5	1	1	1	0	1	0	0
S7	1	1	1	0	0	0	1
S8	1	1	1	0	0	1	0
S9	1	1	0	0	1	0	0

由于多值形式背景不易进行形式概念分析，因此将多值形式背景转换为单值形式背景。在表 2 中用 1 到 10 代替各个语句，用 a 到 g 代替各个属性。并且用“×”来代替表中值为 1 的位置，值为 0 的位置则留白。由此得到单值形式背景，如表 3 所示。

表 3 单值形式背景

	a	b	c	d	e	f	g
1,6	×	×	×				
2	×	×	×		×	×	×
3,10	×	×					
4	×	×		×			
5	×	×	×		×		
7	×	×	×				×
8	×	×	×			×	
9	×	×			×		

对所生成的单值形式背景按照概念格的构造方法将单值形式背景转化为带有父子关系的单值形式背景，原则是基于属性个数的数量进行排序。转化结果如表 4 所示。

表 4 带有父子关系的单值形式背景

	a	b	c	d	e	f	g
3,10	×	×					
1,6	×	×	×				
4	×	×		×			
9	×	×			×		
5	×	×	×		×		
8	×	×	×			×	
7	×	×	×				×
2	×	×	×		×	×	×

2.2Hasse 图

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 4 所绘 Hasse 图如图 1 所示。

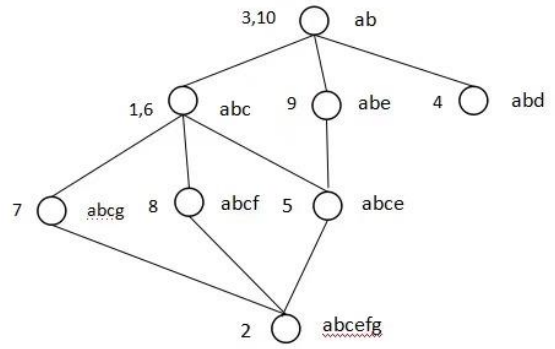

图 1 Hasse 图

Hasse 图中的每个结点表示集合 A 中的一个元素，结点的位置按所在偏序中的次序从底向上排列。即对任意 a、b 属于 A，若 $a<b(a\leqslant b\wedge a\neq b)$，则 a 排在 b 的下边。如果 $a<b$，且不存在 $c\in A$ 满足 $a<c<b$，则在 a 和 b 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。

2.3 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。

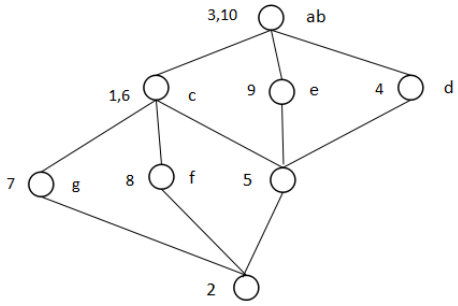

图 2 概念格

3 基于概念格的决策

基于概念格的决策包括概念识别以及概念推理[5]。概念识别，是指从与特定论域对应的概念格中识别其中的形式概念并且识别形式概念之间的关系。观察图 2 所示的概念格，可以识别到：

(1) 语句 9 是语句 3，10 的子概念，即语句 3，10 所包含的成分语句 9 也都包含；

(2) 语句 3，10 是语句 1，6、语句 9、语句 4 的父概念，即语句 3，10 所增加的成分，语句 1，6、语句 9、语句 4 也可以添加。

概念推理，是通过在概念格上的结点之间的移动，根据结点所表示的形式概念之间的关系，进行推理的过程。例如，观察图 2 所示的概念格，则可以得到一个英语语句中必须要含有主语与谓语的规则。

4. 结语

本文通过基于多种英语语句构建形式背景，给出了从概念转化为形式概念、背景转化为形式背景、约简形式背景转化为单值形式背景再构造概念格的整体过程，全面分析其特征和关系。最后进行了形式概念的识别与推理，分析了英语语句中的成分也从中充分体会到形式概念分析以及概念格，在知识发现推理、Web 语义检索和数据挖掘中的重要作用。

参考文献：

[1] [德]B.甘特尔,R.威尔.形式概念分析[M].马垣,张学东等译.北京：科学出版社,2007.

[2] R. Wille, "Methods of conceptual knowledge processing", Formal Concept Analysis 4th International Conference ICFCA 2006, vol. 3874, pp. 1-29, February 13-17, 2006.

[3] Baidu.Formal Concept Analysis http://baike.baidu.com/4660144.htm[OL],2018,11,1.

[4] 黄映辉.智能信息处理课件:形式概念分析_第 4 章形式概念分析[R].大连海事大学,2014.

[5] [德]B.甘特尔,R.威尔.形式概念分析[M].马垣,张学东等译.北京：科学出版社,2007.