●贾君枝　陈幼华

汉语框架网络知识本体构建研究*

摘　要　汉语框架网络知识本体是以框架语义学为理论基础,有丰富的语料库支撑,揭示了概念的本体关系。其获取,是在构建语料库的基础上,利用叙词表、分类表和其他知识分类体系等现有的知识本体,识别领域内外相关的概念并抽取相应属性,建立概念之间的关系,并利用所识别的概念和关系创建新的本体,融合已有的本体和新建本体。图 2。参考文献 4。

关键词　知识本体　汉语框架网络　语义 Web　语料库
分类号　G354

ABSTRACT　Ontology in Chinese language framework network is based on framework semantics and supported by rich lexical databases. It can reveal ontological relationship of concepts. Its acquisition is based on the construction of lexical databases, the utilization of existing thesauri, classification schedules and other knowledge classification systems, the identification of related concepts and the extraction of their attributes to establish relationship among concepts. Then, we can use identified concepts and relations to create new ontology, and integrate present ontology and newly created ontology. 2 figs. 4 refs.

KEY WORDS　Ontology. Chinese language framework network. Semantic web. Lexical database.
CLASS NUMBER　G354

1　语义 Web 与本体论

在语义 Web 的体系结构中,本体论具有核心的地位。本体论通过对概念的严格定义和概念之间的关系来确定概念精确含义,表示共同认可的、可共享的知识,成为语义 Web 中语义层次上信息共享和交换的基础。基于本体论的方法是基于知识的、语义上的匹配,在查准率和查全率上有更好的保证,对于面向 Web 信息的知识检索必将起到关键性的作用。

美国加州大学伯克利分院的 FrameNet 项目就是建立在框架语义学基础之上的计算机词典编纂工程[1],它在语料库的支持下,创建一种词汇语料库,构建英语词汇及其所属框架的计算机可读信息,从语义角度描述词与词、概念与概念之间的关系,将语义分析知识构成网络,通过该语义知识网络,计算机可以明确 Web 文本资源的准确含义,帮助用户获得较好的检索结果。FrameNet 项目的本体论思想及其较强的语义分析能力可以供汉语语义知识本体研究者借鉴。

2　汉语框架网络知识本体特点

2.1　以框架语义学为理论基础

框架语义学是一种描写词语意义和语法结构意义的方法。框架,是用来描写语言的意义的概念,是跟一些激活性语境相一致的一个结构化的范畴系统[2]。它通过词语分析,形式化地表示出句法、语义之间的关系。用于激活框架概念的语境,包括一些相关的实体、一些行为模式,或者一些社会制度背景。该理论认为,理解语言中词语的意义,必须先具备概念结构,即语义框架的知识[3]。语义框架是一种图示化的表示方法,涉及各种参与者、外部条件和其他概念角色,称为框架元素。词语的意义是由语义框架显示出来的,进入语句后,按照一定的原则,选择和突出基本语义框架的某些方面。解释词语的意义和功能,可以从基本的语义框架开始。汉语框架网络知识本体的"概念"以描述对象的类型而言,有简单事实及抽象概念,也可描述静态实体因时间推移的相关概念。框架是基本概念的结构化描述,它将一些语义相同的词汇集中在同一框架下表达语义概念,激活相同的语境。框架描述的内容包括框架的命名和定义、框架元素、框架与框架之间的关系及其一系列相关词汇,如图 1"逮捕"框架,明确定义了"逮捕"框架的含义;描述了"逮捕"场景要具有的语义角色,即框架元素;建立了与其他框架之间的联系,该框架与"故意行为"框架具有父子继承关系,与"刑事诉讼程序"框架具有总分关系;该框架下所聚合的一系列相关的词汇。

2.2　有丰富的语料库支撑

框架的定义主要基于语料库的分析,将承担相同语义角色的词汇收集在一起,通过抽取具体的框架元素来进行描述。同时从语料库中抽取包含某个给定意义的词汇的句子,

* 本文系国家社会科学基金项目"汉语框架网络知识本体构建研究"(06CTQ004)成果之一。

作为例示,通过把与框架元素相关的语义标签指派到包含该词汇的句子中的其他短语上,使挑选出来的句子得到语义标注。如例句"他被指控犯有故意杀人罪而被警察逮捕",其中句子中的"逮捕"为目标词,按照其所在框架元素的属性,"他"对应框架元素"犯罪嫌疑人","故意杀人罪"对应元素"指控罪名","警察"对应"官方"。根据对所引用的语料库中的例句进行标注,可以说明这些模式是如何在真实句子中实例化的。同时根据一个词与句子中的各种短语结合的各种模式可形成不同的关于这个词的配价结构,在此基础上可总结出最终的标注总结报告,简明显示每个词汇在组合上的可能性,进行相应的"配价描述",如图2。

逮捕

框架定义

官方指控被怀疑犯罪的犯罪嫌疑人,检察机关作出逮捕决定,警方执行这一决定,并把犯罪嫌疑人羁押起来,它是最严厉的刑事诉讼强制措施。

框架元素　　例句

官方　他被指控犯有故意杀人罪而被*警察*逮捕。

指控罪名　他被指控犯有*故意杀人罪*被警察逮捕。

犯罪嫌疑人　*他*被指控犯有故意杀人罪而被警察逮捕。

犯罪行为　他因为　*入室盗窃*　而被逮捕。

修饰　张秀兰因触犯了刑律而被　*依法*　逮捕,家中抛下17岁的女儿。

手段

地点　警察在他　*家门前*　　将其逮捕。

法律依据　根据　*《中华人民共和国刑事诉讼法》第十六条*的规定,他被逮捕了。

时间　警察在　*星期三下午*　逮捕了他。

框架关系

父框架:故意行为

总框架:刑事诉讼程序

词元集合

逮捕　逮住　拘捕　缉捕　抓　捉　逮　搜捕　追捕　捉住　捉拿　缉拿

图1　"逮捕"框架

"逮捕"词的配价结构

1. 官方 + *逮捕* + 犯罪嫌疑人
2. 官方 + 时间 + 地点 + *逮捕* + 犯罪嫌疑人
3. 官方 + 时间 + *逮捕* + 犯罪嫌疑人
4. *逮捕* + 犯罪嫌疑人 < 官方 >
5. 时间 + 官方 + *逮捕* + 犯罪嫌疑人
6. 官方 + 时间 + 指控罪名 + *逮捕* + 犯罪嫌疑人

图2　"逮捕"词的配价描述

2.3　揭示了概念的本体关系

概念的本体关系表示概念之间的交互作用[4],在一定程度上体现知识的层级关系。汉语框架网络知识本体在其概念关系描述中,通过域关系、继承关系、总分关系、因果关系、参考关系等将概念的个体在空间或时间上的连接性及其概念与概念之间的父子、整体与部分、概念的实例与概念之间进行充分揭示,使所有框架构成一个完整的有层次的系统。

(1)域关系。框架与框架之间表现为具体框架与抽象框架存在着使用上的联系,通常按照领域依赖程度逐层细分,可分为顶级域、领域域、任务域、应用域等。如实体—文本—法律—合法/非法—犯罪—违法行为之间互为域关系,形成多层次域。实体与文本属于顶级域,其中文本域是带有一定语言符号的实体,与实体域之间存在着使用联系,法律域属于领域范畴,合法/非法域及其犯罪域为任务域,违法行为属于应用域。

(2)继承关系。框架与框架之间表现为概括框架与具体框架的父子关系,比如犯罪行为是一个抽象的概括行为,其中"虐待、纵火、绑架、盗版、掠夺、强奸、走私、偷窃"等框架都属于其子框架,实为具体的犯罪行为。父框架的所有特征在子框架中都会体现出来,子框架更为详细地描述父框架,且受限于父框架,父框架的框架元素在子框架中有等同或者详细的对应。

(3)总分关系。框架与框架之间表现为整体与部分或者全面与某一方面的关系,将某一场景发生过程按照分场景进行描述。有些框架是复杂的,这些复杂框架由几个简单框架构成,在它们之间标明了事件状态与转换的顺序,每一个都既可作为单独框架描述,同时又通过分框架关系与复杂框架相联系。各分框架之间存在并列的时间序列关系,分框架之和等于完整的总框架发生场景。如框架"审讯过程"作为复杂框架,在此过程中的每一步,都有独立的框架,即其分框架与之对应,包括"逮捕、审讯、判决、上诉"框架,同一个复杂框架中的分框架都通过顺序与其他相联系,总框架的元素

可映射到分框架的框架元素，但并不是像继承关系存在着一一对应。

(4)参照关系。框架与框架之间概念相似或者难以进行区分、比较、对比时，有问题的框架将与代表性框架建立参见关系。在代表性框架中，有明确的定义进行区分与对比，提醒用户注意和进行类似概念区分与对比。如"自主运动"与"乘坐交通工具"都属于不同的运动方式，有些相似，通过参见方式建立联接，在定义中并注释了它们之间的区别。

3 汉语框架网络知识本体构建思路

汉语框架网络知识本体的获取是基于文本语料库，即在构建语料库的基础上，利用现有的叙词表、分类表及其他知识分类体系等现有的知识本体，识别领域内外相关的概念并抽取相应属性，建立概念之间的关系，并利用所识别的概念及关系创建新的本体，将已有的本体与新建本体进行融合。

3.1 选择已有的顶级本体

顶级本体描述最普遍的概念及概念之间的关系，与具体应用无关，FrameNet项目基于大量语料库分析的基础上，已构建多个顶级本体，如交易、时间、空间、身体、运动、生活、社会、情绪、认知、机会、交流、感知等。在交流本体中，"交流"的属性有：信息传播者、信息接受者、话题、媒质等；它包含的下位类有：讨论、报告、质问、争论、要求等许多框架。汉语框架网络需借鉴FrameNet项目已有的研究成果，参考它已构建的顶级本体，结合汉语特点进行部分调整。

3.2 识别领域内部概念及其属性

汉语框架网络知识本体的构建工作量较大。考虑到知识本体构建是一个开放的体系，采用自下而上的构建方法，通过重用和共享方式，在此基础上不断进行扩充，最终实现系统化构建。首先针对某个具体领域或应用而创建新本体，现以法律领域为研究对象。

(1)建立法律文本语料库。利用Google下载200篇法律文本，按照知识体系将它们进行人工分类，采用分词和词性标注软件、未登录词识别软件、词频统计软件，对文本库进行分词、词性标注，并进行人工校对后，形成网络文本语料库。

(2)识别法律概念及其相关属性。根据FrameNet已有的86个法律框架，以法律专家的参与为核心，结合现有的叙词表及分类表中法律相关类目，对文本进行语义及句法分析，确立以概念相对应的框架及其框架元素及其所包含的词汇，拟构建包含1000个动词的100个框架库。

(3)实现语义标注。以语料库中例句为标注对象，以词汇为目标词，采用手工标注，对例句中框架元素所在的成分标记框架元素名称、短语类型和句法功能。如下所示，以"逮捕"为目标词，将"逮捕"框架中的框架元素与例句中的短语建立一一对应关系。

<susp-np-subj 苏振兴 nh、w 张秀兰 nh><rea-pp-adva 因 p 触犯 v 了 u 刑律 n 而 c><null 被 p><manr-dp-adva 依法 d><tgt 逮捕 v>、w 家中 nl 抛 v 下 v 17m 岁 q 和 c 14m 岁 q 的 u 两个 m 女儿 n。

若以一个词汇20个例句标注计算，拟构建2万多条语义注释例句。在此基础上，总结现实存在的各目标词的配价模式，用来获取语义角色与表达语义角色的各成分的语法功能之间的关联或关系。

3.3 识别概念之间的关系

FrameNet中概念之间的关系体现为框架之间的关系，以域关系、继承关系、总分关系作为研究重点。框架之间关系的识别主要依赖于框架元素信息的判断。继承关系中，两个框架的框架元素存在着一一对应关系。域关系中，则存在着部分对应及其使用关系。总分关系中，框架元素有部分对应，但更多地体现为框架与框架之间的状态转变与时间连续性，按照各关系的不同特点进行区分，以构建有层次的知识网络。

3.4 合并新旧本体

基于自下而上的方法，在确定概念、概念的属性及其概念间的关系的基础上，创建新本体，判断它在已有顶级本体中的位置。如果在顶级本体中有同样的命名，则考虑新旧本体之间的合并，否则作为新本体扩充到顶级本体中，同时建立新旧本体之间的映射关系。这样反复按照前面的三个步骤，不断地对新本体进行扩充，从法律领域扩展到其他领域，并有效地实现新旧本体的融合与共享。

3.5 编码化处理

形式化描述是本体论的重要构成内容。主要包含两部分内容：①构建汉语框架网络知识库，建立词汇库、框架库、例句库。②对框架与框架元素、框架与框架之间相互关系进行明确定义，选择DAML+OIL或者OWL本体语言，以框架为对象，描述其框架元素及框架间关系，并对例句库中例句进行描述，为今后语义Web的应用奠定基础。

3.6 评估

知识本体构建的最终目的在于促进语义Web的发展，知识本体构建质量的评估取决于应用效果。根据现已建设好词汇库、框架库、例句库等知识本体，结合自动语义标注工具，建立网页上的词汇与汉语框架知识本体中元素的对应，形成Web语义网，探索将这些本体知识运用到搜索引擎中，从语义层面提高其检准率，依此检索结果评价语义检索应用效果，并评价领域相关的本体建设，以适时做出调整，适应语义Web发展的需要。

4 结束语

汉语框架网络知识本体的构建工作量比较大。我们以FrameNet顶级本体为借鉴对象，以法律领域本体构建为中心，在此基础上逐步扩充，并以真实文本语料为依据，可提高研究的可信度。在知识库及语义标注过程中，(转第64页)

其中，g_{jk}代表路径总数。

这样，作者关联密度指数值为0。取值越大，该节点在所在网络中的"交流性"越好。

组间关联集中指数是假设组内所有节点具有相同的作者关联紧密度时该组节点的最小整体关联值，将其定义为：

$$C_B = \frac{2\sum_{i=1}^{g}[C_B(n^*) - C_B(n_i)]}{(g-1)\wedge 2(g-2)}$$

其中，C_B(ni)代表节点i的关联紧密度，C_B(n*)是整个节点集中最大作者关联密度值。

4 总结

本文在对已有研究成果进行综合考察与分析的基础上，提出了共现分析在文本知识挖掘中应用的研究思路，并结合实例加以论证。相信随着共现分析研究与应用的深化，该方法在知识挖掘中将发挥其作用，并将推动文本知识分类、概念空间与Ontology、语义检索、知识地图等领域的研究。

参考文献

1 Kostoff R. N.. Database Tomography: Multidisciplinary Research Thrusts from Co-word Analysis. Proceedings: Portland International Conference on Management of Engineering and Technology, 1991. 2005-08-10

2 王曰芬，宋爽，苗露. 共现分析在知识服务中的应用研究. 现代图书情报技术，2006(4)

3,10 He Qin. Knowledge Discovery Through Co-Word Analysis. [2006-03-29]. http://www. findarticles. com/p/articles/ mi_m1387/ is_1_48/ai_57046530

4 Coulter N., Monarch I., Konda S., Carr, M.. An Evolutionary Perspective of Software Engineering Research Through Co-Word Analysis. [2006-04-20]. http://www. sei. cmu. edu/publications/documents/95. reports/95. tr. 019. html

5 Ding Y., Chowdhury G. G., Foo S.. Bibliometric Cartography of Information Retrieval Research by Using Co-word Analysis. Information Processing and Management, 2001, vol. 37 :817 ~842

6 Morris T. A.. Structural Relationships within Medical Informatics: a Classification/indexing Co-occurrence Analysis, Drexel University(PhD), 2001

7 Wormell I.. Bibliometric analysis of the Welfare Topic. Scientometrics, 2000, 48 (2)

8 Polanco X.. Clustering and Mapping Web Sites For Displaying Implicit Associations and Visualizing Networks. [2006-04-20]. http://www. math. upatras. gr/ ~ mboudour/ articles/web_clustering&mapping. pdf

9 崔雷. 关于从MEDLINE数据库中进行知识抽取和挖掘的研究进展. 情报学报，2003 ，22(4)

11 He Qin. Component study of co-word analysis. University of Illinois at Urbana-Champaign(PhD), 2001

12 Widhalm C., Gigler U., Kopcsa A., Schiebel E.. Co-occurrence and Knowledge Mapping to Indentify Hot Topics and Key Players in the Field of Mobility and Transport. [2006-04-20]. http://www. semantic-web. at/file_upload/root_tmpphpA4EeZH. pdf

13 H·rlesberger M., Dachs B.. Behaviour of the Trans-border Co-operation within the European Framework-Programme. [2006-03-29]. http://www. eicstes. org/EICSTES_PDF/PAPERS/Behaviour. pdf

王曰芬 南京理工大学经济管理学院副教授。通信地址：江苏南京。邮编210094。

宋爽 卢宁 朱烨 南京理工大学经济管理学院硕士研究生。通信地址同上。

(来稿时间：2006-07-06)

(接第58页)利用研制的词汇编辑器、框架元素编辑器、框架关系编辑器和计算机辅助语义标注器等工具，一定程度上可减少工作量，提高工作效率，最终实现基本概念的结构化描述，以机器可读形式对语义知识编码，为信息科学领域本体论的深入研究提供了参考作用，同时为今后自动语义标注、问答系统、信息抽取、信息检索的应用奠定了基础。

参考文献

1 Ruppenhofer Josef, Michael Ellsworth. FrameNet: Theory and Practice. [2005-11-18]. http://framenet. icsi. berkeley. edu

2 Fillmore, Charles J. Frame semantics. In Linguistics in the Morning Calm, Seoul, Hanshin Publishing Co. 1982: 111 -137

3 Fillmore, Charles J. Frames and the semantics of understanding. In Quadern di Semantica, 1985, Vol. 6. 2

4 冯志伟. 现代术语学引论. 北京：语文出版社，1997

贾君枝 山西大学管理学院副教授，博士。通信地址：太原。邮编030006。

陈幼华 上海大学图书馆科技情报所。通信地址：上海。邮编200436。

(来稿时间：2006-06-20)