第11章 人工智能数据架构

本章节将对人工智能产品经理所需要了解和掌握的数据架构做系统性的介绍。

本章节的目标：  
1. 了解通用人工智能系统的数据架构  
2. 熟悉人工智能与知识图谱的关系  
3. 了解人工智能的知识模型  
4. 熟悉典型人工智能产品的数据架构

希望能够通过以上四个问题，能够对人工智能数据架构有一个比较清晰的认识。接下来我们首先来看人工智能数据架构要解决的核心问题是什么。

# **11.1 人工智能数据架构的核心问题**

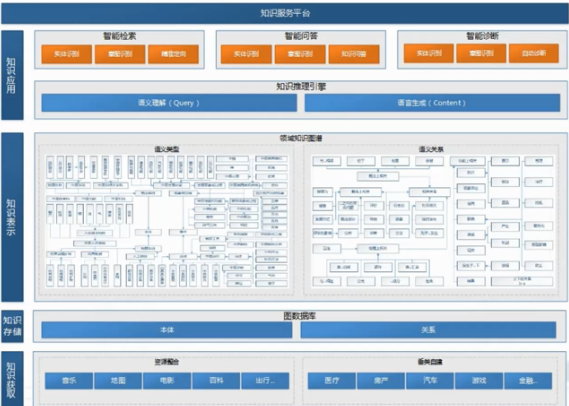
我们要重点解决人工智能数据架构相关的两个问题：

1. 架构框架  
即：通用的人工智能系统的数据架构是如何的

2. 典型框架  
即：目前市场上关于人工智能典型案例的数据架构是怎样的  
我们将在接下来的11.2和11.3节中详细向读者阐述。

# **11.2 通用人工智能系统的数据架构**

通用人工智能系统数据架构框架的四大层次（自底层向上）：知识获取🡪知识存储🡪知识表示🡪知识应用



知识获取：资源整合、垂类自建，其主要考虑人工智能类产品当它所用到的知识图谱知识信息的来源从哪来，主要有两种，资源整合是借助互联网或者其他的开源渠道，对当前已经拥有的资源采用一定的构建或者其他方式，把相应的数据资源整合到一块，具体整合的内容涉及各领域包括音乐、地图、电影、百科等，不管数据在哪只要想要就接进来整合到一起。

垂类自建，是比较累的、投入产出比非常高的一种办法，但是一旦建成可以持续使用，什么时候适用呢？行业领域内没有可借鉴的、可复用的又成熟的数据资源时候，好处在于自己建的拥有绝对的控制权，后续可以交换。

知识存储：图数据库(本体+关系)，这种方式比较主流，它基于文件，好处是可以基于分布式并行计算包括分布式存储框架来进行大量数据高并发的处理，本体对应的是客观存在的事物，关系表示的是事物与事物之间的关系。

知识表示：领域知识图谱(语义类型+语义关系)，知识表示层也即模型层：对知识的建模，有很多种模型的表示方法，目前主流的包括：本体建模，语义网络。往往有时同步使用，即：基于本体的语义网络模型，涵盖的主要内容包括：语义类型，语义关系。

其中：

语义类型：自然领域客观存在的各种事物，它的一个分类关系，用一个专业术语叫：概念(抽象的事物，如汽车)的分类

语义关系：表征概念与概念之间关系的，例如：汽车发动机和轮胎

知识应用：智能检索、智能问答、智能诊断和知识推理引擎。

知识推理引擎：语义理解，语言生成

知识服务平台：智能检索，智能问答，智能诊断

其中：

智能检索：实体识别，当你去搜索一个关键词或者一句话时候，那么通过语义理解可以帮你去理解这句话里面所涵盖的实体对象，例如：宝马X6怎么样，通过语义理解，如果有汽车这个领域的知识图谱，那么我就可以看到，宝马是个品牌，X6是个产品词，这个时候再去给这个人推荐信息就可以不是传统搜索引擎的那种基于关键词的相关，而是有一个搜索结果然后可以定向，定向仅推宝马X6它的详细产品信息，售卖信息等。

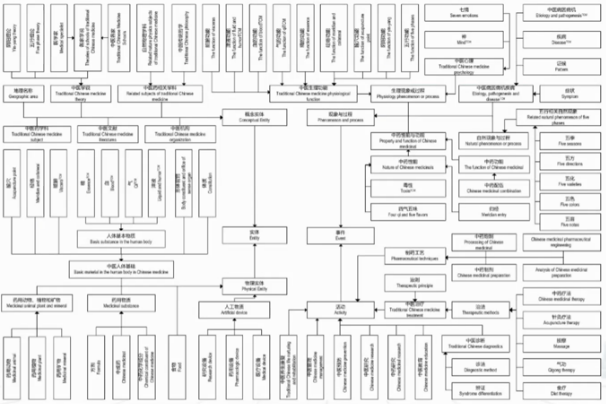
智能问答：我们知道人工智能时代是以语音交互为典型代表的，这个时候人与机器之间就不会再有屏幕触屏鼠标键盘等，这种人机交互是人发起问题，然后智能设备响应你的问题然后给你反馈和响应，在这个时候，基于一句话，翻译成自然语言处理之后的结果，就涉及到实体的识别，意图的识别，另外，针对你想问的内容，例如：感冒了我需要吃什么药，后来需要针对感冒了这样一个病症去寻找对应治疗感冒的药品。

智能诊断：目前来说比较典型的应用场景，主要集中在医疗和医药这个领域，因为智能问答可以在任何一个方向，而智能诊断相对来说就是比较局限的一个领域，主要在医疗这方面，包括通过实体和意图的识别来自动诊断疾病。

综上：这四个层次都是围绕一个主题：知识图谱的建设来展开的。

# **11.3 人工智能系统的典型案例**

这里我们给出的典型案例是TCMLS语义类型（126）



这张图表示的是一个中医医学语言系统（TCMLS），我们分析它的语义网络模型，主要分为实体和事件。

首先，实体的划分：概念实体、物理实体；

其次，事件的划分：现象与过程、活动。

这个分类方法是有典型代表意义的，在这个里面需要注意的一个问题是：实体与事件，动态与静态的内容其实都属于本体。

实体里面概念和物理实体的划分，重在意会，物理实体可以理解为一个客观的存在，概念实体可以理解为一个抽象的存在，当然它也是客观的，只不过它是一个抽象的有具体的物理意义的组成。

事件的构成，现象与过程和活动。人的某一个特定的行为产生的某一个结果我们称之为活动，然后现象不是一种活动而是呈现出来的一种结果，例如感冒了是一种现象也是一个事件，在这里考虑的第二个问题是针对事件而言，无论是一个具体的活动还是一个现象，它都可以归类为事件里面，在TCMLS语义类型里面一共有126种，只不过是在认证的时候它这个版本不一样，然后TCMLS的语义关系，一共是58种，其实是两大类，第一类是相关关系，第二类是上下文关系。

其中，相关关系分为5大类：

空间上的相关，例如：与……相邻，位于……之上，包围，穿过等

概念上相关，……是……的评价，特性，诊断，分析，测量，方法等

时间上相关，与……同时发生，先于……发生等，这里的……是指概念，也就是具体的实体对象

物理上相关，由……组成，部分，由……汇合等，例如：长江是由哪些分支流汇合而成的

功能上相关，例如：谁使用什么等

只需要了解语义关系中相关关系可以分成5大类，每一类包括……相关，具体里面的每一个关系这种类型我们可以在自己的应用场景里面自己来设计，这种可以作为一种参考模型。

# **11.4 想一想**

制约人工智能发展的关键因素都有哪些？

为什么提到这个问题？目前智能音箱很傻，那么之所以有这样的结果是因为当我们在和他进行交互和交流的时候，首先他不能准确的理解我们，然后他不能给到我们想要的答案，所以我们认为他很傻，很白痴。

那么问题的实质是什么？

人和音箱进行交互只是一个手段的差异，然后之所以理解不了你和给不了你想要的服务能力的产出，在于他背后的知识体系的支撑还没有达到一个工业级应用的水平，这个里面的关键点就是知识图谱的建设，也就是从某一个层面来讲，他是完全受限于当前知识图谱的工程建设能力的问题，当然还有其他很多原因，我们可以就这个问题展开思索，同时去做一些相应的总结。

# **11.5 参考资料**