基于互联网的发动机研发知识数据挖掘与应用

Acquisition and Application of Engine R&D Knowledge Data Based on Internet

李昌红 何为 孙明霞 / 中国航发动力所

航空发动机研发需要有目的、有计划、有组织地挖掘互联网的新技术、新动态和新成果，经处理、分析后提供定制化的具有先进性、预测性和前瞻性的知识数据，为发动机研制趋势对比、建模、新产品开发，以及关键技术突破提供数据支撑。

站在巨人的肩膀上创新研发是这个时代的主题，任何组织封闭自己、闭门造车终将被淘汰。向发动机研发人员提供面向全球的知识化环境，离不开先进的外部知识数据和精准情报。为实现发动机研发所需知识数据的采集、存储、分析与利用，与时俱进地满足多样化、高效化、个性化、专深化的用户要求，更好地为管理决策、科研攻关、科研保障建设提供决策支撑和技术引领，开展基于互联网数据的航空发动机研发知识数据挖掘与应用至关重要。

知识数据在发动机研发中的作用

发动机研制过程是一项复杂的系统工程，涉及学科、领域众多，对先进知识和技术需求强烈，我国发动机研制水平与世界先进水平还存在差距，需要学习、借鉴行业内的前沿技术。通过互联网获取的知识数据，经提炼总结形成研发体系指导性文件，可以实现外部知识内部化，内部知识体系化，支撑研发流程活动高效运行，为研发人员提供工作指导，为研发过程中疑难问题提供解决途径，为模型、仿真优化提供数据支撑，提升组织的研发能力，如图1所示。面向部门业务需求，通过立足互联网数据相关技术，构建基于各业务主题的互联网数据挖掘、处理、应用等方法及工具，可以实现知识伴随流程的存储、积累和应用，有效支撑发动机研制。

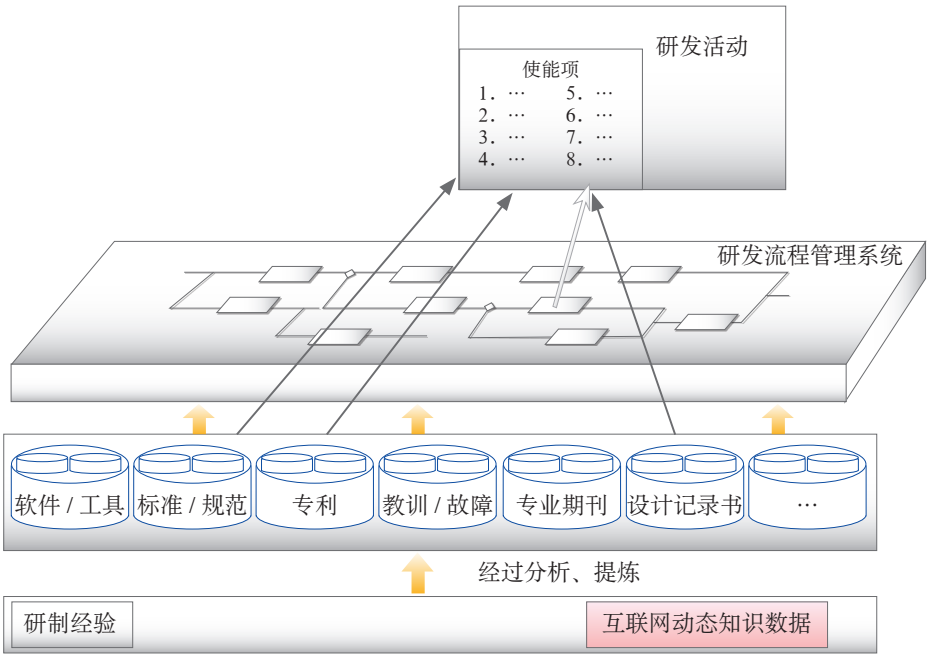

图1 互联网知识数据在发动机研制中的作用

发动机研发知识数据的挖掘

依据发动机研制技术树，分析相关技术对互联网数据的需求，提出数据获取、挖掘的专题清单并进行专题跟踪、挖掘。按主题分类管理，以树形结构管理主题类型，可以同时创建多个采集主题类型，实现按主题分类的建立、修改、删除及浏览等功能。对每项主题可以实现跟踪条件配置、跟踪启动、主题重命名及删除等管理功能。利用搜索引擎对一定范围的网站内容进行定期

的自动采集和挖掘。针对某一具体站点内容更新的规律，可以设置网页内容自动跟踪挖掘的周期。

互联网数据挖掘、处理过程中，为更好地满足发动机研发中的使用需求，须实现多种功能：能够挖掘、分析多种常用国家语言，支持多种编码识别与转换，支持各种常用文档格式的识别与下载，能自动过滤同一文档的不同格式；对于有表格、图片、视频以及音频等非文本信息的网页，能够连同网页中的其他文本信息一起被识别与下载，下载后应保持原有的文档结构与顺序不变；能够实现附件分类挖掘，包括文档、图片、多媒体、2D/3D以及压缩包等；实现网页去噪，将垃圾及无关信息过滤。

对指定网站和栏目进行定点、定期的自动挖掘，实现站点的配置管理。以树形结构管理新闻站点分类，可实现站点新建、重命名、删除等操作；实现挖掘规则设置，对具体的某个站点进行管理，设置自动下载的匹配规则，包括常规设置、采集页面规则、翻页规则等。

发动机研发知识数据的处理

挖掘后的互联网数据，经过处理、加工，变成发动机研制可用、好用的知识数据，提供给设计人员，进行定制的个性化匹配，具体可以包括以下几个方面。

第一，挖掘到的互联网知识数据，根据匹配的专题，按照特定的维度在知识管理平台进行存储、管理，对获取的互联网知识数据进行多维度的分类，如按设计过程、数据类型和研制阶段等，如图2所示。根据分类情况，实现互联网数据挖据结果的自动归类、属性定义等。属性信息还须包括标题、作者、来源、时间和大小等。

第二，去重加工。设置一个文章信息相似度，根据对文档标题和内容的分析判断，能够自动去除高相似度的文章。

第三，精准判定。采用基于学习训练的方法或基于规则的方法将采集到的数据信息自动归到已有的分类体系中，也可进行手工归类调整。一篇文档可被归到多个类目下。

第四，自动关联。根据对文档内容的理解，自动将内容相似的文档建立链接，当用户查看每篇文档时，在原文下方会显示与本文档内容相似的其他文档信息；当用户用鼠标选择某句或某段文字时，系统自动呈现与之相关的文档。

第五，聚类挖掘。可以对当前或某一时段内的相关信息按内容相似性进行聚类，自动生成类别的标题、主题或关键词，以适当方式展示聚类结果，聚类的范围与主题可以自行定义。

第六，自动摘要。通过对文档全文的分析，自动提炼出关键词与摘要。关键词与摘要能反映文档的主要意思，组成摘要的句子应规范可读。

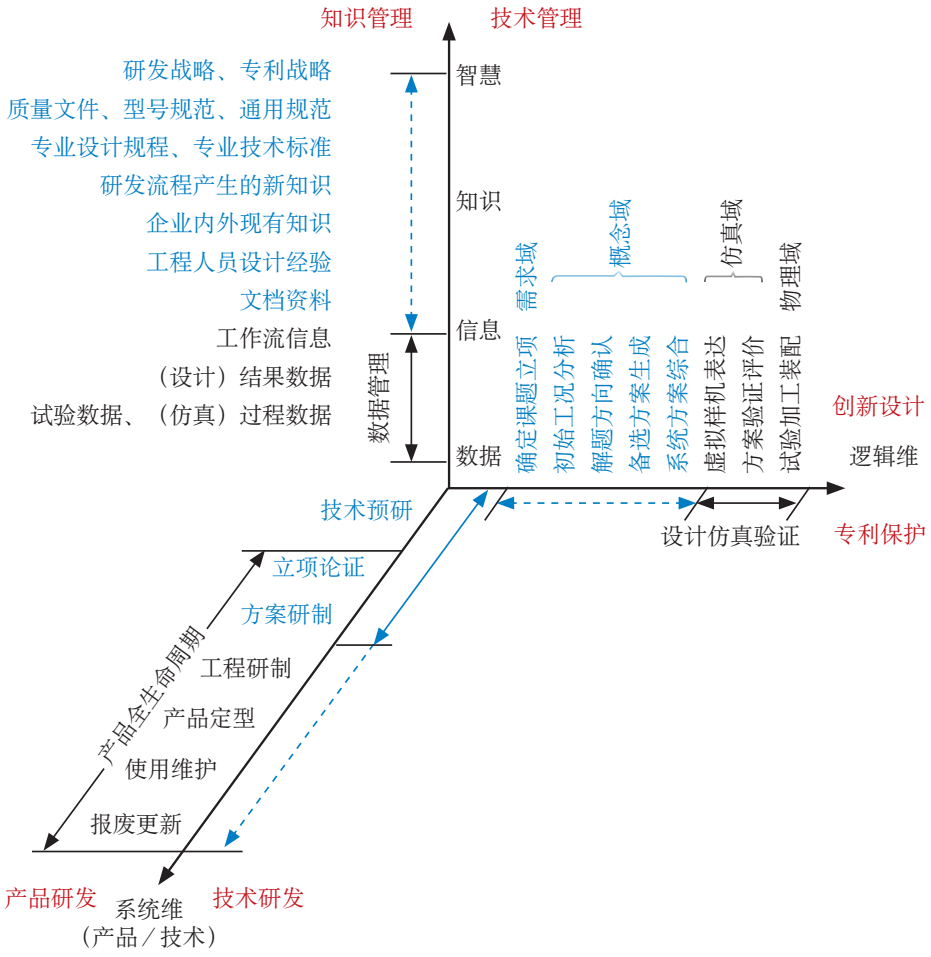

图2　数据属性分类

第七，文档中结构化信息的自动抽取。自动识别并抽取文档中的图表和结构化数据，如设计数据、交付量、订单、价格、企业领导姓名、相关时间、地点等，并存储到相关的数据库中，并且建立数据与原文档的关联。

发动机研发知识数据的分析

通过建立知识矩阵和魔方，实现多维度的分析，包括相似关系分析、时空关系分析和组合关系分析。开展专利趋势分析、分布分析、机构分析和人物分析等。

为直观有效地把控全局，需要实现可视化分析，并具备多种视角以及动态可视化展现功能，方便直观查看数据结果。数据分析应能通过数据线条、大小、颜色反映数据变化趋势并实现点、曲线图、柱状图、饼图、云图等对比分析，可以在曲线上进行标识/标记，绘制两个数据之间的关系，如图3所示。

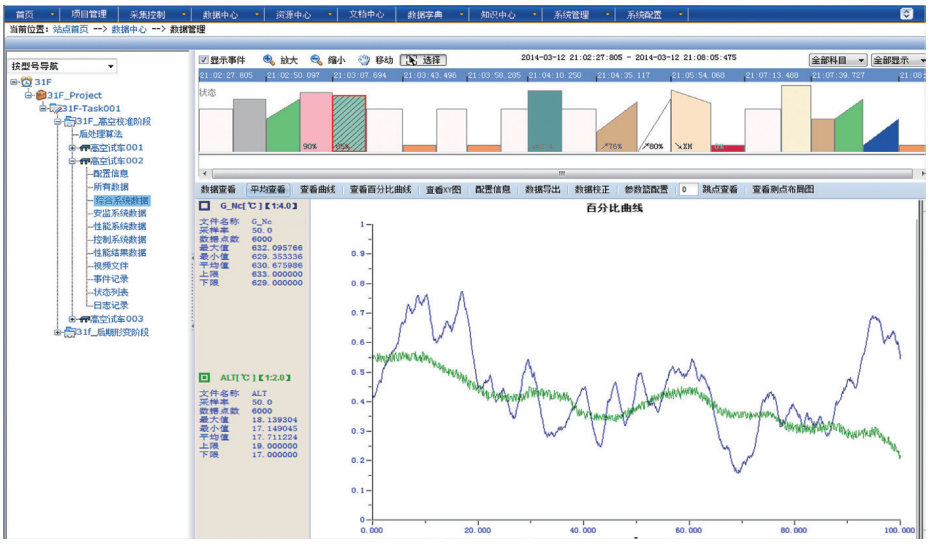

图3　趋势对比分析

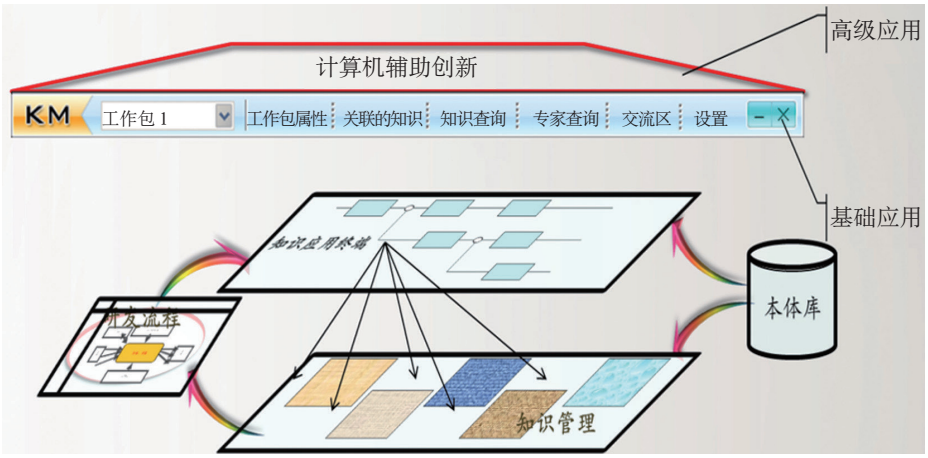

图4　互联网知识数据在发动机研发中的应用

发动机研发知识数据的应用与推送

经挖掘、处理后的互联网数据，结合信息化手段开展应用：实现数据信息的检索，包括一键式检索、分类浏览检索、关键词检索、字段检索、句子检索、全文检索等；实现搜索导航功能，根据关键词内容，实时自动生成相关搜索建议，并以树状结构来展现，帮助找到更相关的搜索结果。在一次检索结果基础上，可进一步用关键词检索、字段检索等方式进行限制检索。检索中的每一条信息均显示标题、日期、来源、相关度以及自动摘要，可选择按日期、相关度或文档类型等方式排列检索结果。在检索结果列表或结果页面中，对标题、摘要以及正文中出现的检索词进行突出显示。

互联网挖掘的知识数据经提炼、总结形成体系指导性文件，根据流程活动、职位通道、关键词等进行知识数据的自动推送，通过与专业设计系统接口的连接，为相关专业系统技术活动提供有效支撑，用于指导发动机的研发、制造、运行维护等全过程，如图4所示。

结束语

在互联网时代，航空发动机研制需利用互联网技术，获取和挖掘前沿、动态的互联网信息，匹配定制的专题云，将获取的信息整理、分类、定制化推送给研发人员，不断沉淀和积累出大量优质知识，实现外部知识内部化，为发动机研制提供好用、可用的知识数据。 航空动力

（李昌红，中国航发动力所，高级工程师，主要从事研发体系相关工作）