

军事训练领域核心本体的构建

蒋 维, 郝文宁, 杨晓恕

(解放军理工大学工程兵工程学院, 南京 210007)

摘 要: 数据是作战指挥训练模拟系统的基础, 权威数据的缺乏、数据检索的困难等, 容易出现数据失控的现象。加强对数据的管理, 建立统一的标准是目前亟待解决的问题。该文通过引入本体有效地解决了上述问题, 本体的建立使得资源数据定义格式有了统一的规范, 在军事训练网中可共享数据。

关键词: 军事训练领域; 本体; 军事本体; 本体构建

Foundation of Ontology in Military Training Field

JIANG Wei, HAO Wen-ning, YANG Xiao-jia

(Engineering Institute of Engineering Corps, PLA Univ. of Sci. & Tech., Nanjing 210007)

【Abstract】 Data are the base of the simulation training system for commander of combat. But a lack of authority data and the difficulties of data search may lead to the phenomenon that the data can not be controlled. It is a key to strengthen the management of the data and establish a standard. This paper uses the ontology to solve the above problem. The foundation of the ontology has a standard for the definition of the resource data. So it is possible that the data can be shared in the Internet.

【Key words】 military training field; ontology; military ontology; ontology foundation

随着新军事革命的开展, 作战指挥训练模拟系统成为高科技练兵的一个重要手段。而作战指挥训练模拟系统离不开数据, 数据是系统的基础。随着作战指挥训练模拟系统研究的深入和作战演习(特别是跨军兵种作战演习)过程中各种数据的急剧膨胀, 对系统的协调运作形成了巨大的压力, 存在权威数据缺乏、数据种类繁多、数据检索困难、数据流向不明、数据缺乏安全性、数据无法共享等问题, 更为严重的情况是数据泛滥或“数据失控”。因此, 加强对数据的管理、建立统一的标准是目前亟待解决的一个问题。早期的研究是通过建立元数据进行管理, 所谓元数据就是管理数据的数据, 但其可扩充性很差。而近年来, 本体的出现使得改善其扩充性成为可能^[1-3]。本文主要研究在军事训练领域引进本体, 即军事训练领域核心本体的建立。文献[4]在有关军事领域建立本体方面作了一些研究, 但是该文中军事训练本体的建立基本都是手工完成的, 本文试图在其基础上引进一些算法, 实现本体建立过程中的部分功能自动完成。

1 军事训练领域本体模型核心概念集的建立

本体论的概念最初起源于哲学领域。它在哲学中的定义为“对世界上客观存在物的系统地描述, 即存在论”, 是客观存在的一个系统的解释或说明, 关心的是客观现实的抽象本质。在人工智能界, 最著名并被引用得最为广泛的本体论的定义是由 Gruber 提出的“本体是概念化的明确的规范说明”, Fensel 对这个定义进行分析后认为 Ontology 的概念包括 4 个主要方面: (1)概念化: 客观世界的现象的抽象模型; (2)明确: 概念及它们之间联系都被精确定义; (3)形式化: 精确的数学描述; (4)共享: 本体中反映的知识是其使用者共同认可的。

1.1 军事领域本体、概念本体和属性本体的定义

在军事训练领域中引进本体建立军事本体, 军事本体又

称为军事知识本体。本文的军事本体有 2 个层次: (1)军事领域本体, 是对军事领域的概念结构进行整体刻画; (2)概念本体和属性本体。下面是军事本体的形式定义:

定义 军事本体是一个六元组

$$O = \{C, P, A, H^c, prop, att\}$$

其中, C 表示军事本体中所有概念的集合; P 表示所有关系的集合; A 表示属性合, $H^c \subseteq C \times C$ 表达了概念之间的层次联系, $H^c(C_1, C_2)$ 说明 C_1 是 C_2 的子概念; $Prop(p) = (C_1, C_2)$ 表示 C_1 和 C_2 概念之间存在 P 联系; 函数 $att: A \rightarrow C$ 将概念与字面值对应起来(如 $range(A) = string$)。

在军事训练领域引进了本体的概念, 能够通过属性集和属性之间的约束的刻画, 更好地描述军事训练领域的数据这一概念。

1.2 本体建立步骤

本体的建立过程如图 1 所示。

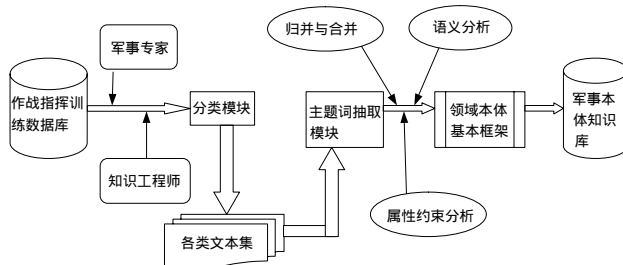


图 1 本体的建立过程

基金项目: 国家自然科学基金资助项目(70371039)

作者简介: 蒋 维(1981 -), 女, 博士研究生, 主研方向: 信息系统安全与管理; 郝文宁, 副教授; 杨晓恕, 博士研究生

收稿日期: 2007-03-30 **E-mail:** lijunling@nju.org.cn

结合以往研究经验,本文提出了建立本体的5个步骤:(1)手工交互阶段;(2)主题词抽取模块;(3)归并与合并;(4)语义分析;(5)属性约束分析。

1.2.1 手工交互阶段

手工交互阶段主要通过军事专家和知识工程师的交互完成。该阶段的目标是进行概念的提取、数据的分析。概念(即类)的确定主要由军事专家和知识工程师参照解放军相关条文(该条文共分29类,收词6562条)分类标准从现有的军事训练数据库中进行。要进行提取概念首先要确定该领域的关键概念,确定上位概念和常用概念。本文分析了作战指挥训练数据库中相关记录,并按军事训练领域对象间逻辑关系进行分类,从而能够将相近的记录归并在一起。军事训练领域数据对象简单分类如图2所示。

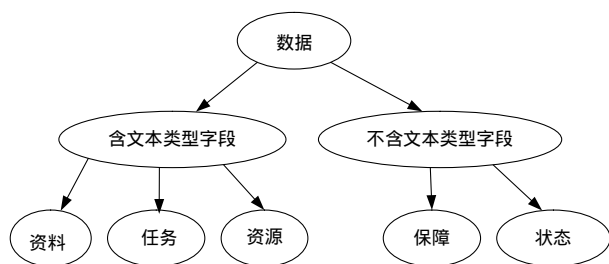


图2 数据对象简单分类

1.2.2 主题词抽取模块

该模块主要是为了提取手工交互阶段已经经过分类后的记录进行主题词的抽取,从而建立每类的主题词集合,本体概念和属性都是从关键词中选取。

该模块分2种情况处理:

(1)数据记录中有一个字段为文本型

读取该字段的内容,然后对其进行分词处理之后可以主题词抽取。关键技术使用了文本分类中的特征抽取和选择,具体见2.1节。

(2)数据记录中没有字段为文本型

对于这种类型的纪录,则是抽取字段名,同时记录它的取值类型和词性。

通过抽取决无重复的主题词,然后可以进行对这些主题词的归并、合并、归类和语义分析。

1.2.3 归并与合并

本体的语义关系有许多种,例如同义关系、上下位关系、包含关系等。上下位关系一般指的是本体与本体之间的关系,通常两个本体之间是父子关系,可以是直接的,也可以是间接的,例如水上坦克本体是装甲车辆本体的子本体,它们之间是继承关系。包含关系指的是一个本体是另外一个本体的属性,例如履带本体被水上坦克包含在内。

同义关系属于本体的语义关系。中文单词的意思含义非常丰富,由于人们看待问题的角度不同,对待同一个概念的词汇表达可能有多,如人们提到“速度”,通常还有“时速”,表达的都是同一个意思,因此正确确定以本体的语义关系,可以扩大本体表达的外延。

该阶段主要是通过SOM网络聚类算法自动寻找主题词集合中的同义词来构造同义关系,具体算法在本文下面章节介绍。

1.2.4 语义分析

该部分主要完成的功能有以下几个方面:(1)分析主体词的词性;(2)了解主题词的注解,确定它的确切含义和所属

类别;(3)构建军事训练领域的领域本体,即军事训练领域的主要框架;(4)构建概念本体和属性集。

主体词的词性在分词的时候已经进行标注了,本文主要考虑两种词性:名词和谓词。最后得到3大类概念集合:名词性概念集合,谓词性概念集合和军事训练概念集合。3个集合的简单举例如下:

名词性概念集合:战斗企图。实例:敌方企图,我方企图。

军事训练概念集合:(1)武器装备;(2)武装力量;(3)战场环境等。

谓词性概念集合:进攻战斗行动。实例:开进,伏击。

军事训练概念集合是本文研究的重点,通过对军事训练概念集合的分析,从主题词集合中共抽取了13个主题词构建了包括作战想定等在内的13个领域本体。然后从余下的主题词集合中选取一部分词作为概念本体,最后的主题词挑选部分作为属性集。

在本题概念集构建完了之后,需要进行属性之间的约束分析。

1.2.5 属性约束分析

本体间的属性往往有一些约束关系,例如:

演习战例本体:继承 战例本体

```
{
    属性: 参战兵力
        : 类型 整数
    属性: 伤亡人数
        : 类型 整数 }
```

存在的约束是:伤亡人数小于参战兵力。属性约束分析只能通过军事专家手工进行。

2 几个关键技术的实现

实现过程的关键技术有:主题词抽取,语义关系处理和规则的生成等。本节主要讨论主题词的抽取和本体间语义关系的处理。

2.1 主题词的抽取

对经过分类后的各类文本集按照解放军相关条文进行分词处理并完成词性的标注。剔出那些出现频率高但是对文本分类不起作用的词,如“的,很,非常,许多”等。

当文本进行分词后,本文使用位置启发式方法计算各个词的权值,再通过设定一定的阈值进行筛选。通常计算各个词的权值都是通过统计的方法实现,如计算每个词在文档中出现的频率。

作为一个本体的属性要求它在训练集文本中出现的概率要满足一定的要求:

$$p_{ij} = p(t_i | C_j) = \frac{t_i \text{在} C_j \text{出现的次数}}{C_j \text{中的词的个数}}$$

其中, t_i 表示第*i*个单词; C_j 表示*C*类中的第*j*篇文本。根据常识知道,一般单词在文本中出现的位置不同,它所含的信息量也不一样。通常单词出现在标题或关键词和摘要中,它的信息量比较大,而在正文中则相对差些。定义了位置基数如下:

$$B_i = C_j \text{ 中单词的个数} / C \text{ 类中的文本数}$$

如果单词出现在标题、摘要、关键词和正文中,分别赋予不同的权重系数,最后得出每个词真正的权重如下:

(下转第212页)