主题词表自动化构建技术

中国科学技术信息研究所

中国科学技术信息研究所
中国科学技术信息研究所

目　录

一、主题词表自动化构建的技术流程框架.....4
二、素材资源采集中的关键技术.....6
三、术语素材深度清洗的关键技术.....13
四、基础词库构建的关键技术.....16
五、术语词库构建的关键技术.....22
六、语义关系建立的关键技术.....26
七、主题词表编制平台中的关键技术.....28
八、语义关系逻辑错误的智能校验与修正.....31

中国科学技术信息研究所
中国科学技术信息研究所

插 图 清 单

图 1 技术流程框架图.....4
图 2 并行分布式采集框架.....7
图 3 多来源素材评价.....9
图 4 术语样本空间.....12
图 5 素材清洗流程.....13
图 6 基于条件随机域的素材清洗技术.....16
图 7 基础词库构建流程.....17
图 8 多层术语语义聚类.....19
图 9 多维度术语质量计算.....21
图 10 术语时序特征示例.....24
图 11 领域术语特征示例.....24
图 12 语义关系模式挖掘示例.....26
图 13 语义关系挖掘流程.....28
图 14 大数据主-从架构下的多任务编制.....29

中国科学技术信息研究所
中国科学技术信息研究所

一、主题词表自动化构建的技术流程框架

传统主题词表编制中人工参与为主，自动化程度较低，造成编制成本高、周期长，质量也无法得到保证。因此尽可能提高自动化程度是当前及未来的主题词表编制趋势，降低人工参与的代价，让编制人员从机器推荐结果中，通过判断和选择为主的方式进行编制，这也符合大数据时代下对大型知识组织系统的构建理念。

主题词表编制在技术层面面临的主要问题是：编制主题词表的素材规模大、多源异构、缺少统一规范等。拟采用当前人工智能较为成熟先进的技术方法，解决编制素材中噪音高、分布不均衡及缺少标注样例等问题，以无监督学习和递进迭代学习方面的技术方法为主，并结合深度学习技术挖掘分析术语特征及语义关系特征，在编制过程中把专家的编制结果作为训练样本，达到专家人工编制与机器自动推荐相互促进相互增益的目的。

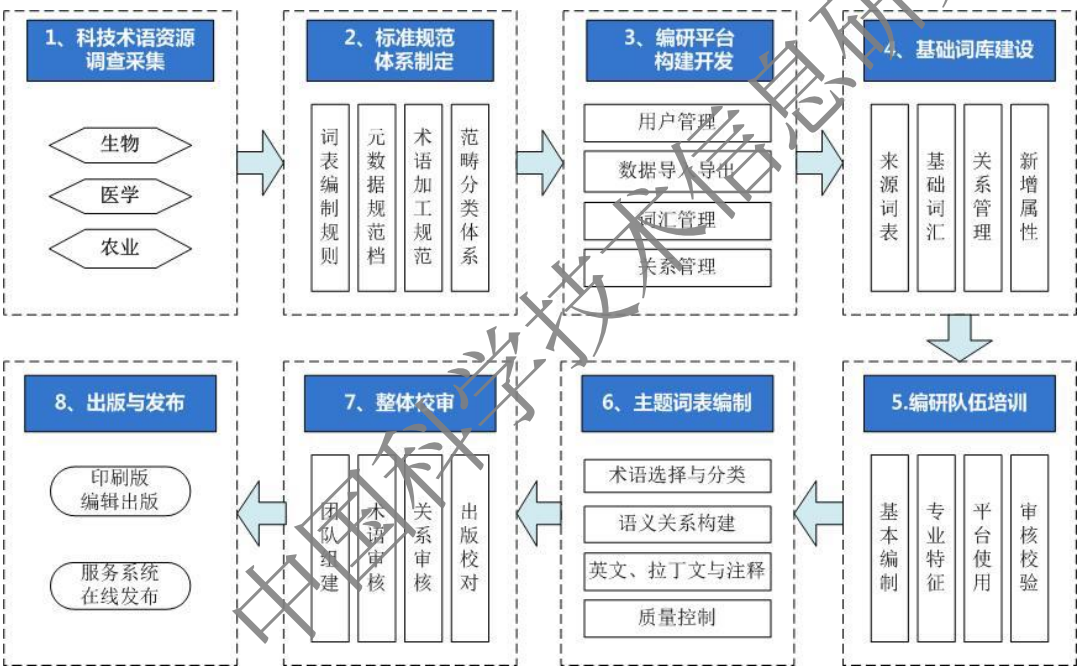

图 1 **技术流程框架图**

- 术语素材来源：来源的广泛性和可靠性决定了术语素材的覆盖面

中国科学技术信息研究所

和质量，获取术语的信息源主要包括词表（主题词表、词典、术语表、词库等）、文献（标题、摘要、关键词、正文）、网络文本（学术技术类的网文、百科、博客、微信公众号等）。

- 多来源术语采集：首先通过采样评价不同信息源的术语质量，识别出低质量的术语并进行清洗去噪，对进入基础词库的素材术语在表达形式上做统一的规范化。
- 标准规范：针对生命科学领域知识分类特征、知识描述特征、术语分布特征和知识组织系统集成应用模式，制订主题词表编制标准和规范，用来指导整个流程中数据管理和操作执行。
- 元数据体系：针对不同学科领域和不同信息源导致的元数据不一致的问题，依据相应的编制标准规范，建立完整的元数据体系，把不同信息源的术语和语义关系都统一在这个元数据体系下。元数据主要包括术语的唯一标识命名规则、属性信息类型和语义关系类型等。
- 范畴分类体系；与元数据体系类似，不同信息源的术语素材通常使用不同的范畴分类，比如中图分类法、杜威十进分类法(DDC)等，首先要建立主题词表自己的范畴分类体系，并进一步将不同信息源的范畴分类体系向其映射，既为术语赋予了统一的范畴分类，又可以依据范畴分类对课题二到课题五划分编制任务。
- 主题词表编制平台：半自动化的编制工具，不同学科领域的编制专家可以以协同方式在自己领域内根据推荐结果，选取术语并构建及修正术语间的语义关系。编制平台主要特点是：自动推荐、

中国科学技术信息研究所