借助深度学习框架 支撑科技高质量发展

习近平总书记在党的十九大报告中指出，“我国经济已由高速增长阶段转向高质量发展阶段”，这是新时代我国经济发展的一个重大判断。创新是引领发展的第一动力，是建设现代化经济体系的战略支撑。充分发挥科技创新在高质量发展中的支撑引领作用，保持经济持续健康发展，是当前和今后一个时期科技工作的重要任务。

我市作为国家自主创新示范区、国家创新型试点城市、首批知识产权区域布局试点城市，在科技创新云服务、知识产权区域布局和知识产权公共运营服务方面已形成一套较为完善的工作机制和运行平台。但伴随科技政策、知识产权等数据的骤增，目前已有的工作机制和平台无法快速地、有效地挖掘出反映政策体系结构、科技发展方向、平台用户潜在需求等信息。

为了有效地支撑科技高质量发展，需要全面加强深度学习框架基础研究和应用研究，促进深度学习基础框架和应用场景相融合。

1. **研究方法及关键流程**

**（一）研究方法**

深度学习是神经网络及其应用的统称，旨在将复杂的神经网络架构应用在数据建模上，从而带来前所未有的准确性，目前已在计算机视觉、语音识别、文本挖掘等众多领域取得突破性进展。深度学习框架能够提供进行深度学习的底层架构、接口，以及大量神经网络模型，减少重复编程的时间和精力，提高深度学习效率。

**（二）关键流程**

本文主要采用的深度学习框架是Doc2Vec，旨在将复杂的文本转换成数字，从而挖掘出语义信息。关键流程包括收集文本数据、构建语料库、中文分词、特征处理和聚类。

**收集文本数据**。文本数据主要来源于两大类，一类是业务上传数据，包括PDF文件、Word文件、图片、Excel文件和特定URL；另一类是由网络爬虫得到。

**构建语料库**。由于直接使用文本内容作为语料库，可能会引起两个类别的文本分到同一类，即存在误分类情况。为了减少这种情况，本文根据换行符等特殊标记将整个文本内容进行分段，通过段落信息**构建语料库**。

**中文分词**。采用Jieba技术对段落信息进行中文分词。

**停用词维护**。中文分词之后的段落存在很多无效的词，比如“在”，“等”，还有一些标点符号，不想在文本分析的时候引入，因此需要去掉，这些词就是**停用词**。

**特征处理**。利用Doc2Vec提取段落向量，Doc2Vec利用神经网络计算文本的特征向量，能更深入地体现出段落文本的语义特征。

**聚类**。对所有段落向量进行Kmeans聚类，通过段落向量计算每个段落的相似度得到每个类别的文本信息，研究人员会对聚类结果进行人工筛选，对于筛选过的语料库进行重新聚类，直到获得研究人员满意的聚类结果，其具体过程如图1所示所示。



图1 基于Doc2Vec的深度学习框架

1. **具体应用**

深度学习框架+应用场景，已成为当今的主流发展模式，成为人工智能发展的基础和核心，也是支撑科技高质量发展的关键一环。目前，深度学习框架Doc2Vec可以应用于政策血缘网络、企业关键技术演进路线、语义检索和智能客服这四个应用场景中。

**（一）构建政策血缘网络，支撑政策制定科学化**

有研究发现，目前我国政策碎片化现象比较严重，主要表现为政策间相互矛盾、政策重叠，政策体系不够完整且缺乏配套、政策缺乏稳定性等现象。借助深度学习框架Doc2Vec对政策文本进行语义分析，实现政策文本聚类，进而对政策网络进行层级划分，实现同功能政策的替代或分解，减少平行政策数量，从而延缓政策碎片化；对于新政策，通过语义检索在语料库中寻找与新政策相似的已存在的政策，进而为政策制定者提供政策定位的辅助信息，从而避免产生新的政策碎片。

构建政策血缘网络，能够形成清晰的政策体系结构。政策血缘关系指的是目前所有有效或已失效的政策都因与宪法之间的祖孙关系而存在内在的联系，例如，在我国的社会保障体系中，政策之间的政策血缘关系如下图所示。

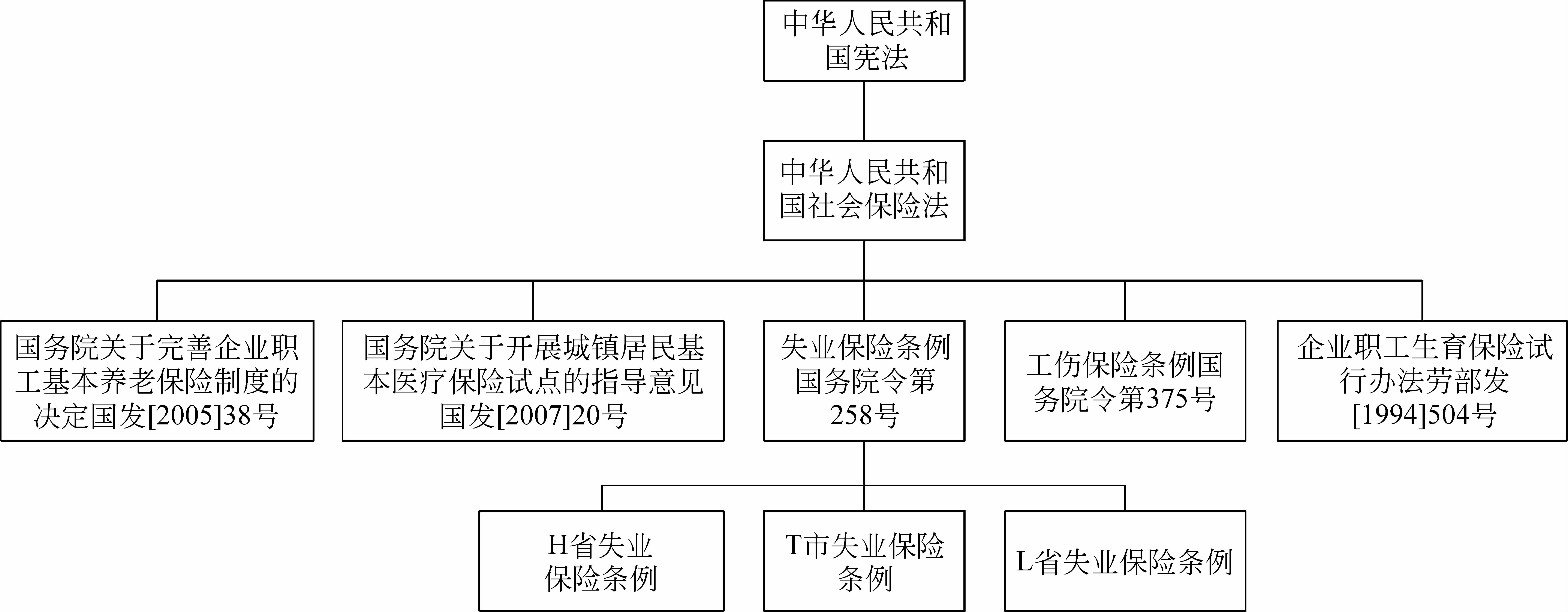


图2 我国社会保障政策族谱树

上图中展示的就是一棵政策族谱树，它描述了图中节点的父子关系。由图可知，树中路径距离较远的节点之间的血缘关系应当较为薄弱，树中路径距离较近的节点之间的血缘关系应当较为紧密。

同时，通过Doc2Vec对政策文本进行语义分析，挖掘出政策间的隐形血缘关系，从而实现隐形血缘政策网络体系。以中国社会保险体系为例，该体系将社会保险分为八大类，其组织形式如下图所示。对上述不同类别的政策子系统，其内部又可根据不同的侧重分出不同的子类，每个子类包含若干相关政策。有些政策的组织形式彼此隔离，互不相关。但是从内容上讲，它们的实施办法相似或相关，这种联系却不能从政策族谱树中体现，这就是政策间的隐形血缘关系。

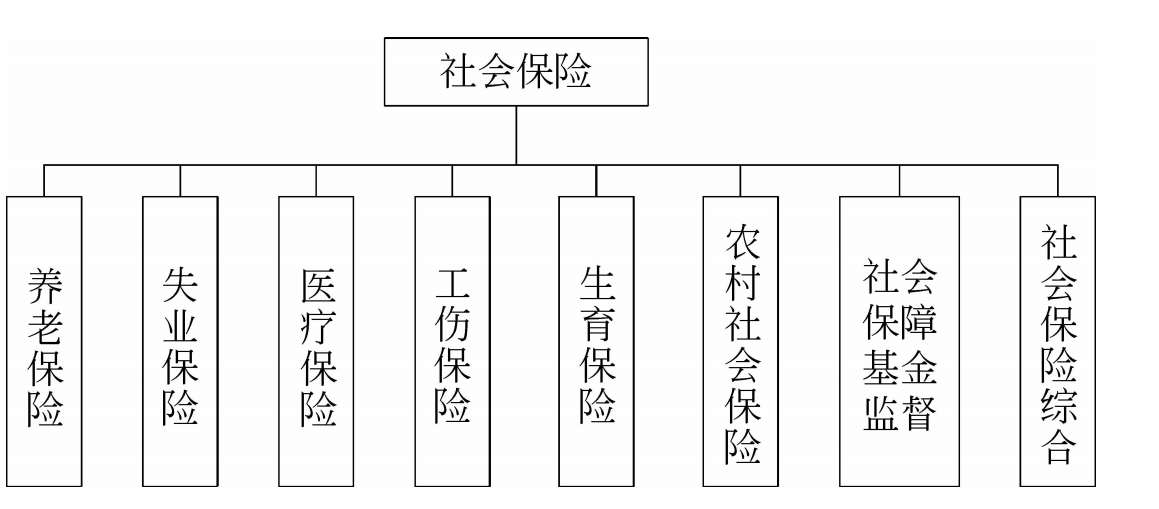


图3 我国社会保险政策分类

通过构建政策血缘网络为政策制定者提供方向指引，同时有效地延缓和避免政策碎片化。

**（二）构建企业关键技术演进路线，准确把握科技发展方向**

伴随着科学技术地快速发展，学科间的交叉融合与技术领域的汇聚发展已成为主流，不同技术领域融合形成的新技术——跨领域技术,同时，它也代表了主要的技术创新源。而专利文献作为反映技术发展过程的核心文本数据，分析专利文献也成为研究企业技术演进路线的重要手段。当前，专利文献的技术分类仍然使用国际专利分类表进行划分。但是，国际专利分类表中缺乏跨领域技术的分类，因此不再单一的使用国际专利分类表进行专利文献的技术分类，而是借助文本分析方法提取专利包含的关键技术，并以此划分专利技术类别。

利用深度学习框架Doc2Vec对专利文献进行文本分析，剖析专利文献所属技术领域、所属关键技术。本文首先将专利按照“文章—段落—关键词”逐级“粉碎”处理，提取专利中的技术类关键词，并根据相似度进行分类，最后借助专家的专业能力从诸多技术类别中审核出关键技术，从而组建关键技术语料库。分析流程图如图4所示。



图4 专利关键技术分析流程图

语料库中的数据主要来源于专利文献，因此语料库中包含多个维度的数据，例如时间、区域、产业、申请单位、发明人、基础技术、前沿技术等多个维度。其中，基础技术维度和前沿技术维度是对关键技术科学含量的划分，关键技术科学含量低判定为基础技术，含量较高判定为前沿技术。语料库维度图如图5所示。



图5 语料库维度图

充分利用多维度数据，获取多种分析结论。借助语料库的数据，可以对关键技术进行七个维度的分析，譬如，以时间为研究对象，可得到关键技术随时间发展趋势；以区域为研究对象，可摸清关键技术的起源地、发展地、需求地等信息；以申请单位为研究对象，则可以获悉掌握关键技术的主要单位等信息。不同维度之间的组合又可以得到更多的信息，例如，以区域和时间为研究对象，可清楚关键技术的发展密集区域和密集期间段；以时间和申请单位为研究对象，可清晰地看到单位的关键技术历史演进图；以申请单位和前沿技术为研究对象，可摸清掌握前沿技术的申请单位。诸如此类，举不胜举。因此，以专利为数据来源的关键技术语料库，可以进行多维数据的多方式结合，从而更加全面地了解关键技术的过去、现在和未来。

通过构建企业关键技术演进路线，充分摸清企业关键技术演变过程，指引企业技术发展方向，增强企业专利成果转化关注度，从而增加技术发展方向准确性，确保专利技术创新性，提高专利的市场转化率。

**（三）搭建语义检索平台，提供知识产权精准检索服务**

知识产权对政府和企业意义重大，但从现状来看，政府和企业在利用知识产权时存在重重的困难，其中主要的问题包括繁重的数据检索和筛选工作，同时还有数据零散，以及各国语言、法律法规的差异造成的检索困难，使得检索效率下降，准确率不高。

通过深度学习框架Doc2Vec能够有效地提升检索的效率及准确性。该方法的大体流程如下：首先通过对用户输入的文本进行基于Doc2Vec的文本向量化，然后从知识产权语料库中寻找与文本相似的文本段落，并返回给用户。通过该方法不仅可以大量减少投入的人力物力，避免了需要专业的专利工程师来编写检索表达式，提高检索的全面性和准确性。其检索流程如图6所示。



图6 知识产权语义检索

充分利用语义检索深入挖掘知识产权中隐含的信息，应用于项目查重、文献查新等场景，为企业发展提供导航，维护企业合法权益，培养市场创新能力。

**（四）建设智能客服系统，为科技创新主体提供高效服务**

为了更好地服务科技创新主体，打造服务型政府，建议在云平台的基础上建设智能客服系统。随着云平台用户的剧增，在使用过程中不可避免产生许多问题。以项目申报系统为例，每当临近项目截止日期时许多用户会操作不当或者对申报系统的不熟悉而产生诸多问题。如果系统可以为用户及时地解答疑问，便会极大地提高用户满意度。但是依靠人工客服无法提供全面的服务，当询问量过多时，明显不能满足需求，且流程繁琐。为解决这一问题，可以利用深度学习框架Doc2Vec对文本进行语义分析形成智能客服系统，迅速反馈给用户准确、简洁的答案。

智能客服系统主要由三部分组成，分别是问句分析、信息检索、答案抽取，如图7所示。1）问题分析将用户语言转换成计算机能够识别的语言，主要包括对问句进行中文分词、词性标注、句法分析，为了能够快速准确找到答案，还要对问句进行分类，最后进行关键词提取和拓展；2）信息检索，首先利用Doc2Vec将常见问题、热点问题进行聚类并形成问题库，然后对用户提出的问题进行相似度判定；3）答案抽取，通过问句相似度计算在问题库中找到与提问问题相似的结果，从高到低进行排序，最终提供答案数据

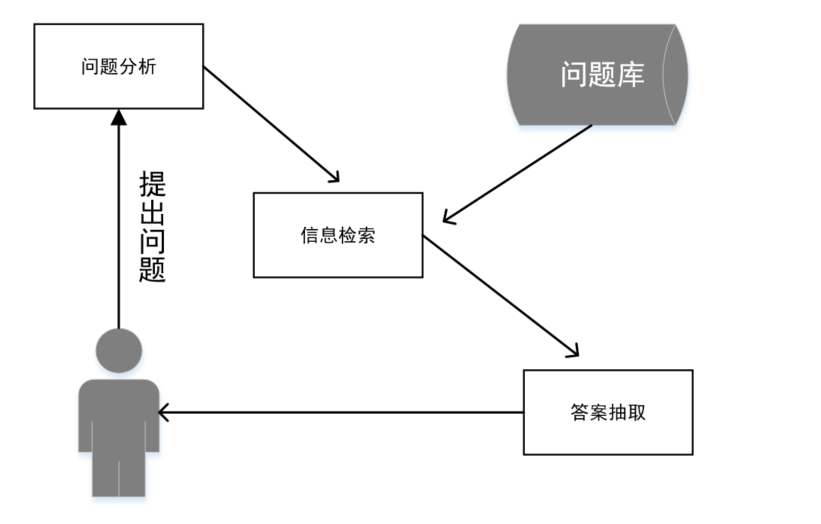


图7 智能客服系统

智能客服系统在云平台为科技创新主体服务的过程中起着显而易见的优势，1）使用便捷、自然的语言交互服务，提升用户体验，降低用户使用难度，让人工智能服务于用户;2）提升服务效率，缩短咨询处理时限，节省服务成本;3）收集用户诉求和行为数据，支撑产品迭代优化。

加快建设人民满意的服务型政府，使得客服系统更加智能化。智能客服为创新主体提供高效服务，为科技高质量发展添砖加瓦。

1. **总结**

本文借助深度学习框架Doc2Vec，积极探索该框架在科技系统中的应用场景。切实做到为政策制定者谋方向，为企业谋权益，为科技创新主体谋服务。进一步促进技术、产业、应用快速发展，抢占战略制高点，支撑科技高质量发展。依靠政策血缘网络为政策制定者指引方向，便于政策系统化管理；通过分析企业关键技术演进路线，摸清产业技术发展态势，指引产业技术发展方向；借助语义检索，实现快速检索和精确匹配，维护企业合法权益，培养市场创新能力；建立智能客服系统，加快建设服务型政府，提升人民群众对政府工作的满意度。