基于 HNC 概念关联性的领域判定研究*

池哲洁[1,2]，张全[2]

（1. 中国科学院研究生院，北京 100049；2. 中国科学院声学研究所，北京 100190）

摘要：在概念层次网络理论中，领域是语境单元的一个要素，而领域判定是语境单元萃取的重要课题之一。本文提出一种利用领域概念以及概念关联式进行领域判定的方法，通过在概念基元层面进行频数统计、概念合并及概念汇总实现领域的判定。对政治、经济、军事三个领域的语料进行测试，结果表明，使用概念关联式能够改进领域判定的效果，其 F_1 值分别达到 90.61%、90.83%、90.99%，比不使用概念关联式的情况分别提高了 7.7%、12.76%、5.01%。最后，与基于关键词方法的对比结果也显示使用概念基元的方法效果较好。

关键词：概念基元；概念关联式；领域判定

Domain Determination Based on HNC Concept Association

Chi Zhejie[1,2], Zhang Quan[2]

(1. Graduate University of Chinese Academy of Sciences, Beijing 100039, China;

2. Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: For Hierarchical Network of Concepts theory, domain is one of the main factors in Sentences Group Unit. Domain determination is an important issue in Sentences Group Unit Extraction. To determine the domain, we proposed a method using domain concepts and concept association expressions, which counted frequencies, merged concepts and summarized concepts in concept primitive space. For politics, economics and military domain, the experimental results show high performance in present method, the F_1 scores reach 90.61%, 90.83% and 90.99% respectively, which are 7.7%, 12.76% and 5.01% higher than the results with no concept association expressions. Finally, compared with the keywords-based method, the concept-primitives-based method shows high performance.

Key words: Concept Primitives；Concept Association Expressions；Domain Determination

1 引言

概念层次网络（简称 HNC）理论[1-3]立足语言概念空间，认为语言概念空间是一个由基层（对应概念基元空间）、第一介层（对应句类空间）、第二介层（对应语境单元空间）和上层（对应语境空间）构成的四层级结构体，在此基础上建立了自然语言的理解和处理的新模式。在概念基元空间中，按人类活动以及其他生命体本能活动、自然界灾祸状态可划分出十大领域类：心理活动及精神状态、人类思维活动、专业及追求活动（第二类劳动）、理念活动、第一类劳动、业余活动、信仰活动、本能活动、灾祸、状态，这一分类的主体是对人类活动所属范畴的分类[2]。

领域用来描述事件的类型，确定事件核心归属的范围，它和情景、背景共同构成语境单元的三要素。领域是语境单元的第一要素，同时，情景和背景也是领域的函数，即领域也能决定情景和背景，说明领域在语境单元空间中起主轴作用。在语境单元萃取（实现从句类空间到语境单元空间提升的过程）中，领域判定是其八大课题中的重要一项，领域信息的获取能够为语境单元萃取提供原始材料。

在已有工作中，对领域的研究主要是在句类和句群层面，韦向峰[4]设计了句类分析平台，在语段对领域、情景及背景等基本信息的获取进行研究；缪建明[5]采用句类表示式的方法将

*本文承 国家高技术研究发展计划(863 计划)“十二五”计划项目课题“基于云计算的海量文本语义计算框架与开放域自动问答验证系统”(2012AA011102)，国家语委“十二五”科研项目“基于概念空间的语义关联研究”(YB125-53)，中科院信息化项目“科技资源聚合服务”，中科院声学所知识创新工程项目“音频内容分解与检索”(Y154141431)资助

领域知识组织起来，形成领域句类知识，为句群处理提供便利。本文尝试一种在概念基元层面实现领域判定的方法：直接利用领域概念，并结合概念关联式，通过频数统计直观的呈现领域信息，最终基于频数比较对领域做出判定。

2 HNC 概念关联式介绍

HNC 概念关联中，除了 HNC 逻辑符号定义的关联外，还存在 10 种沿袭逻辑[6]关联类型（如表 1 所示）。

表 1 10 种沿袭逻辑关联型

关联类型	符号	示例
强关联	≡	a219\10*b\25≡a42
强交式关联	=	3099=107a
强流式关联	<=	jl12<= 53
强源式关联	=>	7103^e46d01=>a60
包含	%=	q701e22%=a72^e21
属于	=%	a228i\3=%a59a
对应	:=	a11e1ne223:=a109
等同	=:	a15=:a13\1d01
定义	::=	73228::=(7322,l83,d22)
虚设	==	a103e22==a143

对以上 10 种沿袭逻辑关联说明如下：

强关联（≡）的两个节点关联性强，有些可视为同一个，示例中 a219\10*b\25（战争资源基建）与 a42（战争）强关联。

强交式关联（=）表示两个节点在 HNC 作用效应链各环节具有交织性表现，它是同一个概念本体从不同观察角度看到的不同映像，如：3099（反复）是 107a（过程周行性）的效应描述，强调实现过程的非单调性。

源、流式关联（=>、<=）展示概念的源流关系，源和流是对偶的两端；流式关联中，前者是流，后者是源，源式关联则相反，如：7103^e46d01（好奇）是 a60（探索与研究）的起因，故 7103^e46d01 强源式关联于 a60。

包含、属于（%=、=%）是一种父集合和子集合的关系，在包含关联中，前者的一部分是后者，属于关联则表示前者是后者的一部分，如：q701e22（讲谈）包括 a72^e21（教）。

对应（:=）一般表示条件关系，如：a11e1ne223（王朝更迭）对应于 a109（王权制度）。

等同（=:）是弱定义式，如：a15（征服）是 a13\1d01（国家、民族之间政治斗争）的最高级形式。

定义（::=）表示一个节点的内涵可以通过其他节点或多个节点的组合来阐述，一般被定义者是单一概念节点，而定义者则是多个概念节点的组合，如：73228（先验理性行为）定义为基于先验理性的理性行为（7322（行为与理性），l83（主客观因素的综合），d22（先验理性））。

虚设（==）表示前者是后者的虚设，前者是为了体现概念的完整性而设置的，其具体延伸见后者，如：a103e22（对外政策）是 a143（外交政策）的虚设。

不同概念关联式反映概念间关联性的强弱以及概念的关联方向，强关联、强交式关联两端的概念关联性强，且作用是相互的；源、流式关联及包含、属于则是弱一级的关联式，且具有方向性。对于同一概念，可能存在多个概念关联式，比如：a12in（治国基本方式的第一要点）和 a53e2m（治国的文武之道）强关联，同时也和 a123e2m3（民政）强交式关联。概念之间存在关联关系，对于不同的概念，则可考虑应用关联关系将它们进行合并，以缩减所要处理的概念，这样能够为后续判断提供便利。本文基于现有的概念关联式，着重考虑包

含领域信息的概念节点，通过对文本中的概念进行频数统计实现文本的领域判定。

3 领域判定算法描述

本文对文本的领域判定主要基于带有领域信息的概念节点并结合概念关联式进行的。通过词语和概念基元的映射关系，将文本的词形对应到概念基元空间中，实现文本领域特征的第一次压缩；获取概念基元统计信息后，利用已有的概念关联式，对概念进行合并，实现文本领域特征的进一步压缩；对于合并后的概念基元，则考虑按概念层次进行汇总，最终形成能够直接用于领域判定的结果。

领域判定算法描述如下：

(1) 对待判定文本进行分词处理，完成分词后，转(2)。

(2) 利用词语-概念基元映射表，基于切分好的词语统计概念基元信息，形成概念基元-频数表，转(3)。

(3) 基于已有的概念关联式，将概念基元信息进行合并处理，形成合并后的概念基元与其频数的对应表，转(4)。

(4) 对合并后的概念基元按概念层次进行汇总，最终形成可直接用于领域判断的概念-频数对应表，转(5)。

(5) 基于最终汇总的概念对文本领域做出判断。

3.1 概念基元信息统计

要基于带有领域信息的概念节点对领域做出判定，必须先获取概念基元信息。本文以词语作为处理单位，利用现有的词语-概念基元映射表进行概念基元统计，因此，此处主要进行的处理是词语切分和词语到概念基元的映射。

本文的分词工作是采用汉语分词系统（NLPIR，又名 ICTCLAS2013）[7]来完成的。该分词系统给定默认的词典，但与本文采用的词语-概念基元映射表有一些出入，部分映射表中的词语不在该分词系统的词典中，为了能够更充分地利用已有的映射信息，我们对词语-概念基元映射表的词语进行处理，抽取映射表中未能被分词系统切分成单一词语的项目，将其整理后作为用户词典加入到分词系统中。

完成词语切分后，利用词语-概念基元映射表将词语与概念基元对应，统计概念基元信息，此处对切分出的所有词语均进行统计，但对于词语-概念基元映射表中未出现的词语则因缺乏映射标准不予考虑。理论上，将词语和概念基元对应需要一个精确的标注过程，考虑到本文是以判定领域为目的，添加非领域信息对结果判定不会产生太大影响，故在此处进行从简处理，将所出现词语对应的概念基元一并统计，以此结果作为概念基元的统计结果。

3.2 概念关联式的应用

据统计，10 种沿袭逻辑的概念关联式共有 3641 个，各类沿袭逻辑的分布为（括号中的数目为该类型关联式的数目）：强关联（436）、强交式关联（908）、强流式关联（595）、强源式关联（293）、包含（25）、属于（160）、对应（780）、等同（127）、定义（282）、虚设（35）。共有 2779 个概念基元挂靠了概念关联式，这些概念中，绝大多数只有一个概念关联式与之对应，少部分概念有多个概念关联式。

不同类型的关联式所体现出的概念关联性强弱程度不一样，这样，对于某一关联式类型，可以赋予其一关联权重 $\mu_{AssociationType}$，表示在特定关联式下，将一概念向另一概念进行合并时，可以保留 $\mu_{AssociationType}$ 的原概念信息。例如，对于强关联（≡）类型，其两个概念节点可视为一个，故可取 $\mu_{\equiv}=1$，认为合并后的概念能够完全保留原概念信息。为各类型关联式赋予关联权重后，将不同概念节点进行合并则具有量化指标，可直接应用于基于频数的计算中。

考虑到不同类型关联式的关联性强弱以及各关联式在领域判定中所能起的作用，本文选取关联性较强的 7 种关联式，分别是：强关联、强交式关联、强流式关联、强源式关联、包含、属于、等同，这些关联式形成一个键－值对结构的关联概念对应表，键项对应待合并概

念节点，值项则是合并后的概念节点。对关联式选取的总体准则是：尽量让包含领域信息的概念节点作为合并后的概念，即尽量让领域概念出现在关联概念对应表中的值项。由于对各类型关联式的关联性没有先验量化指标，并且领域概念对领域判定的贡献可以体现为“有”或“无”这样的布尔选项，本文在计算关联权重时做一简化处理，对各类型关联式其关联权重均取 $\mu_{AssociationType}=1$，对于各类型关联式，本文采用的具体选取准则为：

(1) 对于强关联、强交式关联、等同，出现在两边的概念地位一致，若两端节点均包含领域信息，则不选用；若仅一方包含领域信息，则将不含领域信息的概念向包含领域信息的概念合并，形成的概念关联式加入关联概念对应表中；若两端节点均不含领域信息，则按原形式将概念关联式加入关联概念对应表。

(2) 对于源、流式关联，将源向流合并，形成统一形式的概念关联式，若源端节点包含领域信息，则该关联式不选用，否则，将概念关联式加入关联概念对应表。

(3) 对于包含、属于，将“包含”的后者向前者合并，“属于”的前者向后者合并，形成统一形式的概念关联式，若关联式前端节点包含领域信息，则该关联式不选用，否则，将概念关联式加入关联概念对应表。

(4) 对于一个概念节点有多个概念关联式与之对应的情况，优先选取合并后值项带有领域信息的关联式，若带有领域信息的关联式有多个，则按强关联-强交式关联-包含、属于-源、流式关联的顺序选取。

基于上述选取准则并经过人工校对，本文最终选取了 866 个概念关联式，形成关联概念对应表，应用于概念节点的合并中。

3.3 概念基元按层次汇总

应用概念关联式对概念进行合并后，所得到的概念层次不一，无法直接应用于领域的判定，因此，需要考虑对其中包含领域信息的概念做汇总处理，以形成能够直接用于领域判定的结果。

HNC 的概念具有层次性和网络性，对概念按照概念范畴-概念林-概念树-根概念-概念延伸结构的方法从高层到低层来表示, 后一层面的概念是前一层面概念的延伸，延伸的概念处于更为底层的地位，越往底层，概念表达的意义越具体、越特殊。结合 HNC 关于领域概念的设定，本文将概念基元汇总到概念林层面，依此对文本的领域做出判定。以专业活动领域为例，在概念林层面共有 9 个节点：a0（专业活动基本共性）、a1（政治）、a2（经济）、a3（文化）、a4（军事）、a5（法律）、a6（科技）、a7（教育）、a8（卫保）。

经过不同层次的概念映射及人工筛选，本文形成一个概念基元-概念林对应表，将其应用于概念基元的汇总。对于汇总的概念林层次结果，本文采用简单的频数对比方法对领域做出判定：对汇总概念林结果进行降序排列，选取其中出现频数最高的领域概念所对应的领域作为判定结果，若其中未出现包含领域的概念林节点，则判定为“其他”领域。

4 实验结果与分析

4.1 实验设置

本文的测试语料来源是国家语委现代汉语语料库（以下简称语委语料库），主要考察专业活动中的领域，从语委语料库的中选取政治、经济和军事三个领域的文本进行测试，为每个领域选取 120 篇文本作为测试语料。

实验评价指标采用正确率（Precision，简记 P）、召回率（Recall，简记 R）和 F_1 值，对某个领域文本的判定结果，a 表示确实属于该领域的文本数，b 表示不属于该领域而被误判为该领域的文本数，c 表示属于该领域却没有被判定为该领域的文本数，则正确率、召回率和 F_1 值的计算分式分别为：

$$\text{正确率：}\ P=\frac{a}{a+b}$$

召回率：$R=\frac{a}{a+c}$

$$F_1=\frac{2PR}{P+R}$$

F_1 值综合考虑了领域判定的正确率和召回率，对判定效果的评价更为全面，作为此次判定的评价指标。

4.2 实验结果

对选取的语料采用本文的方法进行测试，同时，对于不应用概念关联式的情况也进行了测试，作为本文方法的对比，所得结果如表 2 和表 3 所示。

表 2 基于概念关联式的领域判定结果

机器判定 / 预定领域	政治	经济	军事	其他	召回率(%)
政治	111	2	0	7	92.50
经济	5	104	1	10	86.67
军事	9	3	101	7	84.17
正确率(%)	88.80	95.41	99.02		

表 3 不使用概念关联式的领域判定结果

机器判定 / 预定领域	政治	经济	军事	其他	召回率(%)
政治	97	14	1	8	80.83
经济	11	89	1	19	74.17
军事	6	5	92	17	76.67
正确率(%)	85.09	82.41	97.87		

两种方法 F_1 值的对比情况如图 1 所示。

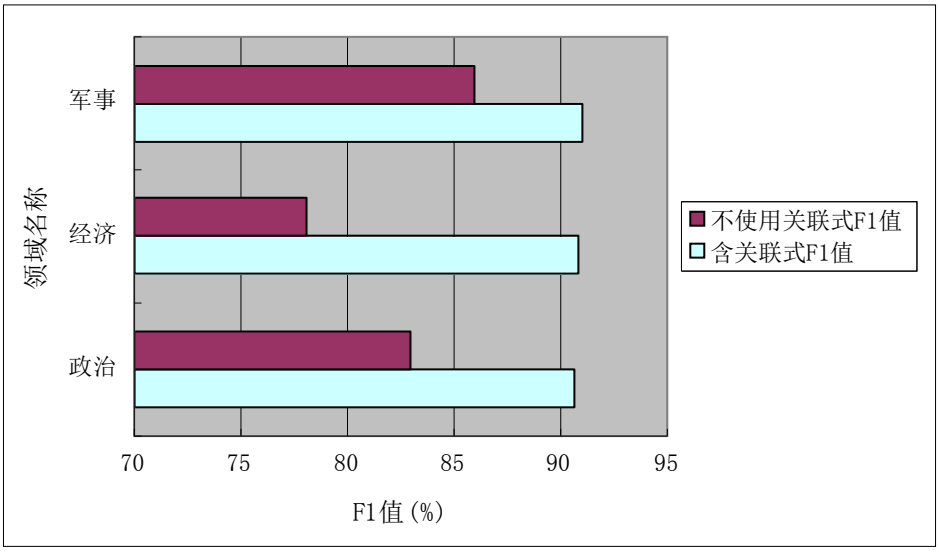

图 1 两种判定方法 F_1 值对比情况

从以上结果可以看出：本文添加概念关联式进行领域判定的方法效果较好，相比于不使用概念关联式的情况，在判定正确率和召回率方面均有不同程度的提升，F_1 值改进明显（采用本文方法，政治、经济、军事三个领域的 F_1 值分别为：90.61%、90.83%、90.99%，不使

用概念关联式的情况：82.91%、78.07%、85.98%，各领域的 F_1 值分别提升了 7.7%、12.76%、5.01%），说明概念关联式的使用在领域判定中能够发挥积极作用。同时，采用本文提出的判定方法进行测试，其 F_1 值均在 90%附近，说明该方法在政治、经济、军事三个领域的稳定性较好。另外，无论使用概念关联式与否，军事领域的判定正确率都很高，说明这一类领域概念对领域的区分度强，其他领域被误判为该领域的可能性低。

在实验中我们发现，绝大多数情况下，不使用概念关联式能够做出正确判定的文本，加入概念关联式后其判定结果仍然是正确的，且作为判定指标的频数信息会往预定领域倾斜；但也存在一些文本，加入概念关联式后，对领域的判定起到了干扰作用，即不使用概念关联式时判定结果是正确的，使用概念关联式反而得到其他领域的结果（该情况，军事领域出现 1 篇，经济领域出现 2 篇）。究其原因，与所选取的概念关联式中不同领域的关联式分布不均衡有关；同时，对于一些文本，其本身就存在多个领域交叉的情况，而本文在结果判定中，只选取频数最大的领域的做法略显粗糙，对于多个领域的频数结果相当，只简单取一项的做法欠缺合理性。我们相信，通过合理选取概念关联式，构造均衡的概念关联式对应表，并制定综合考虑多领域结果的判定方法能够提升领域判定的效果。

4.3 与基于关键词的领域判定的对比

本文最后还设计了一种基于关键词的领域判定方法，用于和前文采用的基于概念基元的判定方法进行对比。基于关键词的领域判定方法操作如下：①、从训练语料中为各领域提取关键词；②、对待判定文本进行词频统计；③、将词频统计结果与各领域关键词对应，选取得分最高的领域作为判定结果。步骤①中的关键词提取采用 χ^2 统计量[8]实现，其计算公式为：

$\chi^2 = \dfrac{N(|ad-bc|-\frac{N}{2})^2}{(a+b)(a+c)(b+d)(c+d)}$，其中，$a$ 表示某词在指定领域中的频数，b 为该领域各文本的总词频，c 表示该词在参照领域（非指定领域）中的频数，d 表示参照领域的总词频，$N=a+b+c+d$；为领域中各个词语计算 χ^2 值，并按 χ^2 值从大到小排序，选取提名靠前的 K 个（K 人为指定，本文取 K=200）词语作为指定领域的关键词；训练语料从语委语料库中选取，各领域分别选取 300 篇文档。在领域判定上，我们同样使用频数统计信息，与基于概念基元的方法一致。步骤③为每个关键词赋予相同权重，以各领域所有关键词的总频数作为最终得分，领域判定基于得分排序，取最高得分作为判定结果。采用 4.1 的测试语料，其判定结果如表 4 所示。

表 4 基于关键词的领域判定结果

机器判定 / 预定领域	政治	经济	军事	其他	召回率(%)	F_1 值
政治	95	9	7	9	79.17	77.23
经济	17	67	1	35	55.83	68.37
军事	14	0	82	24	68.33	78.09
正确率(%)	75.40	88.16	91.11			

从表 4 中可以看出，基于关键词的方法在领域判定中效果不如基于概念基元的方法，其在政治、经济、军事三个领域的 F_1 值分别为 77.23%、68.37%、78.09%，均低于基于概念基元而不使用概念关联式，更低于使用概念关联式的方法，这说明基于概念基元的方法在领域特征提取和凝练上是有优势的。

5 小结

本文基于 HNC 设计的领域概念，利用概念关联式通过频数统计、概念合并以及概念汇总进行领域判定。采用本文提出的方法，对领域判定直接，不需要建立复杂的模型，且无需训练，具有很强的适用性。通过对特定领域的语料进行测试，发现本文方法表现良好，相比

于不使用概念关联式的方法，其性能提升明显。不过，本文也存在需要改进和完善的地方，主要包括：

(1) 在频数统计阶段，由词语到概念基元映射的过程中，本文选取全部概念的方法会对判定结果产生干扰，因此有必要进行概念基元精确标注的工作。

(2) 在关联式选择上，本文选取的概念关联式在各领域中分布不均匀，对领域判定也会造成一定影响。因此，需要制定合理的选取准则，以充分发挥概念关联式的作用。

(3) 本文对不同类型概念关联式的关联权重采取统一赋值处理，没能体现出各关联式的关联强度差异，有必要通过实验为不同的概念关联式赋予权重，为使用概念关联式提供量化标准。

(4) 本文简单地基于汇总结果的频数信息判定领域的方法略显粗糙，丁泽亚[9]曾经利用关联规则挖掘的方法从分类语料中获取与类别关联的概念及概念组合，可以考虑以此为参考，加入领域相关联的概念组合信息，制定新的评分方式；另外，对于领域交叉型文本的判定也需要特别考虑，以制定该类型文本的判定准则。

(5) 本文实验中所涉及的领域较少，在后续工作中有必要扩大领域范围，进一步验证概念关联式在不同领域的判定中能否够发挥作用。另外，本文的工作只是在概念基元层面进行的，仅依靠概念进行处理有时效果不佳，因此有必要将判定工作扩展到句子层面，在句类空间中利用领域句类知识对本文的工作进行补充。

以上均是下一阶段可尝试的工作，希望通过上述改进，领域判定能够取得更好的效果，从而在语境单元萃取方面发挥重要作用。

参考文献

[1] 黄曾阳.HNC（概念层次网络）理论[M].北京：清华大学出版社，1998.
[2] 黄曾阳.语言概念空间的基本定理和数学物理表示式[M].北京：海洋出版社，2004.
[3] 苗传江.HNC（概念层次网络）理论导论[M].北京：清华大学出版社，2005.
[4] 韦向峰.基于HNC理论的扩展句类分析平台研究[D].中国科学院声学研究所博士学位论文，2005.
[5] 缪建明.专业活动领域句类的设计与知识表示[D].中国科学院声学研究所博士学位论文，2007.
[6] HNC自然语言理解处理网站.HNC理论全书[OL].http://www.hncnlp.com/.
[7] 张华平. NLPIR汉语分词系统下载包[OL]. http://ictclas.nlpir.org/newsdownloads?DocId=352.
[8] 杨惠中.语料库语言学导论[M].上海：上海外语教育出版社，2002
[9] 丁泽亚.利用语言概念空间的文本分类研究[D].中国科学院声学研究所博士学位论文，2012.