专家知识抽取与规则化描述

# 知识抽取方式

从知识抽取过程的自动化程度来看，知识抽取可以分为人工抽取、半自动抽取和自动抽取三种方式。自动化程度的不同反映了知识抽取系统所具有的推理能力的不同。

## 1.1 人工抽取

人工抽取是由知识工程师与领域专家共同工作、交流来提炼知识、对所需要的知识内容及类型进行定义，即人工编写规则，然后知识工程师根据预定义的规则手动进行分析、综合、整理、归纳并匹配出相关知识后将知识以特定形式存入知识库。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，所得到的知识大多可以直接用于解决问题。

## 1.2 半自动抽取

半自动知识抽取指利用一些知识获取工具来辅助知识工程师把知识原材料或专家描述的知识内容翻译成所需的知识形式并存入数据库。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通过对应用实例与实际问题进行建模从而发现一些专家尚未形式化甚至未发现的新知识和新规律，从信息源中自动提取出实体、属性和关系。

这种方式构建出模型的代价较小，鲁棒性更好，但是由于训练语料的数量和质量直接影响着这种方法的抽取性能，知识畸变（言语描述和具体行动间形成的错位现象）也影响着知识获取的精准性，自动抽取方式还未成熟。

# 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

图数据中最重要的信息就是图的节点、边的信息和边之间的关系，可以使用关系型数据库或图数据库来存储这些信息，具体如下。

图数据库是基于图论实现的一种新型NoSQL数据库，其数据库存储结构和数据的查询方式都是以图论为基础的。图论中图的基本元素为节点和边，在图数据库中对应的就是节点和关系。图数据库仅使用一个模式和一组起点，就可以围绕这些初始起点探索相邻数据，收集和汇总来自数百万个节点和关系的信息，并保持搜索范围之外的任何数据不变，因此图数据擅长管理高度连接的数据和复杂的查询。

在使用图数据库存储图数据时，可以使用当前流行的图数据库Neo4j。Neo4j有效地将属性图模型实现到了存储级别，还提供了完整的数据库特性，使其更适合生产场景中数据使用。

## 2.2 表数据的存储

表数据中的各种数据被划分为一个个字段，每条数据对于同一个字段可能会有着不同的值，在存储表数据时可以使用关系型数据库。

关系型数据库是建立在关系模型基础上的数据库，其中的数据被组织成二维表格的形式，因此十分适合用来存储表结构的数据。在关系型数据库中，数据由不同的表组成，每个表包含有不同的字段，还可以对各个字段施加多种限制，其结构如下表1所示。

表1 关系型数据表结构

|  |  |  |
| --- | --- | --- |
| 字段1 | 字段2 | 字段3 |
| A | E | I |
| B | F | J |
| C | G | K |

在使用关系型数据库存储图数据时，可以使用当前流行的MySQL数据库管理系统。相比于其它的关系型数据库管理系统，MySQL具有性能高、成本低、可靠性好等多种优点，很适合作为中小型系统的数据库。