专家知识抽取与规则化描述

1知识抽取方式

1.1人工抽取

人工抽取是指首先对所需要的知识内容及类型进行定义，即人工编写规则，然后专家根据预定义的规则匹配出相关知识。这种方式首先要先构建大量的实体及关系抽取规则，然后再将规则与文本字符串进行匹配，识别出命名实体及它们之间的关系。早期的知识抽取大都采用这种方式，在小数据集上可以达到很高的准确率和召回率。

1.2半自动抽取

半自动知识抽取主要是先由用户手工标注出训练数据，然后利用机器学习方法从训练数据中学习得到抽取规则，用来从具有相同结构的其他内容中抽取出相应的信息。

采用半自动抽取时可以把文档当作字符串进行训练，如果文档是比较结构化的还可以提取出结构化文档中的树结构去训练。

使用半自动抽取在学习到抽取规则后可以有较好的效果，但是手工标注训练数据代价高昂，不仅要标注一定数量的训练数据，而且要选择适当的训练数据，会比较繁琐。因此这种方式只适合在数据量较小或数据格式统一时使用，在数据量较大且数据格式不统一时较为麻烦。

1.3自动抽取

自动提取利用数据挖掘、机器学习、深度学习等人工智能技术从信息源中自动提取出实体、属性和关系，直接以词向量作为系统的输入，通过神经网络来实现实体的识别。 这种方式构建模型的代价较小，鲁棒性更好，但是训练语料的数量和质量直接影响着这种方法的抽取性能，需要寻找出合适的数据作为训练数据。

2存储规范

规则化输出结果主要是以图和表的形式，以下对这两种形式数据的存储进行说明。

2.1图数据的存储

2.2表数据的存储

2.3存储规范（方法、根据规则化的输出结果如图、表，每种方法推荐一种工具）