雷达知识图谱

项目方案书

编写人：刘硕、赵创钿、欧远昶

日期：2016年6月13日

目录

[1. 背景 3](#_Toc453621913)

[2. 大体框架 3](#_Toc453621914)

[3. 本体构建过程 4](#_Toc453621915)

[3.1初步：手动 4](#_Toc453621916)

[3.1.1主题分类表 4](#_Toc453621917)

[3.1.2 类属性表 5](#_Toc453621918)

[3.2 自动抽取 5](#_Toc453621919)

[3.2.1 模板 5](#_Toc453621920)

[3.2.2 文本抽取（知识元） 5](#_Toc453621921)

[3.3 人工验证、修正 6](#_Toc453621922)

[4. 参考文献 6](#_Toc453621923)

## 1. 背景

本项目针对国防工业出版社的出版的雷达学科众多出版物，希望从中抽取出实体与概念，构建一个关于雷达系统的知识图谱，将不同出版物中的知识资源整合起来。基于本体的知识抽取技术是当今知识抽取的新方向与研究热点，本项目的目标也是对雷达知识构建出其本体结构。这样信息检索就可以进一步发展为知识检索,信息服务也可以发展成为知识服务。文本知识抽取正是在这样的背景之下,引起了越来越多国内外学者的关注,同时也形成了一个极具挑战性的研究领域。雷达主题的出版物种类不一，有诸如雷达手册之类的工具书，也有介绍雷达原理之类的教材类书籍，其内容重点各异，有的讲述雷达的探测技术，有的讲述雷达试验的原理、内容、方法及试验数据处理。虽然便面上不同书籍描述的是雷达的不同方面的知识，但实际上它们的关系时互补的，从知识图谱的角度来看是零散的知识资源，将这些知识与概念自动抽取出来，并且形成有结构和关联关系的本体，目前并没有一个完善的直接可用的系统。我们的研究对象就是这些非结构化的自由文本,研究目标是从自然语言描述的文本内容中抽取知识。涉及的相关技术和领域包括自然语言处理、文本挖掘、人工智能等。目前,文本知识抽取急切需要解决三个难题或者说需要回答三个问题:

(1)抽取什么，即解决文本中知识的定义问题。这就要明确在这些出版物中哪些内容是我们需要构建的雷达知识系统中有用的概念，明确所要抽取的目标是什么,才能对所抽取的结果进行评价。

(2)怎样抽取，即建立文本知识抽取模型的问题。让计算机从自然语言描述的文本中自动抽取出人们所需的知识,是一项复杂的工程。所以需要建立一个系统性的知识抽取模型并按照一系列步骤，实现从手动、半自动到自动的过程。

(3)为什么抽取，即抽取出来的知识如何表示,如何建立知识库的问题。从自由文本中抽取出来的知识需要采用一种具有逻辑性的知识表示形式存入知识库中，即构建本体，这样才能进行知识推理和知识创新，使知识服务成为可能。

## 2. 大体框架

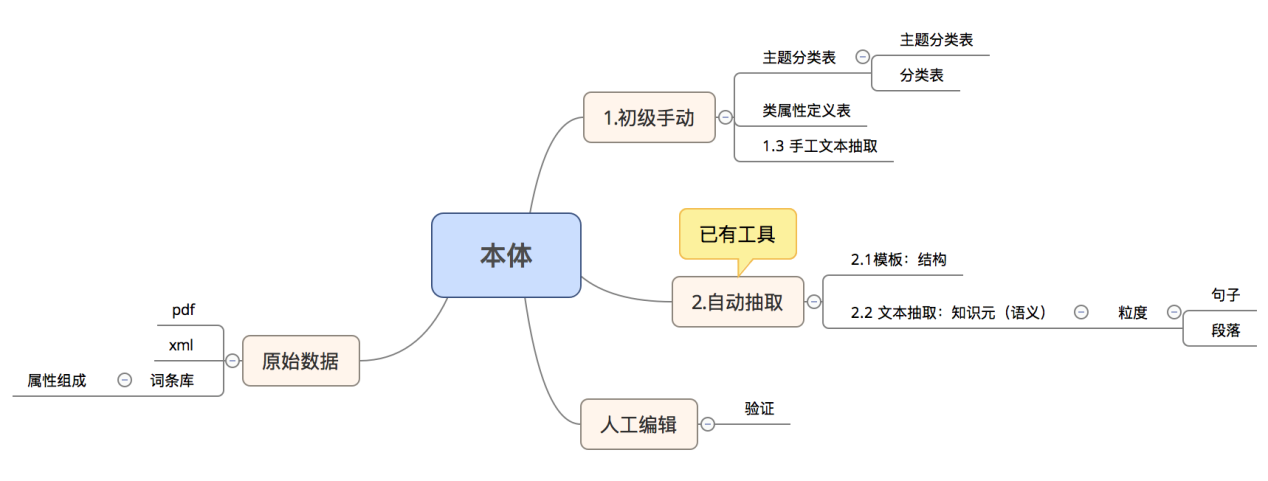


图1 大体框架

## 3. 本体构建过程

### 3.1初步：手动

#### 3.1.1主题分类表

叙词表，将文献作者、标引者和检索者使用的自然语言转换成规范化的叙词型主题检索语言的术语控制工具，亦称主题词表、检索词典。它是一种概括某一学科领域，以规范化的、受控的、动态性的叙词(主题词)为基本成分和以参照系统显示词间关系，用于标引、存储和检索文献的词典。

本项目的主题是雷达及其技术，所以主题表中的词汇都与雷达有关。

根据《汉语主题词表》的结构，我们知道了雷达主题词表的主要结构。

正式叙词的款目内容一般包括：

1. 款目叙词项：在该主题表中，还有对应的英文译名。
2. 标注项：包括款目叙词所属的范畴号、款目词的序号、族首词符号、部分款目词的含义注释等。
3. 参照项：
   * + “代”项（非正式叙词）。符号为“D”；
     + “分”项（下位叙词），符号为“F”，
     + “属”项（上位叙词），符号为“S”，
     + “族”项（族首词），符号为“Z”，
     + “参”项（相关叙词），符号为“C”。
     + 一些叙词没有参照项，即是没有无关联词。

非正式叙词的款目项与正式叙词基本相同，标注项中，一般没有注释项，参照项只有“用”项（应用的正式主题词），符号为“Y”，指出代替它的正式叙词。、

|  |
| --- |
| 雷达回波〔1101〕  款目主题词  Radar echo  英文译名  （代项）非正式主题词  *D*雷达反射  分项下位主题词  *F*月球雷达回波  杂乱回波  属项上位主题词  *S*回波  *C*雷达回波箱  参项相关主题词 |

图2 款目事例

主题词表转换成xml、owl格式。基于主题词表以及四本书的知识将主题词表转换成本体模型。

例子：

#### 3.1.2 类属性表

### 3.2 自动抽取

信息抽取(Information Extraction)，是自然语言处理的一个重要分支，主要是从大量无结构或半结构的文字资料中自动抽取特性信息，以作为数据库存取之用的技术。

严格地说，信息抽取的对象除了文字信息，还可以包含图片、音频、视频等多媒体信息，此处仅取文本领域。

过程：

* 输入：文本信息。书籍的Xml版本
* 设定：特性信息模板。有针对性（如针对会议通知，可以设定模板：时间／地点／主讲人／论题）。
* 输出：格式化的特性信息，方便存储、检索等进一步操作

#### 3.2.1 模板（结构）

#### 3.2.2 文本抽取（知识元）

信息抽取的方法可以简单分为两类：基于规则和基于机器学习。发展至今，基于机器学习的方法是信息抽取技术的主流技术。

* 基于规则

优点：高精度（针对特定领域）

缺点：依赖语言学家的手工劳动（劳动量大，易出错）、可移植性低

* 基于机器学习：

优点：可移植性高，规则全面

缺点：样本量不足（语法中总存在不常见现象）、针对具体问题不易修改、标注样本工作量巨大

命名实体识别的过程通常分两步：

* 识别实体边界
* 确定实体类别

英语中的命名实体具有比较明显的形态标志，如人名、地名等实体中的每个词的第一个字母要大写等，所以实体边界识别相对来说比较容易。

中文内在的特殊性决定了在文本处理时首先必须进行词法分析，中文命名实体识别的难度要比英文的难度大，“分词”成为很重要的一个环节。

命名实体识别常用的算法有四种：

1. 隐马尔可夫模型（Hidden Markov Model, HMM)
2. 最大熵（Maximum Entropy, ME)
3. 条件随机场（Condition Random Fields, CRFs)
4. 支持向量机（Support Vector Machine, SVM）

四种算法优缺点简单比较：

1. 隐马尔可夫模型(HMM)优点是训练和识别时速度快，缺点是理论上的准确率相对于最大熵和条件随机场低。
2. 最大熵模型(ME)，优点是结构紧凑，具有较好的通用性，主要缺点是训练时间复杂性非常高，另外由于需要明确的归一化计算，导致开销比较大。
3. 条件随机场(CRFs)提供了一个特征灵活，全局最优的标注框架，缺点也是收敛速度慢，训练时间长。

### 3.3 人工验证、修正

通过人工查看，查找出不合理的地方并进行修正。

## 4. 参考文献