基于形式概念分析的词义理解研究

胡蕊
（大连海事大学 信息科学技术学院，辽宁 大连 110310）

摘 要: 探讨了形式概念分析在自然语言理解中的多义词分析及义素分析中的应用。在对多义词进行分析时，根据词性与词义之间的二元关系，构造词性与词义的决策背景，进而发现了词性与词义之间的决策规则，并对这些决策规则进行了解释；在对义素分析进行研究时，根据语言对象及其语义特征之间的二元关系构造形式背景，并在此背景上分析语言对象，实例表明了形式概念分析既可以很容易地对词语进行分类，又可以很直观地反映词类之间的关系，是进行义素分析的一种有效工具。

关键词: 形式概念分析，概念格，多义词，义素分析
中图法分类号 TP393 **文献标识码** A

Study on Word Meaning Comprehension Based on Formal Concept Analysis

HU Rui
(Faculty of Information Science and Technology，Dalian Maritime University，Dalian 110310，China)

Abstract: We discussed polysemy analysis and semantic analysis based on formal concept analysis．On polysemy analysis，we generated a decision context based on relations between word meanings and the corresponding parts of speech．And then we extracted the decision rules from the context and gave them explanations．On semantic analysis，we formed a formal context based on relations between linguistic objects and their semantic features，and analyzed them according to formal concept analysis．Experiments show that we can not only classify the linguistic objects easily，but view relations between different word classifications intuitively，and that formal concept analysis is an efficient tool for semantic analysis．

Keywords: Formal concept analysis，Concept lattice，Polysemy，Semantic analysis

1 引言

目前，在自然语言学(计算语言学)中，如果可以通过汉语词法、句法、语义等知识库的学习发现其中隐含的规律，将有助于计算机对现代汉语的理解。而形式概念分析上的规则提取可以有效提取知识库中的隐藏的规则，因此利用形式概念分析对语言知识进行研究是一个很有意义的课题。

本文利用语言学中的语言对象及其语义特征构造对象与属性间的二元关系，根据此二元关系构造语言学上的形式背景，并对汉语中常见多义词进行分析，得出多义词的义项和以该义项为中心词素所构成词的词性之间的决策关系。其次，将形式概念分析的方法用于义素分析，在语义场中对词进行了分析，并通过实例验证了方法的可行性。

2 形式概念分析的基本概念

一个形式背景 K 是一个三元组：$K=(G,M,I)$，其中 G 为所有对象的集合，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。对于 $g\in G,m\in M,(g,m)\in I$ 表示“对象 g 具有属性 m”。

设 $K=(G,M,I)$ 为一形式背景。对于集合 $A\subseteq G$，记：

$$A^{'}=\{m\in M\mid (g,m)\in I,\forall g\in A\}$$

相应地，对于集合 $B\subseteq M$，记：

$$B^{'}=\{g\in G\mid (g,m)\in I,\forall m\in B\}$$ 。

为方便起见，我们用 $g^{'}$ 表示 $\{g\}^{'}$，用 $m^{'}$ 表示 $\{m\}^{'}$。

设 $K=(G,M,I)$ 为一形式背景，$A\subseteq G$，$B\subseteq M$，称 $C=(A,B)$ 为 K 的一个概念，如果 $A^{'}=B$ 且 $B^{'}=A$，此时称 A 为 C 的外延，B 为 C 的内涵。我们用 $\underline{B}(K)$ 记 K 的所有概念组成的集合。

设 $K=(G,M,I)$ 为一形式背景，$C_1=(A_1,B_1)$，$C_2=(A_2,B_2)$ 是 K 的两个概念，规定：

$$C_1\le C_2\Leftrightarrow A_1\subseteq A_2(\Leftrightarrow B_1\supseteq B_2)$$

此时，C_2 称为 C_1 的超概念，C_1 称为 C_2 的子概念，概念间的关系称为泛化和例化关系。

显然，关系“≤”是集合 $\underline{B}(K)$ 上的一个偏序，它可诱导出 $\underline{B}(K)$ 上的一个格结构，可以证明，它是一个完备格，相应的下确界和上确界定义为：

$$\underset{t\in T}{\wedge}(A_t,B_t)=(\bigcap_{t\in T}A_t,(\bigcup_{t\in T}B_t)^{''})$$

$$\underset{t\in T}{\vee}(A_t,B_t)=((\bigcup_{t\in T}A_t)^{''},\bigcap_{t\in T}B_t)$$

其中 $(A_t,B_t)\in\underline{B}(K)$，T 是指标集。此完备格称为形式背景 K 的概念格，在没有歧义的情况下，仍然记为 $\underline{B}(K)$。

概念格的图形可视化可以用 Hasse 图来表示。生成图的方法如下：如果 $C_1\le C_2$，且格中没有概念 C_3 使得 $C_1\le C_3\le C_2$，那么就存在一条从 C_1 到 C_2 的边。在图中，我们使用黑色的格点表示形式概念，通过线段表示了概念之间的泛化和例化关系。

3 形式概念分析在多义词分析中的应用

词义是词所固有的内容，就是词所代表的被人们用来称说的事物。词义与词并不一定是一对一的：一个词可以有一个意义，也可以有多个意义。多义词的每一个意义都是一个语义单位。词典释义中的所谓义项，就是指词义中能够确定下来的这类单位。当以多义词作为中心词素构词时，使用的是中心词的不同义项，并且构成词的词性也不一定相同。在自然语言中，多义词现象极其普遍，人们对多义词的研究成果较多。

我们将利用自然语言中词义与词性之间关系，构建合适的决策形式背景，进而得到词组的词性与中心词的义项之间的决策规则关系。

我们以“浅”为中心词素构词来进行说明。从构成词的词性方面看，可得到“名词”、“动词”、“容词”、“副词”，我们取前 3 种常见词性作为条件属性；从构成词所使用的义项上看，“浅”包含了“距离小”、“浅薄”、“浅显”、“程度不深”、“颜色淡”、“历时短”等 6 种义项。我们采用前 4 种常用义项作为决策属性。由此构成一个关于浅的决策背景，如表 1 所列。

表 1 “浅”构成的决策背景

	名词	动词	形容词	距离小	浅薄	浅显	程度不深
浅海	1	0	0	1	0	0	0
短浅	0	0	1	1	1	0	0
深浅	1	0	1	1	0	0	1
浅薄	1	0	1	1	1	1	1
粗浅	0	0	1	0	1	0	0
浅陋	0	0	1	0	1	1	1
浅明	0	0	1	0	0	1	1
浅显	0	0	1	0	0	1	0
浅尝	0	1	0	0	0	0	1

为了表述方便，以下将用表中的“a”，“b”，“c”分别表示“名词”、“动词”、“形容词”这 3 个词性；“d”，“e”，"f"，“g”分别表示“距离小”、“浅薄”、“浅显”和“程度不深”这 4 个义项。“1”，"2"，“3”，“4”，“5”，“6”，“7”，“8”，“9”分别代表“浅海”、“短浅”、“深浅”、“浅薄”、“粗浅”、“浅陋”、“浅明”、“浅显”和“浅尝”9 个对象。我们可以通过观察它们所在行及所具有的各列属性值，来判断该构成词所具有的词性以及中心词素所使用的义项。“1”表示“具有该列属性值”，“0”表示“不具有该列属性值”。比如，由“浅海”所在的行及所取的各列属性值，我们可以看出“浅海”是名词，其中“浅”使用的是“距离小”这一义项。

生成对应的概念格，如图 1 所示 。

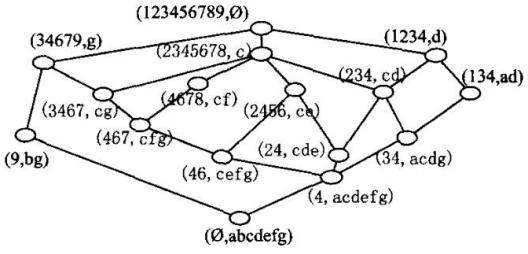

图 1 表 1 的概念格的 Hasse 图表示

我们可以通过概念格上的决策规则提取方法从概念格图 1 上提取决策规则。所获得的决策规则如表 2 所列。

表 2 图 1 中提取的决策规则

序号	决策规则	实用规则
1	{a}→{d}	名词→距离小
2	{b}→{g}	动词→程度不深
3	{ac}→{g}	名词＋形容词→程度不深

以上是以“浅”作为中心词素来说明形式概念分析方法在多义词理解上的可行性。应用此种方法可以使计算机有效地理解自然语言中的词义。但是由于语言的复杂性，如何找到合适的词组来构造形式背景，是我们需要考虑的问题；其次，要为所有的多义词建立相应的决策背景，并在此基础上提取决策规则是一项既繁琐又庞大的工作，如何进一步完善与精简该工作，也是一个值得进一步研究的课题。

4 形式概念分析在义素分析中的应用

义素分析是现代语义学的一个重要成果，在语言学研究中已被广泛地研究。它从词的内部微观层次，揭示了词义之间的区别和联系，能较好地解释语义的组合与聚合规律。另外，义素分析通过对词义的细致分析，为词义分析的形式化和精确化提供了一种新方法。

将形式概念分析应用在义素分析上，可以使义素分析更好地进行形式化分析，同时对词的分类、搭配等具有较好效果。

用形式概念分析的方法进行义素分析大致可以分为 4 个步骤。

(1)选取语义相关的词作为义素分析对象；

(2)提取待分析对象合适的语义特征。通过对词的语义分析，提取可以较好地反映待分析对象的一些词或论述的语义特征，根据这些语义特征可以有效地对义素对象进行归类。

(3)构建义素分析的形式背景。以步骤(1)中所选取的对象为背景的对象，以步骤(2)中提取的语义特征为背景属性，构建义素分析的形式背景。

(4)根据步骤(3)中所建立的形式背景构造概念格，并用格的节点来对词进行分类。

在动词中，有一些词表示了说话者对自己所说内容的态度，将由这些词构成的语义场称为主张动词语义场。下面将利用形式概念分析对主张动词语义场进行分析。首先选取一些动词作为分析对象，这里选取了“鼓励”、“引导”、“诱导”、“激励”、“教导”、“煽动”、“教唆”、“指导”这 8 个词。其次，我们既要选取可以表示这些动词的描述作为语义特征，又要找到区分这些动词的描述的语义特征，这些词在使用中意思大致相当，但在反映说话者对所说内容的态度上有所不同。这里我们提取到“说话者支持自己所说内容”、“说话者不支持自己所说内容”、“说话者批判自己所说内容”和“说话者不批判自己所说内容”4 种特征。根据这 4 种特征来对上述对象进行义素分析，我们可以构建主张动词语义场的形式背景，如表 3 所列。

表 3 主张动词语义场的形式背景

	说话者支持自己所说内容	说话者不支持自己所说内容	说话者不批判自己所说内容	说话者批判自己所说内容
鼓励	1	0	1	0
引导	0	1	1	0
诱导	0	1	0	1
激励	1	0	1	0
教导	0	1	1	0
煽动	0	1	0	1
教唆	0	1	0	1
指导	0	1	1	0

由表 3 可以看到，对象“鼓励”与“激励”都只具有“说话者支持自己所说的内容”和“说话者不批判自己所说内容”这两个特征，也就是说，从义素分析的角度上讲，它们具有相同的义素特征。从形式背景角度出发，对象“鼓励”和“激励”都满足属性“说话者支持自己所说内容”和“说话者不批判自己所说内容”，而不满足属性“说话者不支持自己所说内容”和“说话者批判自己所说内容”。因此我们采用对象约简的观点，把表 3 中第 1 行和第 4 行进行合并(约简)，同理也可以对其它对象进行合并，进而得到表 3 的约简形式背景表 4。为了便于表示，我们分别用“a”，“b”，“c”，“d”表示表 3 中的“说话者支持自己所说内容”、“说话者不支持自己所说内容”、“说话者不批判自己所说内容”及“说话者批判自己所说内容”这 4 个语义特征，使用“1”，“2”，“3”表示表 3 中具有完全相同语义特征的 3 个对象组。

表 4 表 3 的形式化背景

	a	b	c	d
1	1	0	1	0
2	0	1	1	0
3	0	1	0	1

从表 4 可以得到其概念格，如图 2 所示。

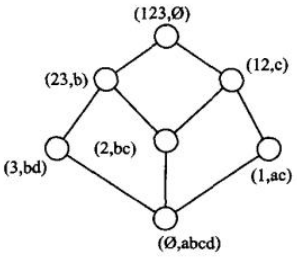

图 2 表 4 对应的概念格

我们从概念格上，可以很容易地对词语进行分类，并且依据语义特征细化的程度不同，可以获得不同的分类效果。通过概念格上的概念间泛化例化关系，可以更好地区分词义，进而可以根据所要表达的具体语义特征选出最合适的一类词或一个词，也可以从概念间的关系对词语进行更好的理解。

义素分析实际上是对词义进行形式化描述，通过形式化的描述使我们可以更直观地认识和理解词义。理解词义是自然语言理解过程中所必需的阶段，如何更好地理解词义，对建立机器翻译系统、人工智能系统等都有着非常重要的意义。义素分析已在语言学中有了广泛的应用，如何采用恰当的方法进行义素分析对于语言学的研究极为重要，而采用形式概念分析对义素分析进行研究是一个新的课题。如何更好地将形式概念分析的方法应用到义素分析中，是我们仍需继续研究的课题。

5 总结

本文利用语言学中的语言对象及其语义特征构造对象与属性间的二元关系，根据此二元关系构造语言学上的形式背景，并对汉语中常见多义词进行分析，得出多义词的义项和以该义项为中心词素所构成词的词性之间的决策关系。其次，将形式概念分析的方法用于义素分析，在语义场中对词进行了分析，并通过实例验证了方法的可行性。

目前，已有一些自然语言工作者将粗糙集的方法用于自然语言分析，但将形式概念分析的方法应用到自然语言理解领域的研究并不多见。本文将形式概念分析的方法用于多义词理解和义索分析中，并探讨了该方法的合理性。如果能将形式概念分析的方法更好地应用到自然语言处理中，不论是对语言学研究还是自然语言理解都将具有更大的实用价值。

参考文献

[1]曲开社，梁亮，梁吉业等．形式概念分析的概念之间包含度理论[J]．计算机科学，2009，36(2)：210—219

[2]孔昭琪．关于多义词研究中的几个问题[J]．山东师大学报，1993，1：77-80

[3]符淮青．词义单位的划分和义项[J]．辞书研究，1995，1：75—83

[4]田兵．多义词的认知语义框架与词典使用者的接受视野[J]．现代外语，2003，26(4)：339—350

[5]Ganter B，Wille R. Formal Concept Analysis：Mathematical Foundations[M]. Berlin：Springer-Verlag，1999

[6]Qu K S，Liang J Y，Wang J H，et al. The algebraic properties of concept lattice[J]. Journal of Systems Science and Information，Research Information Ltd U K，2004，2(2)：271～277

[7]何象林，孔鸿滨，姚绍文．基于语义的工作流数据模式研究[C].国际信息技术与应用论坛论文集．2009

[8]陆歌皓，李昆蔓，夏寿民．DSC--新的业务流程建模方法[J]．计算机科学，2008，35(8专刊)

[9]曲开社，闰俊霞，翟岩慧．GM偏序图的构建和基于GM偏序图的规则提取[J]．计算机工程，2007，43(36)：51—54

[10]HAN Jia- wei，Kamber M. 数据挖掘：概念与技术［M］．范明，孟小峰译．北京：机械工业出版社，2007:268-269.

[11]Skowron A. Extracting laws from decision tables: a rough set approach[J]. Computational Intelligence, 1995, 11(2): 371-388.

[12]Berry M W, Drmac Z, Jessup E R. Matrices, vector spaces, and information retrieval[J]. SIAM review, 1999, 41(2): 335-362.

[13]李艳霞，史一民，李冠宇．基于概念格的K—Means算法研究[J]．计算机工程与设计，2011，32(2)：656-658.

[14]Godin R, Mineau G, Missaoui R. Incremental structuring of knowledge bases[C]. Proc. of KRUSE. 1995, 95: 179-193.